Quantitative Analysis and Characterization of Wikipedia Requests

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

ABSTRACT

Our poster describes the quantitative analysis carried out to study the use of the Wikipedia system by its users 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. By filtering and classifying a large sample of the requests directed to the Wikimedia systems over 7 days we have been able to identify important information such us the targeted namespaces, the visited resources or the requested actions. The results found include the identification of weekly and daily patterns, and several correlations between different 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. DESCRIPTION AND MAIN GOALS

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]. Nowadays, the Wikipedia contains approximately 8 millions of articles distributed in 250 different language editions. All the Wikipedia editions, as well as others projects 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] and [2]). In particular, we have found just one report[3] involving the specific topics of Wikipedia traffic or Wikipedia patterns of use.

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.

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.

2. METHODOLOGY OF THE STUDY

The study is based on the analysis 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. 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 selected a period covering exactly one week, from 00:00:00 April 7th 2008 (Monday) to 23:59:59 April 13th 2008 (Sunday). This resulted in more than 175 million of Squid log lines which were parsed by an ad-hoc

WikiSym '08, September 8-10, Porto, Portugal
Copyright 2008 ACM 978-1-60558-128-3/08/09 ...$5.00.

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

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 twenty most visited editions (according to the number of requests directed to them) were considered.

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.

For this work, we have considered requests corresponding to articles in the main and discussion namespaces, as well as those corresponding to the user and user discussion namespaces.

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.

3. QUANTITATIVE RESULTS

More than 175 million of lines from the Squid log files corresponding to the analysis period were processed as part of this work in order to establish time and kind-of-use patterns such us described below.

3.1 Time patterns

We have traced the distribution of the requests across the days of the week resulting in Figure 1. As we can observe the number of accesses remains somewhat similar during the first five days -which correspond to the working days- ,but is lower during the weekend days (times are CET, and therefore “weekend” includes, e.g., for time zones in America, the last hours of Friday).

3.2 Use patterns

With respect to the kind of contents requested by the users, we have first classified URLs according to the namespace to which they are directed. Figure 2 shows how the number of requests directed to the analyzed Wikipedias is correlated with the number of requests directed to articles in the main name namespaces in the same Wikipedia projects.

When talking about user requested actions, the most frequently requested action is the edition with a remarkable rate of the history reviews. We remark the low percentage (less than the 7% of all the Wikipedia accesses) of requests resulting in save operations. This means that the ratio edition/save, which illustrates how active and collaborative users of a particular Wikipedia are, is remarkably poor.

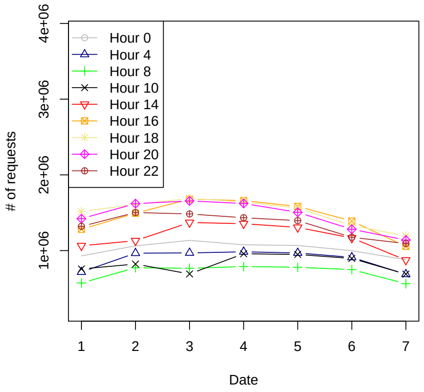

Figure 1: Hourly distribution of the requests made to the Wikipedia over the considered 7-days period.

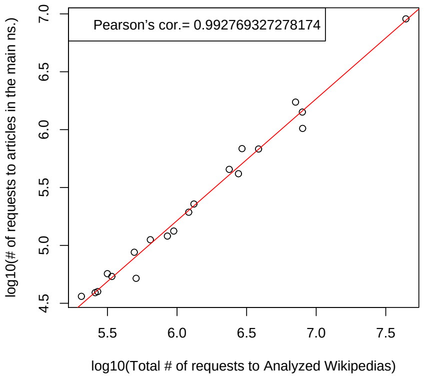

Figure 2: Total number of requests in analyzed Wikipedias against number of requests directed to articles in the main namespace.

4. CONCLUSSIONS

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.

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

[3] G. Urdaneta, P. Guillaume, and M. Van Steen. Wikipedia workload analysis, September 2007.