

A QUANTITATIVE EXAMINATION OF THE IMPACT OF FEATURED ARTICLES IN WIKIPEDIA

Antonio J. Reinoso, Jesus M. Gonzalez-Barahona

GSyC/Libresoft, Universidad Rey Juan Carlos, Madrid, Spain {ajreinoso, jgb}@gsyc.urjc.es

Rocío Muñoz Mansilla

Department of Applied Computer Science, UNED, Madrid, Spain rmunoz@dia.uned.es

Israel Herraiz

School of Engineering, Universidad Alfonso X el Sabio, Madrid, Spain herraiz@uax.es

Keywords: Wikipedia, featured articles, usage patterns, traffic characterization, quantitative analysis

Abstract: This paper presents a quantitative examination of the impact of the presentation of featured articles as quality content in the main page of several Wikipedia editions. Moreover, the paper also presents the analysis performed to determine the number of visits received by the articles promoted to the featured status. We have analyzed the visits not only in the month when articles awarded the promotion or were included in the main page, but also in the previous and following ones. The main aim for this is to assess the attention attracted by the featured content and the different dynamics exhibited by each community of users in respect to the promotion process. The main results of this paper are twofold: it shows how to extract relevant information related to the use of Wikipedia, which is an emerging research topic, and it analyzes whether the featured articles mechanism achieve to attract more attention.

1 INTRODUCTION

Wikipedia continues to be an absolute success and stands as the most relevant wiki-based platform. As a free and on-line encyclopedia, it offers a rich collection of contents, provided in different media formats and related to all the areas of knowledge. Undoubtedly, the Wikipedia phenomenon constitutes one of the most remarkable milestones in the evolution of encyclopedias. In addition, its supporting paradigm, based in the application of collaborative and cooperative efforts to the production of knowledge, has been object of a great number of studies and examinations.

This significant relevance has made Wikipedia to become a subject of increasing interest for the research community. However, this research usually concerns quality and reliability aspects about the Wikipedia contents (Priedhorsky et al., 2007; Olleros, 2008) or examines its growth or evolution (Suh et al., 2009). By contrast, very few studies have been devoted to analyze how Wikipedia users visit and browse the encyclopedia.

This paper presents an analysis aimed to determine the attention, measured in number of visits, that featured articles attract when they are included

as quality content in the main pages of different Wikipedia editions. Furthermore, we have also considered the differences in traffic to featured articles during the discussion of their promotion in order to assess if different communities exhibit different dynamics when looking for consensus.

The rest of the article 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 prior to a proper discussion about them. Finally, we propose some ideas for further work and present our conclusions.

2 THE DATA SOURCES

Visits to Wikipedia are issued in the form of URLs sent from users' browsers. These URL's are registered by the Wikimedia Foundation Squid servers in the form of log lines after serving the requested content. Squid servers are a special kind of servers performing web caching which are used by the Wikimedia Foundation as the first layer to manage the overall traffic directed to all its projects. Part of the informa-

tion they register is sent to universities and research centers interested in its study.

2.1 THE WIKIMEDIA FOUNDATION SQUID SUBSYSTEM

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. These log lines are sent in a UDP-packet stream to our facilities where they are conveniently stored.

The Wikimedia Foundation Squid servers use a customized format for generating their log lines. However, we do not receive all of the registered information, basically in consideration to users' privacy, but just several fields of the log format. The most important field we receive is the URL which constitutes the submitted request. In addition, the date of the request and a field indicating whether it causes a write operation to the database are also included.

2.2 FEATURED ARTICLES

Featured articles are considered the best articles all over the Wikipedia. In order to be promoted to this status, the articles, first, have to be nominated and included in an special page as candidates to featured articles. Usually and prior to the their nomination, future candidate articles pass through a peer revision process in which reviewers make suggestions to improve their quality.

Featured article have to meet a set of criteria apart from the requirements demanded to every Wikipedia article. These criteria cover from a clear and comprehensive writing of the article to a proper structure and organization. Other aspects such us stability, neutrality as well as length and citation robustness are also considered.

In what our research is concerned, we analyze the impact of featured articles in two very different ways. First, we consider the influence of the promotion of articles to the featured status in their number of visits. Then, we also study the impact of the presentation of a featured article as an example of high quality content in the main page of some editions of Wikipedia. Regarding this, our main goal is to determine some kind of pattern which can serve to model the traffic to an article after being considered as featured.

3 METHODOLOGY OF THE STUDY

The analysis presented here is based on a sample of the Wikimedia Foundation Squid log lines corresponding to two different sets, each made up of three months: Mars, April and May in one set, and September, October and November, in the other one. As we receive the 1% of all the traffic directed to the Wikimedia Foundation projects, this results in more than 8,200 million log lines to process for the considered months.

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 top-six visited editions have been considered.

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.

Browsing the special pages of each Wikipedia edition devoted to its featured contents, we obtained the featured articles promoted during April and October 2009. Moreover, we extracted the featured articles appearing in the main page during the same months. Then, we queried the database resulting from the processing of the Squid log lines to look for the number of visits corresponding to those articles during the aforementioned months as well as during the previous and the following ones in the aim of finding out what impact have the two featured mechanisms on the visits that articles get. We did all the analysis shown here using the GNU R statistical package (R Development Core Team, 2009).

4 STATISTICAL ANALYSIS

Before starting to analyze the impact of the featured articles, we characterized the volume and activity of each one of the considered Wikipedias. This will allow to cross correlate the results and conclusions with the dimension of the population under study. In this way, we decreasingly ordered the considered Wikipedias by size in number of articles and by number of received visits. As a result, we obtained

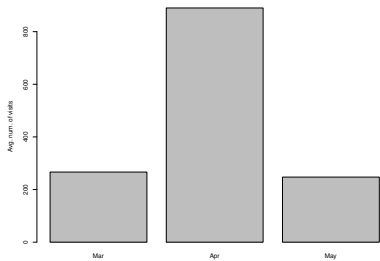


Figure 1: Average number of visits for today’s featured articles in the English Wikipedia during April 2009

that both ordinations were different. It is interesting the case of the Spanish Wikipedia, the smallest in size but the second in number of visits.

Figure 1 shows the average number of visits (or mean) for the featured articles presented in the main page of the English Wikipedia during April. At a first glance, it seems clear that the so-called “today’s featured articles” attract a great amount of attention in the month they are included in the main page, compared with the previous and the following ones.

If we analyze now the same metric applied to the articles just promoted to the featured status in April and November, we obtain that those articles do not receive always the highest number of visits in the month they are promoted as today’s featured articles did. This is probably due to the effect of the internal mechanism for promotion that entails a reviewing, a nomination and a consensus process. In this way, the different dynamics exhibited by each community of users in the promotion process are reflected in the visits that the articles attract.

If we focus on the case of the English Wikipedia, at a first glance, it seems that level of visits during April and October was higher than it was during the corresponding previous and following months, when the level of visits remained quite similar. It seems that, in both periods, the bulk of visits correspond to the months when articles are displayed in the main pages in all the Wikipedias except the Spanish one that presents a similar behavior in all the months.

To find out whether the differences in the median values for all the samples are negligible or not, we use a statistical test. Because the median values seem to be highly skewed in the box, the first step is testing whether the samples are extracted from a Normally distributed population. Depending on the result, we will choose a different statistical test to compare visits in different months.

For example the tables 1 shows the results of the Normality test for the visits to the featured ar-

Lang.	Sept		Oct		Nov	
	<i>W</i>	<i>p</i>	<i>W</i>	<i>p</i>	<i>W</i>	<i>p</i>
DE	0.94	0.64	0.96	0.85	0.91	0.33
EN	0.95	0.19	0.92	0.03	0.95	0.16
ES	0.94	0.63	0.91	0.30	0.97	0.85
FR	0.82	0.03	0.89	0.22	0.87	0.13

Table 1: Normality tests for featured articles displayed in the main pages. Only the month of September for the French Wikipedia and the month of October for the English one seem to be Normal ($p < 0.05$). The rest of samples are non-Normal.

Lang	S / O		O / N		S / N	
	<i>U</i>	<i>p</i>	<i>U</i>	<i>p</i>	<i>U</i>	<i>p</i>
DE	13	0.01	63	0.05	32	0.47
EN	140	0.00	645	0.00	337	0.37
ES	33	0.53	46.5	0.62	36	0.72
FR	25	0.19	52	0.34	38	0.86

Table 2: Results of the Wilcoxon rank-sum test for all the samples. In the English Wikipedia, the month of October gets more visits ($p < 0.05$). In the English and German Wikipedias, the month of October receives more visits. In the rest, the level of visits is similar. S: September, O: October, N: November

ticles displayed in the main page of the English (EN) and , Spanish (ES), German (DE) and French (FR) Wikipedias during the second considered set of months. The value of the *W* column is the Shapiro-Wilcox statistic, which indicates whether the sample is normal if and only if the *p* value is lower than a certain threshold (0.05 most often). For the presented figure only the month of September for the French Wikipedia and the month of October for the English Wikipedia present Normal distributions.

This non-normality of the samples implies that we have to test the median rather than the mean values, because the mean is highly biased for this kind of samples. Because of this issue, we decided to use a Wilcoxon rank-sum test (also known as Mann-Whitney-Wilcoxon test) to find out whether or not the appearance of a featured article in the main page implies a greater number of visits to those articles. This test is not sensitive to the normality of the data.

Table 2 shows the results of the test. The column labeled *U* shows the value of the statistic, and the column *p* shows the level of significance. These results indicate that featured articles displayed in the main pages attracted more visits during October only in the case of the English Wikipedia.

When examining the promoted articles, none of the central months attracted a number of visits significantly higher than the next ones. Again the expla-

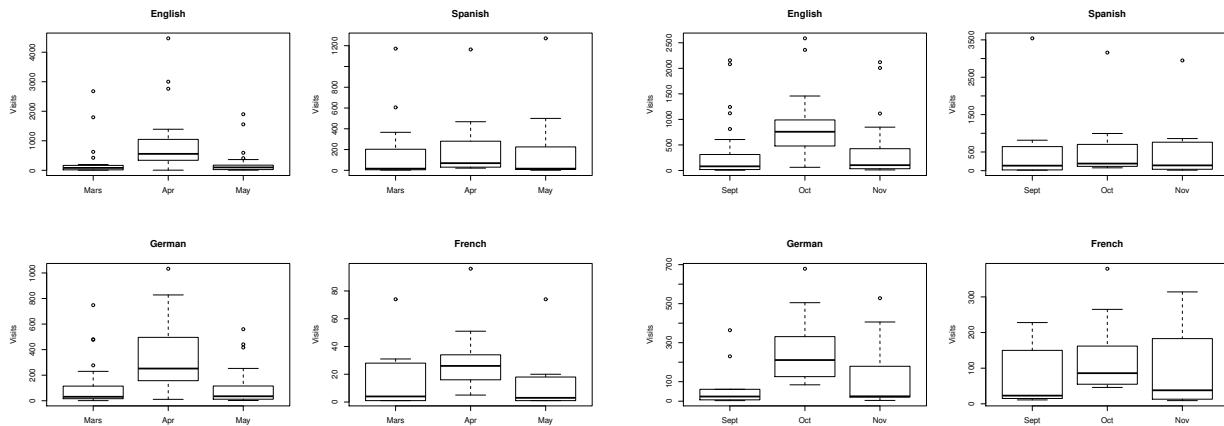


Figure 2: Boxplot of the visits to featured articles included in the main pages of the considered Wikipedias.

nation may reside in the different way of conducting when developing the promotion process.

4.1 SUMMARY OF RESULTS

For the featured articles displayed in the main pages of the four considered Wikipedias (by number of articles and by traffic), we have assessed whether their inclusion as example of featured content has an impact on the visits that those articles get. After an statistical analysis comparing the number of visits for the months right before and right after the month when the articles appear in the main page, our results indicate that such an impact surely happens for the English Wikipedia. Interestingly, we found the same impact in the German Wikipedia for the month of April. On the other hand, our study shows how the different dynamics involved in the articles' promotion process are reflected in very different patterns of visits to the articles awarded with the featured status.

5 CONCLUSIONS AND FURTHER WORK

Wikipedia, the largest wiki-based platform in the world, is a source of information for millions of people around the world. One of the resources of Wikipedia to improve the users experience is the featured status. Best articles are awarded with this status and they are included in the main pages as a sort of recognition.

Our results indicate that we can observe that increased attention only constant in the English Wikipedia. The German Wikipedia also presents a

significant increase of the visits to their featured articles shown in the main page in one of the considered periods. In the case of articles promoted to the featured status, our results show that there is not a common pattern of conducting in the promotion process. This subject clearly deserves an in-depth examination in the search for common features in all the processes.

REFERENCES

- Olleros, F. (2008). Learning to trust the crowd: Some lessons from wikipedia. In *e-Technologies, 2008 International MCETECH Conference on*, pages 212–216.
- Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., and Riedl, J. (2007). Creating, destroying, and restoring value in wikipedia. In *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, pages 259–268, New York, NY, USA. ACM.
- R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Suh, B., Convertino, G., Chi, E. H., and Pirulli, P. (2009). 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. ACM.