Most Popular Contents Requested by Users in Different Wikipedia Editions

Antonio J. Reinoso[1], Juan Ortega-Valiente[2], Rocío Muñoz-Mansilla[3] and Carlos León[2]
[1]*LibreSoft Research Group, URJC, C/ Camino del Molino, Fuenlabrada, Spain*
[2]*Department of ICT Engineering, UAX, Avda. de la Univerisdad, 1, Vva. de la Cañada, Spain*
[3]*Department of Computer Science and Automation, UNED, C/ Juan del Rosal, 16, Madrid, Spain*

Keywords: Wikipedia, Use Patterns, Traffic Characterization, Content Categorization.

Abstract: This paper aims to analyze how the most requested and contributed contents in Wikipedia may significantly vary depending on the considered edition. The on-line Encyclopedia has become a prolific research topic, mainly in aspects related to the assessment of its contents and in its evolution forecasting. However, very little effort has been devoted to deal with the kind of use given to Wikipedia by its visitors, either occasional or subscribers. Thus, our work aims to explore the utilization made of Wikipedia through a classification of the most requested and contributed contents in some of its editions. This way, we will be in position of determining which type of contents attracts the highest numbers of visits and contributions in these editions and which can be a good indicator of the use given to them by their respective community of users. Apart from the subsequent comparison purposes, such examination may reveal interesting topics such as the transmission of tendencies over the different Wikipedia editions, as well as particular user patterns exhibited by the corresponding communities of users.

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.

Wikipedia is organized in about 285[1] editions, each corresponding to a different language. All these editions add up more than 22 million articles[2], which correspond to encyclopedic entries about particular subjects, events or people. Wikipedia articles not only address topics from academic disciplines, such as scientific or humanistic subjects, but also from areas related to music, sports, current events and so forth. Focusing on the audience, the overall set of Wikipedia editions attracts approximately 15,000 million visits a month[3]. This fact can be seen as an absolute argument for the popularity gained by Wikipedia and contributes to reinforce its massive acceptance by the Internet community.

As a result of this relevance, Wikipedia has turned into a subject of increasing interest for researchers [4]. This way, quantitative examinations about its articles, authors, visits and contributions have been performed in different studies such as (Ortega et al., 2007) and (Tony and Riedl, 2009). The quality and reliability of the offered information has propitiated a prolific research field where several techniques and approaches have been conducted: (Korfiatis et al., 2006), (Giles, 2005) and (Chesney, 2006). In addition, Wikipedia's growth tendencies and general evolution have also been largely addressed in studies such as (Capocci et al., 2006) and (Suh et al., 2009). Several other works have focused on particular aspects, such as motivation: (Kuznetsov, 2006) and (Nov, 2007), con-

[1]http://meta.wikimedia.org/wiki/List_of_Wikipedias (Retrieved on 6 June 2012)

[2]http://meta.wikimedia.org/wiki/List_of_Wikipedias#-Grand_Total (Retrieved on 6 June 2012)

[3]http://stats.wikimedia.org/EN/Sitemap.htm (Retrieved on 6 June 2012)

[4]http://en.wikipedia.org/wiki/Wikipedia:Academic_-studies_of_Wikipedia (Retrieved on 6 June 2012)

Reinoso A., Ortega-Valiente J., Muñoz-Mansilla R. and León C..
Most Popular Contents Requested by Users in Different Wikipedia Editions.
DOI: 10.5220/0004173104150418
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 415-418
ISBN: 978-989-8565-30-3

sensus: (Kittur et al., 2007), (Suh et al., 2007) and (Viégas et al., 2007) or vandalism: (Priedhorsky et al., 2007). By contrast, few studies (Urdaneta et al., 2007), (Reinoso et al., 2010) or (Reinoso, 2011) have been devoted to analyze the manner in which users interact and make use of Wikipedia.

Therefore, our main objective is the classification and categorization of the most solicited content of Wikipedia, as we consider that this is directly related to the use given by users to the Encyclopedia. Unlike many other previous studies, which are fundamentally based on surveys conducted for specific populations, such as (Konieczny, 2007) or (Willinsky, 2007), our methodological approach is based on the analysis of a sample of the requests sent to Wikipedia by users. With this, we consider that the focus of our analysis is significantly widened: users, Wikipedia editions, temporal period...

The rest of the paper is organized as follows. Next section presents the most relevant aspects of the methodology conducted to develop our analysis. After this, some important results are adequately presented. Finally the section dedicated to present our conclusions and ideas for further work finishes the paper.

2 METHODOLOGY

The Wikimedia Foundation has deployed a layer of special Squid servers to deal with the incoming traffic directed to its several wiki-based projects. Squids work as reverse proxy servers, and perform web caching to avoid the operation of both web and database servers placed behind them. When requested contents can be found in their caches, users receive their response directly from the Squids. Otherwise, Squids ask web servers for the solicited resource and, once obtained, they send it to the user. Independently of the way in which contents sent to users are obtained, Squid servers write down in a log line data related to each particular served request.

The analysis presented here is based on a sample of the log lines stored by Squid servers. All the Squid servers arranged by the Wikimedia Foundation pack and send their log lines to a central aggregator host. In this system, lines are managed by special log processors that can write them to a given destination or pipe them to other processes. In both cases, a sampling factor is configured to determine the number of lines to be processed. Our sample corresponds to the whole 2009 and a sampling factor of the 1% was used to extract it. As a result, we managed to receive one in every hundred requests composing the traffic to the several projects maintained by the Wikimedia Foundation. To summarize, more than 14,000 million log lines have been analyzed as a part of this work.

Receiving log lines from a centralized system is specially relevant for our analysis as it means that our sample is made up of lines from all the Squids deployed by the Wikimedia Foundation as a part of its Content Delivery Network (CDN), which is composed, at a glance, by two large groups of Squid clusters (located, respectively, in USA and in the Netherlands) to whom requests are directed using geographical DNS balancing policies. This guarantees the heterogeneity of our data feed and prevents our results from localized effects or trends due to sociocultural particularities that may arise if we only examined log lines from particular Squids or groups o them.

Once the log lines have been received in our systems, they become ready to be analyzed by the tool developed for this purpose: The *WikiSquilter project*[5]. The analysis consists in a characterization performed in a three-step process: parsing, filtering and storage. Firstly, log lines are parsed to extract relevant informational elements from the users' requests. Secondly, these informational elements are filtered to determine if the corresponding requests fits the directives of the analysis. Finally, information fields from requests considered of interest are normalized and stored in a database for statistical examinations.

Important information concerning users' requests, like their date or if they led to a write operation in the database, can be directly obtained from the log lines fields. Nevertheless, most of the data needed for our analysis is embedded in the URL constituting each request. Therefore, URLs have to be parsed in order to extract their relevant informational elements:

1. The Wikimedia Foundation project 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 title of every requested article.

This information allow us to isolate the requests directed to the Wikipedia project from all the ones that compose our sample, as we have just focused on targeting specific editions. Specifically, we have considered only the top-ten editions regarding both their number of articles and visits. These editions are the German, English, Spanish, French, Italian, Japanese, Dutch, Polish, Portuguese and Russian ones.

As the articles' titles are available, we obtained the ones corresponding to the 65 most visited and

[5]http://sourceforge.net/projects/squilter (Retrieved on 6 June 2012)

contributed articles in the German, English, Spanish and French Wikipedias during 6 random months from 2009, and assigned them to an specific set of categories based in a previous one described in (Spoerry, 2007). At the moment, this is the only step that has to be performed manually, as there are no categorization systems that can be fed only with the few words forming a title. The list of categories is presented below:

- Entertainment (ENT) and Current Issues (CUR).
- Politics and War (POL) and Geography (GEO).
- Information and Communication Technologies (ICT).
- Science (SCI) and Arts and Humanities (ART).
- Sexuality (SEX).

3 ANALYSIS AND RESULTS

This section presents our most remarkable findings in respect to the categorization of the most visited and contributed articles in the considered Wikipedia editions. Such results can be considered as representative of the type of use made of these editions by their corresponding communities of users. Regarding visits as those requests devoted just to get the information contained in the Wikipedia's articles and that do not entail any kind of contribution nor any other type of action, Table 1 presents the categories of contents most repeatedly demanded in the analyzed editions.

According to this table, we can see how in all the studied Wikipedias, except the Spanish one, the category related to *Entertainment* topics attracts most of the visits. However, in the Spanish edition, *Scientific* articles are the ones that attract most of the users' attention. We estimate that this fact could be related with the high percentage that ICT-related *Information and Communication Technologies* contents grab in the Spanish Wikipedia which is considerably higher than in the other analyzed editions. As the *Entertainment* category corresponds to those topics related to movies, celebrities, video games, music bands, etc., results from Table 1 could suggest that Wikipedia is not considered as a primary source for academic or scientific information by users in the editions where this category is the most popular one. Considering the common range of ages to which these types of contents are directed to, these results could also allow to infer that a large number of Wikipedia visitors in these editions are young people. This fact could be reinforced by the great percentage achieved by articles related to sexual topics in some of the Wikipedia editions with the highest percentages in

Table 1: Categorization of the 65 most visited pages in the German, English, Spanish and French Wikipedias.

Category	DE	EN	ES	FR
MAIN	1,54%	1,54%	1,54%	1,54%
CUR	9,23%	17,85%	5,23%	11,08%
GEO	24,62%	7,69%	13,23%	21,85%
ICT	7,08%	5,23%	**12,31%**	6,15%
ENT	**31,08%**	**44,92%**	16,00%	**27,69%**
POL	9,85%	8,92%	5,23%	6,77%
SCI	5,54%	3,38%	**24,00%**	4,31%
ART	4,31%	0,92%	20,92%	13,85%
SEX	6,77%	8,92%	0,31%	0,00%
UND.	0,00%	0,62%	1,23%	1,54%

Entertainment-related contents. Unfortunately, it is not possible to go beyond these mere speculations with the data currently available to us, as Wikimedia Foundation's strong policies about individual's privacy and confidentiality rights do not allow us to get any demographical information about Wikipedia visitors. On the other hand, Table 2 presents the categorization resulting from classifying the most contributed articles in the same Wikipedias. As it can be shown, articles related to the *Entertainment* category gather most of contributions in the English and Spanish Wikipedias. In the English edition, articles related to current events rank in the second position, whereas in the Spanish edition the same position is occupied by articles corresponding to geographical topics. In the German edition, articles devoted to current events receive the highest number of contributions whereas *Humanistic* articles are in the second position. Finally, in the French Wikipedia articles related to *Humanities* are the most contributed while those corresponding to *Entertainment* are in the second position.

It is common to assume that *Entertainment* and *Current* topics should be the ones most easy to contribute to as they do not require users to deal with academic subjects nor a previous education, training or context on a given matter. However, it is really interesting to note how, for example, though scientific articles are the most popular in the Spanish edition of Wikipedia, is in the German one when they achieve the highest percentage of contributions. This means that users of the Spanish Wikipedia do consume a lot of scientific information but the contributions, in the opposite, do not target at all the same subjects.

As we can see, there are significant differences not only in the topics that receives most of users' attention and contributions in the analyzed Wikipedia editions, but also in the topics most visited and edited for each edition individually considered. This gives us an idea of the different types of use given to the Encyclopedia in the different communities of users and the different patterns of use depending of the considered kind of request even inside a particular edition.

Table 2: Categorization of the 65 most contributed pages in the German, English, Spanish and French Wikipedias.

Category	DE	EN	ES	FR
MAIN	0,00%	0,00%	0,00%	0,00%
CUR	**19,69%**	**25,23%**	5,23%	9,23%
GEO	15,38%	9,85%	**17,54%**	23,69%
ICT	7,69%	2,15%	1,85%	0,92%
ENT	14,77%	**36,31%**	**46,46%**	**25,23%**
POL	12,62%	9,54%	6,46%	7,38%
SCI	7,38%	1,54%	7,08%	4,62%
ART	**17,23%**	14,46%	13,85%	**27,69%**
SEX	0,31%	0,00%	0,00%	0,31%
UND.	4,92%	0,92%	1,54%	0,92%

4 CONCLUSIONS AND FUTURE WORK

As our analysis revealed, there are considerable differences among the types of contents most repeatedly visited and contributed in the different editions of Wikipedia. If we focus on particular editions, we can, even, found significant differences amongst the topics that grab users' attention and those that receive the most of the contributions. This can be regarded as different patterns of use characterizing the utilization made of the Encyclopedia by the different communities of users. Such an analysis of behavioral features would benefit from the inclusion of other aspects such as sociological or sociocultural considerations.

In addition, we could extend our analysis with the development of an automatic categorization system capable of performing a wide-scope classification. Nowadays, some of our best efforts are directed towards this target. In addition, topics involved in other types of requests, such as history reviews or searches, could be also categorized to obtain a more complex profile of requested contents. For sure, some form of users related information, though hashed or anonymized, will contribute to define more accurate patterns of use as well as visitors/contributors profiles.

REFERENCES

Capocci, A., Servedio, V. D. P., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., and Caldarelli, G. (2006). Preferential attachment in the growth of social networks: the case of wikipedia.

Chesney, T. (2006). An empirical examination of wikipedia's credibility. *First Monday*, 11(11).

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.

Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. (2007). He says, she says: conflict and coordination in wikipedia. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462, New York, NY, USA. ACM Press.

Konieczny, P. (2007). Wikis and wikipedia as a teaching tool. *International Journal of Instructional Technology & Distance Learning*, 1.

Korfiatis, Nikolaos, Poulos, Marios, Bokos, and George (2006). Evaluating authoritative sources using social networks: an insight from wikipedia. *Online Information Review*, 30(3):252–262.

Kuznetsov, S. (2006). Motivations of contributors to wikipedia. *SIGCAS Comput. Soc.*, 36(2).

Nov, O. (2007). What motivates wikipedians? *Commun. ACM*, 50(11):60–64.

Ortega, F., Gonzalez-Barahona, J. M., and Robles, G. (2007). 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.

Priedhorsky, R., Chen, J., Shyong, Panciera, K., Terveen, L., and John (2007). Creating, destroying, and restoring value in wikipedia. *MISSING*.

Reinoso, A. J. (2011). *Temporal and behavioral patterns in the use of Wikipedia*. PhD thesis, Universidad Rey Juan Carlos. http://gsyc.es/ ajreinoso/phdthesis.

Reinoso, A. J., Ortega, F., Gonzalez-Barahona, J. M., and Herraiz, I. (2010). A statistical approach to the impact of featured articles in wikipedia. In *International Conference on Knowledge Engineering and Ontology Development*, Valencia, Spain.

Spoerry, A. (2007). What is popular in wikipedia and why? *First Monday*.

Suh, B., Chi, E. H., Pendleton, B. A., and Kittur, A. (2007). Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 163–170. IEEE.

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, 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.

Urdaneta, G., Pierre, G., and van Steen, M. (2007). A decentralized wiki enginge for collaborative wikipedia hosting. In *Proceedings of the 3rd International Conference on Web Information Systems and Technologies*, pages 156–163.

Viégas, F. B., Wattenberg, M., Kriss, J., and van Ham, F. (2007). Talk before you type: Coordination in wikipedia. In *MISSING*, pages 78–78.

Willinsky, J. (2007). What open access research can do for wikipedia. *First Monday*, 12(3).