A quantitative approach to the use of the Wikipedia

Antonio J. Reinoso, Jesus M. Gonzalez-Barahona, Felipe Ortega and Gregorio Robles
Libresoft Research Group, Universidad Rey Juan Carlos
Tulipan s/n, 28933, Mostoles, Madrid. SPAIN
{ajreinoso, jgb, jfelipe, grex}@gsyc.es

ABSTRACT

This paper presents a quantitative study of the use of the Wikipedia system by its users (both readers and editors), with special focus on the identification of time and kind-of-use patterns, characterization of traffic and workload, and comparative analysis of different language editions. The basis of the study is the filtering and analysis of a large sample of the requests directed to the Wikimedia systems for six weeks during the range of months comprising November 2007 to April 2008. In particular, we have considered the twenty most frequently visited language editions of the Wikipedia, identifying for each access to any of them the corresponding namespace (sets of resources with uniform semantics), resource name (article names, for example) and action (editions, submissions, history reviews, save operations, etc.). The results found include the identification of weekly and daily patterns, and several correlations between several actions on the articles. In summary, the study shows an overall picture of how the most visited language editions of the Wikipedia are being accessed by their users.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*clustering, information filtering, retrieval models search process, selection process.*

General Terms

Measurement, Languages

Keywords

Wikipedia, quantitative analysis, temporal patterns, request characterization, workload analysis

1. INTRODUCTION

Wikipedia has successfully grown into a massive collaboration tool, based on the wiki paradigm as a new way of producing and accessing intellectual works. The popularity of the Wikipedia has not stopped growing, being currently the 8th most visited web site on the Internet according to Alexa's ranking[1]. The Wikipedia contains approximately 8 millions of articles distributed in 250 different language editions. It is composed of several editions, each one associated to a different language and with a particular subdomain. Thus, any URL beginning with "http://en.wikipedia.org" refers to the English edition of the Wikipedia, whereas URLs beginning with "http://es.wikipedia.org" will refer to the Spanish edition.

All the Wikipedia editions, as well as others such us Wikiversity, Wiktionary or Wikiquote, are being maintained by the Wikimedia Foundation. This non-profit organization offers data feeding to research groups interested in their projects and activities.

Despite this significant relevance of the Wikipedia on the current websites scenario, there are few studies describing the overall operation of the Wikipedia ([1], [3] and [5]). In particular, we have found just one report [6] involving the specific topics of Wikipedia traffic or Wikipedia patterns of use.

In this paper we are reporting the use of operation-related data from the Wikipedia in order to analyze the overall use of the system. This kind of analysis provides a better understanding of the system and allows establishing a model of utilization. In addition, data collected may lead to technical improvements in the operation of the Wikipedia system that could be applied to other, similar Internet-based systems.

Our research work has focused on finding temporal access patterns and other sort of characterization, such as the ones based on the namespace of the articles requested or on the different actions requested by users in the Wikipedia.

In this way, in our analysis we first filter for general requests directed to a set of specific Wikipedia editions, and then specify some namespaces and actions in order to establish some kind of correlation between the two measurements.

[1] `http://www.alexa.com/data/details/-traffic_details/wikipedia.org`

The results of our analysis and characterization work include the appreciation of several visiting patterns depending on the hour of the day and on the day of the week as far as several other comparisons among the various kinds of contents and actions requested by users.

The rest of the article is structured as follows: First, we provide a brief description of the data sources considered, and then we introduce the methodology followed to conduct our work. After this, we show the quantitative results and discuss them. Finally we present our conclusions and propose ideas for future work.

2. THE WIKIPEDIA SYSTEM AND LOGS

This section is aimed to describe the data sources considered for our analysis and their relationships with the Wikipedia system. In particular, the Squid-based subsystem, which receives all the users' requests, will be introduced. Since the logs obtained from this subsystem are the fundamental data source for the study, they will also be described in detail. Some important aspects of how information is organized in the Wikipedia system (e.g., namespaces) will also be presented.

2.1 The Wikipedia Squid subsystem

Most of the pages requested to the Wikipedia by non-logged users can be served avoiding database requests and HTTP server operations. This is the main reason for the Wikipedia system to use a Squid-based caching front-end for improving performance in common activities. The subsystem is composed of a set of Squid servers which are capable of handling most of the requests made by non-logged users. In particular, during high load peaks because of a broad request of media, the Squids manage almost all the traffic without the need to access the rest of the system. In any case, they handle all user requests received by the Wikipedia system.

The Squid subsystem implements a HTTP reverse proxy caching. There are three Squid server clusters: a primary cluster (located in Tampa, Florida) is placed in front of the Apache web servers, databases and media storage systems that support Wikipedia. Two secondary clusters (located in Amsterdam and Seoul) perform only web caching. The Squids at Wikipedia are currently running at a hit-rate of approximately 85% for text and 98% for media, using CARP (Cache Array Routing Protocol) [2].

As content serving has a different access/communication pattern than media serving, each Squid cluster has been split into task-oriented groups [4]. Moreover, each cluster uses a multiple-tier approach with a front layer using the CARP algorithm for URI-based distribution, and a second layer of Squid servers that deals with requests from clients.

This hash-based distribution reduces the number of cached copies of objects and allows an efficient handling of node failures by the redistribution of requests across other active machines, as a function of a weight assignment policy.

The normal operation rate of a Wikipedia Squid server is over 1000 HTTP requests per second (although it is possible to reach peaks of 2500 HTTP requests/second) [2].

2.2 Wikipedia Squid logs

Squid servers save all related information about HTTP transactions (requests by user browsers) in a file called *access.log*. This is a line-based file, with each line corresponding to a request from a client. Squid usually records all HTTP accesses, except for those disconnected before any data could be sent. Squid also registers all ICP (but not HTCP) transactions, unless disabled with the appropriate directive.

Log lines are buffered and sent to a centralized host, henbane.yaseo.wikimedia.org, in 1450-byte UDP packets[2]. A program called udp2log logs into the files or pipes specified in a configuration file which defines a sampling factor (the number of lines that will be received before sending one to the specified destination) for each destination (feed).

2.3 The Wikipedia Squid log format

Log lines currently offered by the Wikimedia Squid system to our feed do not contain all the Wikipedia Squid log format fields, but just those marked as received in Table 1. In particular, any field compromising the privacy of Wikipedia users is not included. In addition to those aforementioned fields contained in the Wikipedia Squid log format, our feed receives two more fields:

1. A reference number included by the log2udp program when it sends information; it is used to track the loss of packets.
2. A field indicating if the request caused a save operation on the data base storage system.

It is important to remark that each of the lines of the Squid logs corresponds to an user request and, therefore, can be used to track and characterize how users (both readers and editors) are interacting with the Wikipedia system.

2.4 Wikipedia namespaces

Namespaces are prefixes before an article or page title (like "User:" or "Talk:") that allow to have these pages under multiple names, but serving for various purposes. They can be viewed as folders that separate different basic types of information or functionality and become a useful tool to establish a distinction, for example, between a concrete content and its discussion issues. While new namespaces can be added, the number of namespaces in a wiki is typically relatively low[3].

Wikipedia uses eighteen built-in namespaces: the main namespace, where article titles have no prefix, and other seventeen, each with its own prefix. In addition, there are two custom namespaces, with their own prefixes[4].

Page titles in MediaWiki are composed of two parts, an optional namespace name and the remainder of the title, separated by a colon {:}. Articles in the main namespace are those most commonly requested. It is important to mention that if a page title contains a colon, but the initial part of

[2] `https://wikitech.leuksman.com/view/Squids`
[3] `http://en.wikipedia.org/wiki/MediaWiki`
[4] `http://en.wikipedia.org/wiki/Wikipedia:Namespace`

Table 1: Relevant Wikipedia Squid log format fields.

Field	Description	Received
Squid Hostname	Squid server that writes the log line.	
Sequence number	Sequence number of the writing operation for the Squid Server.	
Time since epoch (ms. resolution)	Timestamp	Yes
Request service time (ms.)	Total spent time to serve the logged client request.	Yes
Client IP address	Client IP address	
Reply size including HTTP headers	Number of bytes transferred to the client (includes overheads) because of TCP/IP headers	Yes
Request method	Request method: HTTP- or ICP-specific.	Yes
URL	URL containing the request.	Yes
MIME content type.	MIME header corresponding to the URL.	

the title is not one of the pre-defined namespaces, that page is considered to be in the main namespace.

3. METHODOLOGY OF THE STUDY

The study is based on the analysis of the Wikipedia Squid log feed received by GSyC/Libresoft. It consists of a 10% sample of the lines logged by the udp2log program at henbane.yaseo.wikimedia.org, therefore corresponding to a 10% of the total traffic directed to all the projects maintained by the Wikimedia Foundation. The packets containing the Squid log lines are received by a syslog-ng client in our facilities, which buffers and, finally, writes them to a log file which is rotated and stored daily. Since each log line corresponds to an user request, our interest in them is based on the fact that their analysis will provide description patterns of how people use the Wikipedia.

From this feed, we have selected six complete weeks. Each week belongs to a different month between November 2007 and April 2008. We have considered each week as a period starting at 00:00:00 on Mondays and ending at 23:59:59 the next Sunday. The analysis of the overall group of weeks has involved more than a 1 billion Squid log lines, which have been parsed by an ad-hoc Java-written multithreaded application in order to store some of their most relevant fields into a MySQL relational database.

To ensure that the study involved mature and highly active language editions of the Wikipedia, only the requests corresponding to the editions in Table 2 were considered (the lines not corresponding to them were filtered out). In fact, these editions are the top-twenty regarding only their volumes of Squid log lines which may also serve to provide an estimation of the total traffic managed by the Wikipedia servers.

The number of requests in the Squid log files directed specifically to Wikipedia resources corresponds each week to a percentage in the range of 40%-50% of the total requests. The rest of the URLs are directed to other Wikimedia Foundation projects (such as Wiktionary, Wikiquote, etc.) and to static files such as images or templates. The editions of the Wikipedia filtered for this study summarized more than the 96% of the total traffic directed to all the language editions of the Wikipedia. This fact points out that only a small portion of all the Wikipedia requests is not being considered in our analysis.

The parsing process has been fine-tuned to avoid the loss of lines caused by errors when inserting log line fields into the database. From all the lines processed, just around 6,000 could not be inserted into the database due to wrong characters in the URL.

Some of the parameters and the indicators we were looking for could not be extracted directly from the Squid log fields. In particular, we had to parse URLs looking for:

1. The targeted Wikimedia project, such us Wikipedia, Wiktionary, Wikiquote, etc.
2. The language edition of the Wikipedia.
3. The namespace to which the request is related.
4. The action (edit, submit, history review...) requested by the user (if any).
5. The type of static files when available.
6. The title of the article.
7. The user page name.

The namespace included in the requests allows to study the distribution of the accesses within a given language and, most important, may serve as a good indicator of the activity of users, since contributors usually read specific namespaces, such as the discussion namespace.

For this work, we have considered requests corresponding to articles in the main and discussion namespaces, as well as those corresponding to the user (i.e., prefixed with "User:" in the English Wikipedia) and user discussion (prefixed "User_talk:" in the same Wikipedia) namespaces. It is important to notice that namespace prefixes are usually translated for each language edition. For example, the discussion namespace is referred to as the "Talk:" namespace in the English Wikipedia, as the "Diskussion:" namespace in the German Wikipedia and as the "%E3%83%8E%E3%83%BC%E3%83%88:" namespace in the Japanese Wikipedia. This fact has been taken into account carefully when developing the parser algorithm in order to avoid inacurate data as a result of a wrong classification of requests directed to the same particular namespace in each filtered edition of the Wikipedia.

Table 2: Filtered editions of the Wikipedia.

Code	Language
EN	English
DE	German
ES	Spanish
JA	Japanese
FR	French
PL	Poland
IT	Italian
PT	Portuguese
NL	Dutch
RU	Russian
TR	Turkish
ZH	Chinese
SV	Swedish
HE	Hebrew
FI	Finland
AR	Arabic
ID	Indonesian
HU	Hungarian
CS	Czech
NO	Norwegian

Table 3: Squid Log lines corresponding to each week under study.

Week	Log file date	Total lines	Percent
Week 1	05-11-07 to 11-11-07	187,604,039	17.95%
Week 2	10-12-07 to 16-12-07	165,083,902	15.79%
Week 4	07-01-08 to 13-01-08	159,229,450	15.23%
Week 4	04-02-08 to 10-02-08	171,789,912	16.43%
Week 5	05-03-08 to 11-03-08	182,482,947	17.46%
Week 6	07-04-08 to 13-04-08	179,244,933	17.15%

Apart from URLs that request articles in the mentioned namespaces, we have also classified users' URLs requesting some specific actions. In particular, we have focused on those requesting editions, submissions, savings and history reviews.

4. QUANTITATIVE RESULTS

As we mentioned before, more than 1 billion of lines from the Squid log files corresponding to the analysis period have been processed as part of this work. The number of lines corresponding to each week is shown in Table 3. As we can observe, every week present a very similar number of requests and none of them has a significant relevance in the overall workload. This is because we were looking for weeks which did not correspond to special periods such us holidays; those affected by relevant events or situations were also disregarded.

First of all, we will compare the evolution of the requests to all the editions of the Wikipedia with the global traffic to all the Wikimedia Foundation projects. This comparison results in Figure 1 where the evolution of the requests to the editions of the Wikipedia used for our analysis is also represented. As we can observe, the Wikipedia requests do not decrease in the same way that the general Wikimedia Foundation projects do in the second week, so both evolutions are relatively independent. In particular, this picture confirms that the language editions of the Wikipedia considered for our analysis behave in exactly the same way in which all the editions do. The closeness between the lines representing all the Wikipedia editions and those considered for this work reinforces our affirmation of a very irrelevant loos of data due to the not-considered Wikipedia editions.

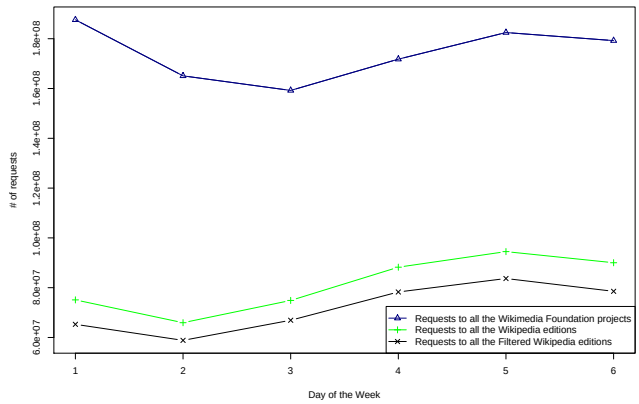

Figure 1: Evolution of the number of requests made to all the Wikimedia Foundation project, to all the editions of the Wikipedia and to those considered for this analysis.

We will also consider the distribution of the Wikipedia requests over the days of each week under study. Figure 2 illustrates that the number of accesses remains quite similar in each day of all the weeks but Monday when two of the weeks have a significant decrease of the number o requests. Regarding the evolution of the number of the requests across the week days, all the weeks present a somewhat similar behavior decreasing the number of requests as the week advances. Saturday is the day with less number of visits to the Wikipedia in all the weeks and this number increases again on Sunday also in all periods. Considered times are CET, and therefore "weekend" includes, e.g., for time zones in America, the last hours of Friday).

Figure 3 shows the evolution of the requests to the Wikipedia along the hours of the day in every week under study. The number of requests decreases during the night and starts to grow in the early morning (as always, time is CET). After the lunchtime off-peak the number of visits rapidly increases to reach its maximum at approximately 19:00 hours when it starts reducing again.

One of our goals was to compare the number of requests directed to each particular edition of the Wikipedia in order to be able to determine the weight of each language edition in the overall server system workload. Thus, Table 4 presents the percentage of requests to each analyzed edition of the Wikipedia (counted as the total number of requests directed to it).

The resources and actions requested by the users to each considered edition of the Wikipedia were also analyzed. In

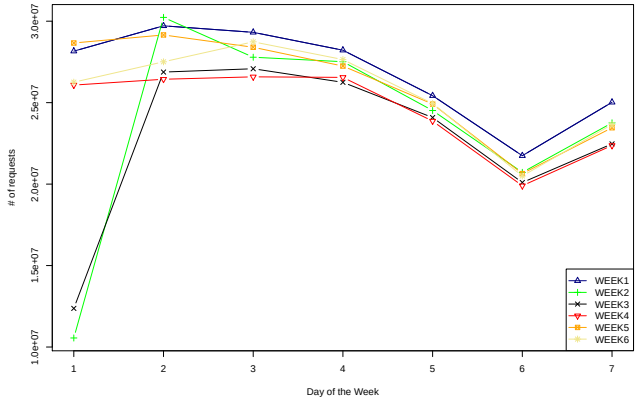

Figure 2: Daily distribution of the requests to the Wikipedia in each week under study.

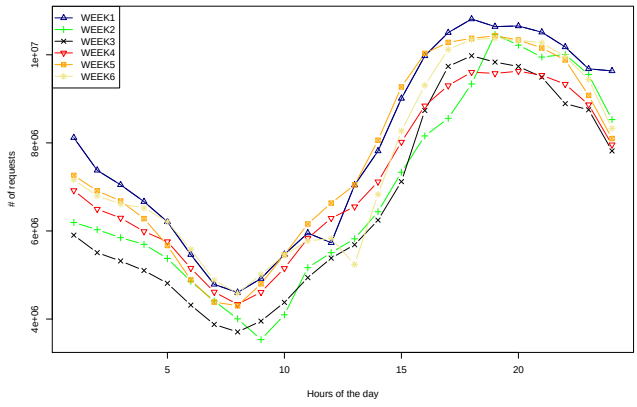

Figure 3: Hourly distribution of the requests to the Wikipedia in each filtered week.

Table 4: Requests to each analyzed edition of the Wikipedia.

Code	Percent
EN	49.45%
ES	8.96%
DE	8.91%
JA	7.97%
FR	4.32%
PL	3.29%
PT	3.10%
IT	2.66%
NL	1.48%
RU	1.36%
TR	1.06%
ZH	0.96%
SV	0.72%
HE	0.57%
FI	0.55%
NO	0.38%
CS	0.35%
AR	0.30%
HU	0.29%
ID	0.24%
OTHERS	3.85%

this way, we have classified the URLs issued when users made their requests for viewing articles according to the filtered namespace to which they were directed. On the other hand, we have also focused on the URLs specifying any filtered action which had been requested by the user. Table 5 shows the results of this first classification. It is important to note that all requests specifying any filtered action are grouped together whether the action involves, or not, an article in filtered namespaces.

If we relate the total number of requests directed to the analyzed Wikipedias to the number of requests directed to articles in the main namespaces in the same Wikipedia projects we will obtain a correlation between the two measurements, as shown in Figure 4. In the same way, Figure 5 shows the correlation between the total number of requests directed to each language edition and the number of those directed to articles in the discussion namespace.

The great difference between the number of requests directed to articles in the main namespace and those directed to articles in the other considered namespaces can be clearly seen from Figure 6.

Talking about filtered actions requested by the users, Table 6 summarizes and classifies them whereas Figure 7 represents the same classification for each analyzed week in order to determine different users behavioral patterns. The graphic shows that the proportion of each kind of action remains practically unchanged for all weeks.

We thought that an interesting analysis and some kind of comparison could be made focusing on the most visited language editions. For this reason, we have studied and compared the volume of the total traffic directed to the top-eight

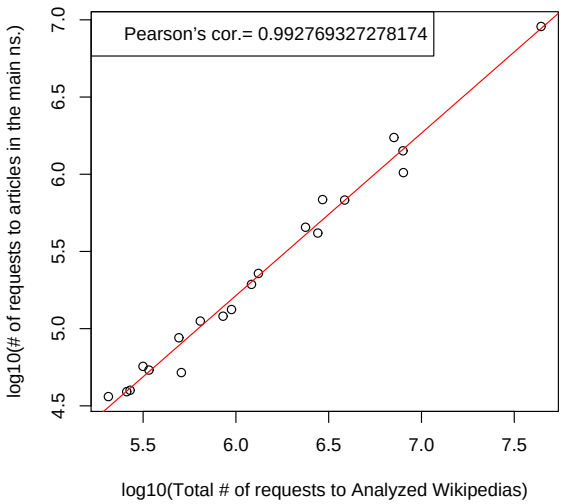

Figure 4: Total number of requests in analyzed Wikipedias against number of requests directed to articles in the main namespace.

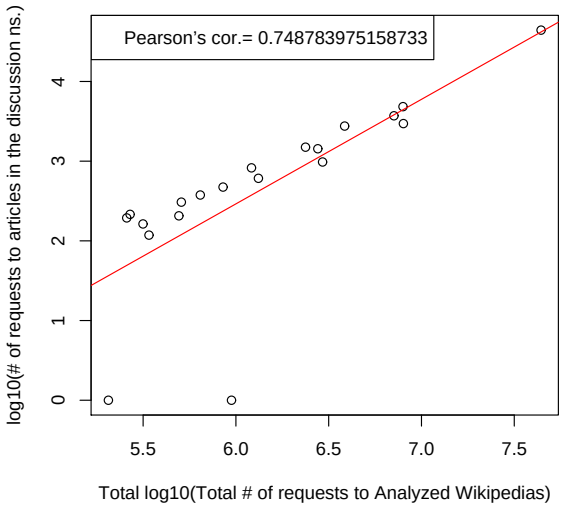

Figure 5: Total number of requests in analyzed Wikipedias against the number of articles in discussion namespace.

Table 5: Contents and actions requested to the considered editions of the Wikipedia.

Content	Total requests
Main Namespace Articles	20.14%
Discussion Namespace Articles	0.38%
User Namespace Articles	0.24%
User Discussion Namespace Articles	0.23%
URLs requesting filtered actions	0.60%

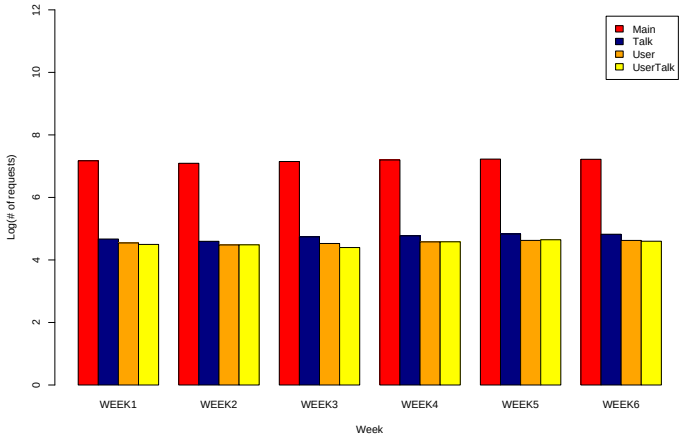

Figure 6: Requests directed to each filtered namespace.

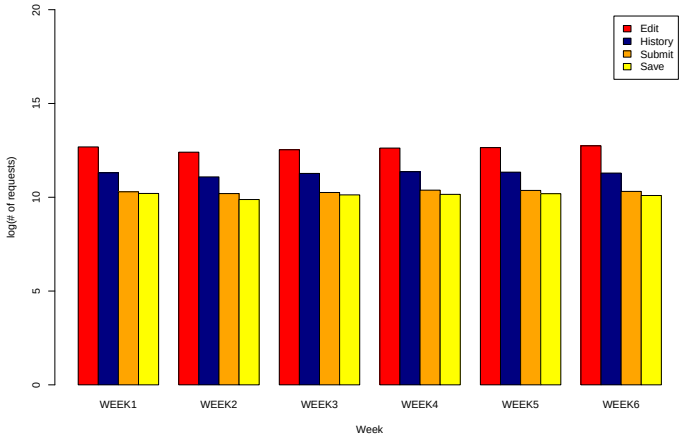

Figure 7: Filtered requested actions.

Table 6: Analyzed Wikipedia requested actions.

Requested action	Percent
edit	69.18%
history	18.26%
save	6.87%
submit	5.69%

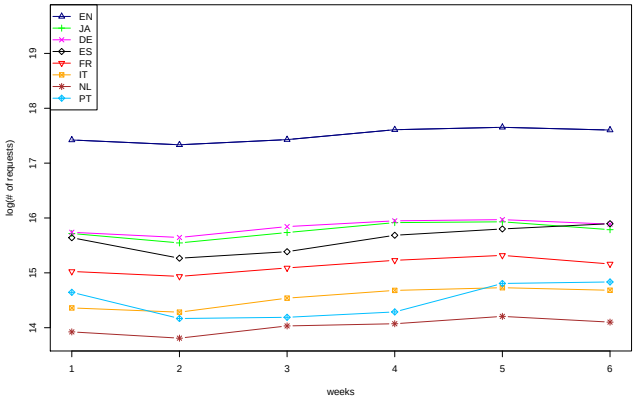

Figure 8: Total Requests to some of the filtered Wikipedias.

more requested editions (Figure 8) and also the traffic resulting of every request sent by the users of the same editions in order to read articles corresponding to the main namespace (Figure 9).

Both graphics show the hegemony of the English edition of the Wikipedia and verify the correlation presented in Figure 4, which means that a higher number or requests usually implies a higher number of users reading articles in the main namespace.

We found also very useful to classify the requested actions from each considered edition of the Wikipedia. Table 7 present the corresponding results.

We have considered combinations of namespaces and requested actions in order to search for relationships. For example, we have compared the total number of edition requests on articles in the main namespace and in the discussion namespace. In almost all editions, approximately between 80% and 90% of all edition requests were directed to articles in the main namespace, whereas the rest involved articles in the discussion namespace. A very similar ratio was observed in the case of "save" operations.

In order to evaluate the participation degree of users, we have defined ratios between requests to articles in the main namespace and requests that result in a save operation involving these articles. Results are presented in Table 8 and can be interpreted as the number of requests for reading an article in the main namespace corresponding to each save operation of these articles. Language editions with higher values denote a weak level of participation of their users who

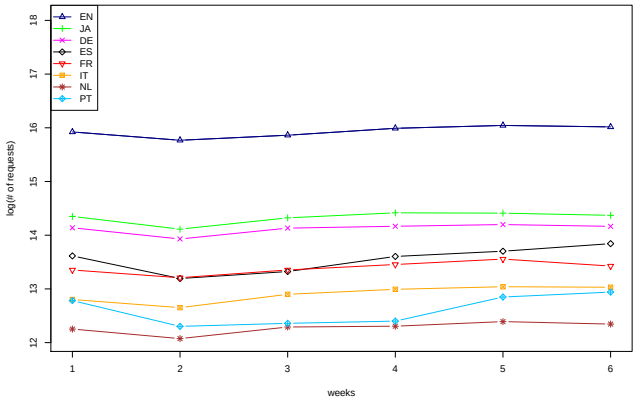

Figure 9: Requests to read articles in the main namespace in some of the filtered Wikipedias.

Table 7: Percentage of the Filtered requested actions in each edition of the Wikipedia.

Code	Edit %	Save %	Submit %	History %
AR	87.77	3.53	1.84	6.87
CS	76.07	7.80	7.17	8.97
DE	65.98	6.77	9.61	17.63
EN	66.37	7.14	4.58	21.91
ES	78.45	4.34	2.50	14.70
FI	62.18	10.23	14.91	12.68
FR	65.87	9.25	7.62	17.26
HE	67.40	12.69	7.90	12.01
HU	78.48	9.01	5.47	7.04
ID	77.89	7.91	9.04	5.16
IT	78.53	6.63	3.79	11,05
JA	86.35	1.76	2.17	9.73
NL	68.70	8.37	7.92	15.00
NO	66.99	9.98	12.92	10.12
PL	79.12	5.24	6.93	8.72
PT	67.92	4.83	5.71	21.54
RU	73.46	9.27	4.72	12.55
SV	59.95	10.89	11.38	17.78
TR	83.85	5.16	2.43	8.57
ZH	85.91	5.05	2.85	6.19

Table 8: Read/save operation ratio for articles in the main namespace.

Code	read/save ratio for articles int the main namespace
EN	619.83
JA	1455.52
DE	758.97
ES	546.99
PL	810.21
FR	305.79
IT	345.95
PT	267.71
NL	336.43
RU	220.51
TR	411.58
ZH	216.14
SV	258.12
FI	270.50
CS	251.12
NO	184.52
HE	211.45
AR	182.92

are limiting their interaction with the Wikipedia to read operations.

5. DISCUSSION

Many interesting conclusions can be obtained from the analysis of the presented data. In this section, we present some of the most notorious.

Figure 3 shows how the Wikipedia visits are progressively increasing as the day advances from 10:00 to 20:00 and progressively reducing as the day advances from 00:00 to 10:00. Given that all the times are CET, from this pattern we can infer the timing of the user habits of the Wikipedia, and in part, the geographical origin of a large fraction of the users. So, arbitrarily choosing CET 00:00 as the starting of the day, we can say that at present time, mostly Americans are using it, still in their evening hours, and East Asians, in the morning. As the Earth rotates, Europeans get involved in their morning hours, causing the "morning peak" at about 11:00, alongside with those East Asians in their evening. The largest peak of use, however, comes in the "afternoon", while Europeans are in their afternoon hours, and Americans (in their morning) join them. Of course other parts of the world also use the Wikipedia, but the daily pattern, together with the most visited languages, seem to suggest that this is the case. A more detailed geo-targeting study of IPs would help to make a more detailed study, of course.

A second consequence of this graphic would suggest that the Wikipedia is most used during working hours, since peaks correspond to times when people are usually working in the parts of the world where a high proportion of users come from. This is in part reinforced by Figure 2, which shows how requests to the Wikipedia tend to reduce during weekends, when people are usually off-duty.

Significant variations between different days can be explained by the relative short period considered for its analysis and needs to be contrasted in future works.

With respect to the use of the different editions, the traffic analysis shows the absolute hegemony of the English edition, receiving almost 50% of the total visits to all the Wikipedia projects, but also shows how the order according to visits is not the same than according to number of articles. For example, Spanish, which is the tenth by number of articles, is second (during the analyzed weeks) by number of visits.

The same analysis allows to determine that the greatest number of Wikipedia accesses is directed to articles in the main namespace, whereas the number of visits to the discussion namespace is rarely notorious. This would also provide a first metric for the relationship between accesses just interested in the content of the Wikipedia as opposed to those interested in how that content was produced (which is somewhat reflected by the discussion pages).

When talking about actions, we remark the percentage of requests resulting in save operations compared to all others. This means that requests introducing contributions suppose less than 7% of all the Wikipedia accesses.

Regarding each particular Wikipedia, the most frequently requested action is performing an edition, with a remarkable rate of the history reviews. Assuming that the goal of each article edition is its corresponding contribution, the ratio edition/save operations is remarkably poor and has to be studied in a more detailed way.

As it was expected, the largest numbers or editions and save operations are directed to articles in the main namespaces, since they are receiving also the largest number of visits.

Finally, the read/save operation ratio illustrates how active and collaborative users of a particular Wikipedia are. In fact, Wikipedias whose ratio is meaningful large are the ones where users make a very poor number of contributions in comparison to the number of visits that only read content.

6. CONCLUSIONS

The logs received from the Wikipedia Squid systems have allowed us to build several patterns describing when and how people are visiting the Wikipedia. In fact, we have been able to accomplish our main goal of finding relationships between the number of accesses to a particular Wikipedia and the content requested. Finally, we could determine the contribution and activity level of analyzed Wikipedia users by studying some ratio involving the actions requested.

7. ACKNOWLEDGMENTS

We wish to thank Wikimedia Foundation, especially Tim Starling, for providing us with all the data used in our analysis tasks. We want also thank to the GSyC/Libresoft systems administration group for its cooperative availability and its technical collaboration.

8. REFERENCES

[1] T. B. Adler and L. de Alfaro. 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, 2007. ACM Press.

[2] M. Bergsma. Wikimedia architecture presentation, 2007.

[3] D. L. Mcguinness, H. Zeng, P. P. da Silva, L. Ding, D. Narayanan, and M. Bhaowal. Investigations into trust for collaborative information repositories: A wikipedia case stud. In *Proceedings of the Workshop on Models of Trust for the Web*, 2006.

[4] D. Mituzas. Wikipedia: Site internals, configuration, code examples and management issues, December 2007.

[5] R. Priedhorsky, J. Chen, Shyong, K. Panciera, L. Terveen, and John. Creating, destroying, and restoring value in wikipedia. November 2007.

[6] G. Urdaneta, P. Guillaume, and M. Van Steen. Wikipedia workload analysis.