A STATISTICAL APPROACH TO THE IMPACT OF FEATURED ARTICLES IN WIKIPEDIA

Antonio J. Reinoso, Felipe Ortega, Jesus M. Gonzalez-Barahona
GSyC/Libresoft, Universidad Rey Juan Carlos, Madrid, Spain
{ajreinoso, jfelipe, jgb}@gsyc.urjc.es

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

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

Abstract: This paper presents an empirical study on the impact of featured articles on the attention that Wikipedia's articles attract, and how this behavior differs in different editions of 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 the corresponding request submitted by any Wikipedia user. The analysis has been conducted regarding the six most visited editions of the Wikipedia and has involved more than 4,100 million log lines corresponding to the traffic of September, October and November 2009. The methodology of work has mainly consisted on the parsing of the requests sent by the users and on their subsequent filtering according to the study directives. Relevant information fields has been finally stored in a database for persistence and further characterization. The main results of this paper are twofold: it shows how to use the the traffic log to extract information about the use of Wikipedia, which is a novel research approach without precedences in the research community, and it analyzes whether the featured articles mechanism achieve to attract more attention or not.

1 INTRODUCTION

Wikipedia stands as the most important wiki-based platform and continues providing the overall society with a great set of contents and media related to all the branches of knowledge.

Most of the previous researching involving the Wikipedia is concerned about the quality and reliability of its contents [Priedhorsky et al., 2007, Olleros, 2008, Javanmardi et al., 2009] or focus on describing its growth tendency and evolution [Tony and Riedl, 2009, Suh et al., 2009]. By contrast, the study presented here focuses on the impact on the number of visits of the consideration of a set of articles as featured. Moreover, the purportedly impact is analysed for different editions of the Wikipedia.

The rest of the article is structured as follows: first of all, we describe the data sources considered and then we introduce the methodology followed to conduct our work. After this, we present our results prior to discuss them. Finally we propose some ideas for further work and present our conclusions.

2 THE DATA SOURCES

This section aims to describe the data sources involved in our analysis and particularly focuses on the log lines containing information about the requests submitted to the Wikipedia by its users. This kind of information is offered by the Wikimedia Foundation to universities or education centers interested in it for research purposes. Therefore, the following sections present the principal issues related to these log lines.

2.1 THE WIKIMEDIA FOUNDATION SQUID SUBSYSTEM

Squid servers may be used to speed up web servers by caching the contents repeatedly requested to them.

Field	Description
GMT time	Current GMT time
Request service time (ms.)	Total time spent to serve the logged request
Request method	Request method (GET, POST, etc.)
URL	URL containing the request.

Table 1: Received fields from the Wikimedia Foundation Squid log format.

Under this approach, Squid servers are said to work as reverse proxy servers.

As a part of their job, Squid systems do log information about every request they serve whether the corresponding contents stems from their caches or, on the contrary, are provided by the web servers.

2.2 THE WIKIMEDIA FOUNDATION SQUID LOGGING FORMAT

The Wikimedia Foundation Squid servers use a customized format for generating their log lines. However, we do not receive all this information but just those fields summarized in Table 1.

2.3 FEATURED ARTICLES

Featured articles are considered the best articles all over the Wikipedia. These articles 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 organisation. Other aspects such us stability, neutrality as well as length and citation robustness are also considered. However, the purportedly quality of the Wikipedia featured article has sometimes been subject of discussion [Lindsey, 2010].

In what our research is concerned, the consideration of an article as featured can have a notable influence over the number of visits to the article during a period next to its promotion and may also affect to the number of contributions received during the same period.

3 METHODOLOGY OF THE STUDY

The analysis presented here is based on a sample of the Wikimedia Foundation Squid log lines corresponding to three months: September, October and November, 2009.

This analysis has focused just on the traffic directed to the Wikipedia project and to ensure that the study involved its most mature and highly active language editions, only the requests corresponding to the most visited ones have been considered.

3.1 THE WIKISQUILTER PROJECT

The analysis of the received log lines 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 aims and directives. As a result of both processes, necessary data to conduct a characterization are obtained and stored in a relational database for further analysis.

Not all the information suitable to perform an appropriate characterization can be obtained directly from the log lines. More in the detail, the application parser is devoted to determine the following information elements:

1. The Wikimedia Foundation project.
2. The language edition of the project.
3. The namespace of the article.
4. The action requested by the user (if any).
5. The searched topic (if any).
6. The title of the article.

4 STATISTICAL ANALYSIS

In this section we present a statistical analysis of the impact of the featured status on the attention that featured articles get.

4.1 VOLUME AND ACTIVITY OF THE WIKIPEDIAS

The visits to the considered editions of the Wikipedia are shown in figure 1, and include all kind of visits, like page views, editions, searches, etc.

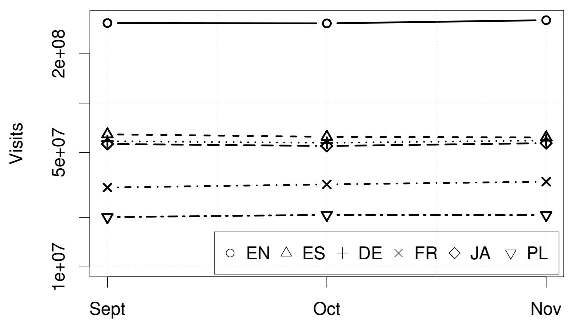

Figure 1: Number of visits in the Wikipedia editions under study.

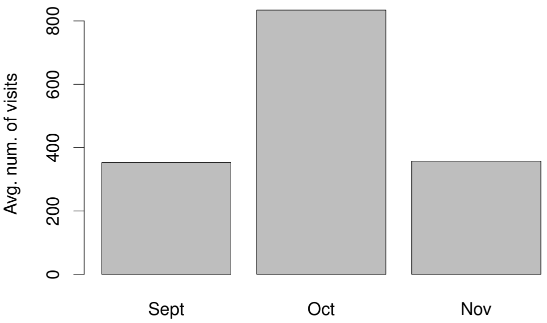

Figure 2: Average number of visits for November's featured articles in the English Wikipedia

4.2 IMPACT OF FEATURED ARTICLES ON VISITS

Searching on the different editions of the Wikipedia, we obtained a sample of all the featured articles corresponding to the month of October 2009. We extracted the number of visits for those articles during the months of September, October and November, with the aim of finding out what impact has the featured article mechanism on the visits that article gets. We did all the analysis shown here using the GNU R statistical package [R Development Core Team, 2009].

For instance, figure 2 shows the average number of visits (or mean) for the November's featured articles of the English Wikipedia. At a first glance, it seems clear that featured articles attract attention during the month where they are awarded the featured status, compared to the previous and following months.

Figure 3 shows a boxplot of all the visits to the featured articles of the samples under study. The boxplot shows visits to articles that were awarded the featured status on October, and it tracks visits to those articles the previous and following month (September and November). In the boxplots, the main box shows the bulk of data (those values between the 25 and 75 percentile), and the median is highlighted with a line inside the box. Outliers (values with very extreme values) are marked with circles outside the box. For instance, if we focus on the case of the English Wikipedia, at a first glance, it seems that level of visits during October was higher than it was during the months of September and November. However, in September and November, the level of visits was similar.

Interestingly, featured articles get more visits only for the case of the English Wikipedia. For the rest of samples, the differences in the median are very small. Even more interestingly, for the case of the Polish Wikipedia, featured articles get more visits the following month *after* being marked as featured. We do not explore this issue in this paper, although it clearly deserves further and deeper research.

5 FURTHER WORK

The identification and quantization of the requested namespaces and the different types of actions may serve to provide an exhaustive understanding of the way in which the users are interacting with the Wikipedia.

Finally, obtaining the searched topic when the request submitted by the user implies a research operation can help to analyse the use of the Wikipedia as a searching engine or to evaluate the impact of the searches on the most visited contents.

6 CONCLUSIONS

Our results indicate that we can observe that increased attention only in the English Wikipedia, while in the rest of the top six Wikipedias the featured status does not affect the level of visits, neither in the month when the status is awarded, nor in the previous or following months.

We have found an exception to this, though. The case of the Polish Wikipedia is interesting, because

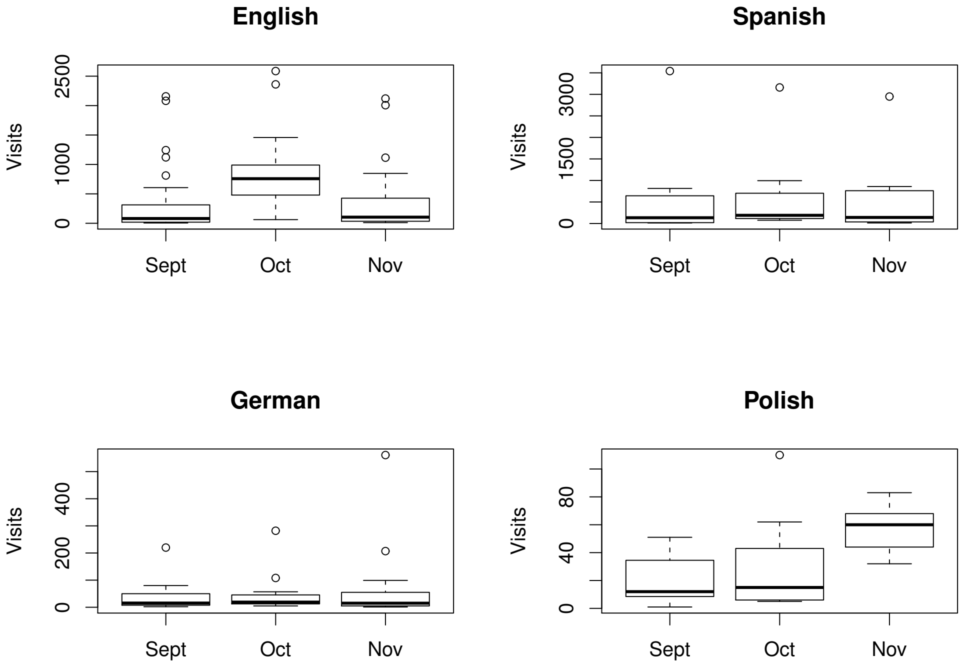

Figure 3: Boxplot of the visits to featured articles for the samples under study.

featured articles gain attention in the month right after the featured status was awarded, this is to say, the Polish Wikipedia's response to the featured *stimulus* is much slower than in the case of the English one.

REFERENCES

[Javanmardi et al., 2009] Javanmardi, S., Ganjisaffar, Y., Lopes, C., and Baldi, P. (2009). User contribution and trust in wikipedia. In *Collaborative Computing: Networking, Applications and Worksharing, 2009 CollaborateCom 2009. 5th International Conference on*, pages 1 –6.

[Lindsey, 2010] Lindsey, D. (2010). Evaluating quality control of Wikipedia's feature articles. *First Monday*, 15(4).

[Olleros, 2008] Olleros, F. (2008). Learning to trust the crowd: Some lessons from wikipedia. In *e-Technologies, 2008 International MCETECH Conference on*, pages 212 –216.

[Priedhorsky et al., 2007] 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 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 et al., 2009] Suh, B., Convertino, G., Chi, E. H., and Pirolli, 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.

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