THE TOP-TEN WIKIPEDIAS
A Quantitative Analysis Using WikiXRay

Felipe Ortega
Grupo de Sistemas y Comunicaciones, Universidad Rey Juan Carlos, Tulipan s/n 28933, Mostoles, Spain

Jesus M. Gonzalez-Barahona
Grupo de Sistemas y Comunicaciones, Universidad Rey Juan Carlos, Tulipan s/n 28933, Mostoles, Spain

Gregorio Robles
Grupo de Sistemas y Comunicaciones, Universidad Rey Juan Carlos, Tulipan s/n 28933, Mostoles, Spain

Keywords: Wikipedia, quantitative analysis, growth metrics, collaborative development.

Abstract: In a few years, Wikipedia has become one of the information systems with more public (both producers and consumers) of the Internet. Its system and information architecture is relatively simple, but has proven to be capable of supporting the largest and more diverse community of collaborative authorship worldwide.
In this paper, we analyze in detail this community, and the contents it is producing. Using a quantitative methodology based on the analysis of the public Wikipedia databases, we describe the main characteristics of the 10 largest language editions, and the authors that work in them. The methodology (which is almost completely automated) is generic enough to be used on the rest of the editions, providing a convenient framework to develop a complete quantitative analysis of the Wikipedia.
Among other parameters, we study the evolution of the number of contributions and articles, their size, and the differences in contributions by different authors, inferring some relationships between contribution patterns and content. These relationships reflect (and in part, explain) the evolution of the different language editions so far, as well as their future trends.

1 INTRODUCTION

Wikipedia is one of the most important projects producing collaborative intellectual work in the last years, and has gained the attention of millions of users worldwide. It is also one of the most popular sites on the Internet (for instance, being ranked by Alexa as the 11th most visited website, with over 50 million requests per day)[1].

Three reasons are usually mentioned to explain the success of Wikipedia. The first one is that its articles and contents are based on the contribution of anyone willing to improve them, with little to no restrictions. Many people would argue that this model could not produce good quality compared to peer-review model generally found in scientific publications. Nevertheless, an article published in Nature (Gigles, 2005) showed that the accuracy of Wikipedia is very close to other *traditional* printed encyclopedias such as Britannica. Therefore, if it were possible to create accurate articles with this open contribution model, it could be probably considered as a new method for collecting human knowledge, with an unparalleled breadth and detail.

The second reason is the ease of use of Wikipedia. Its contents are collected, presented and managed mainly with MediaWiki[2], a libre software[3] developed and maintained by the own project. This software offers simple-to-use and intuitive tools for editing articles, adding figures and multimedia content, and also for article reviews and discussions.

The third advantage of Wikipedia is that all textual

[1] Information extracted from http://www.alexa.com/search?q=wikipedia.org on March 23rd, 2007.

[2] http://www.mediawiki.org

[3] Through this paper, we will use the term libre software to refer both to free software and open source software (according to the respective definitions by the Free Software Foundation and the Open Source Initiative).

Ortega F., M. Gonzalez-Barahona J. and Robles G. (2007).
THE TOP-TEN WIKIPEDIAS - A Quantitative Analysis Using WikiXRay.
In *Proceedings of the Second International Conference on Software and Data Technologies - Volume ISDM/WsEHST/DC*, pages 46-53
DOI: 10.5220/0001330100460053

contents are licensed using the GNU Free Documentation License (GNU FDL). This makes the content freely available to all users, and allows reprints by any third parties as long as they make them available under the same terms. Other contents, such as photographs and multimedia are subject to specific copyright notices, most of them sharing the same philosophy and principles of the GNU FDL.

However, although its success in sharing knowledge, Wikipedia may face some serious challenges in the near future. The most disturbing is the rapid growth in system requirements that has to be dealt with, mainly due to Wikipedia's enormous size. The English version has already surpassed the 1.5 million articles mark (1,697,653 articles as of March 21st 2007). The main consequence of this growth is that Wikipedia is beginning to consider how to expand its system facilities in order to not become a victim of its own success.

To evaluate to what extent Wikipedia is growing, and what scenarios the project will likely face in the future, a detailed quantitative analysis has to be designed and performed. This analysis would make it possible to know the evolution of the most important parameters of the project, and the construction and validation of growth models which could be used to infer those scenarios. In this paper, we propose a methodology for performing such kind of quantitative analysis. The growth of the whole project is affected by four different factors:

- *System infrastructure*: Currently, most of the traffic served by Wikipedia comes from a cluster in Florida, maintained by the Wikimedia Foundation. In the past months, several mirror projects have been set up in Europe (France) and Asia (Yahoo! cluster in Singapore). Many other supporters maintain minor mirror sites all over the world. However, the size of Wikipedia seems to grow much faster than the project facilities, with the risk of overloading the servers and consequently producing a slowdown in the service.
- *Software evolution*: MediaWiki is the core software platform of the Wikipedia Project. This tool, essentially developed in the PHP programming language, is responsible for retrieving the contents (text, graphics, multimedia...) from the database for a certain language, and delivering them to the Apache web servers, to satisfy the requests made by users. A very active community of developers supports the evolution of this software package, adding the functionalities users ask for and boosting the performance.
- *Evolution of articles and contents*: Articles are the core of Wikipedia. There is one article for each different topic, and topics are selected through consensus among the wishes of users. Authors can also discuss article contents through special talk pages, thus reaching consensus about what should and should not be included in them. So far, the content of the articles may include text, graphics, photographs, math formulas and multimedia. They reflect the authors' interests and level of contribution (some languages gather more articles than others), so this factor is in close relationship with the last one.
- *Changes and contributions by the community*: the expansion of Wikipedia is also affected by the contributions from editors and developers. Due to its collaborative nature, Wikipedia strongly depends on the work of volunteers to maintain its current rate of growth. If, for some reason, editors change their current behavior, or developers begin to decrease their rate of software contributions, this will definitely affect Wikipedia's future possibilities.

In this paper, we analyze Wikipedia focusing on the last two factors. We concentrate our efforts in gaining knowledge about the Wikipedia community of authors, in the ten most important language versions of the encyclopedia, and the evolution of the articles we find in each of them. The selection of the ten most important languages has been done regarding the total number of articles.

2 BACKGROUND: PREVIOUS RESEARCH ON COLLABORATIVE PROJECTS

Although the process of collaborative content creation is relatively new, collaborative patterns have already been analyzed thoroughly in other technical domains. Libre software is a very good example of those collaborative environments. Several useful conclusions can be learned from a careful examination of their functional features. Wikipedia is, in some sense, a *libre contents* project. Its articles are subject to the GNU FDL, reflecting much of the same philosophy that we find in libre software. It should be interesting to check how much these two worlds show similar behaviors.

For example, a very popular concept introduced by Raymond (Raymond, 1998) for the libre software development is the *bazaar*. Libre software projects tend to a development model that is similar to oriental bazaars, with spontaneous exchanges and contributions not leaded by a central authority, and without a

mandatory scheduling. These methods can be seen as opposite to typical software development processes, as these are more similar to how medieval cathedrals were built, with very tight and structured roles and duties, and centralized scheduling.

But it was not until almost the year 2000 when the research community realized that lots of publicly available data about libre software could be obtained and analyzed. Some research works, including (Ghosh and Prakash, 2000) and (Koch and Schneider, 2002) showed that a small group of developers were the authors of a large amount of the available code. Mockus et al. (Mockus et al., 2002) performed a research work about the composition of the developers communities of large libre software projects. They verified that a small group of developers (labeled as the *core group*) was in charge of the majority of relevant tasks. A second group, composed by developers who contribute frequently, is around one order of magnitude larger than the core group, while a third one, this one of occasional contributors, is about another order of magnitude larger.

Other interesting research works include (Godfrey and Tu, 2000) about the growth in size over time of the Linux kernel. Godfrey et al. showed that Linux grew following a super-linear model, apparently in contradiction to one of the eight *laws* of software evolution (Lehman et al., 1997). Although not yet confirmed, this may be indicative of a superior growth for open collaborative development environments than with closed industrial settings commonly used. Other research works have focused their attention on the study of Linux distributions, where hundreds to thousands of libre software programs are integrated and shipped. Especially the case of Debian (Gonzalez-Barahona et al., 2001; Gonzalez-Barahona et al., 2004; Amor et al., 2005a; Amor et al., 2005b) is very interesting in this regard as it is a distribution built exclusively by volunteers.

A methodological approach of how to retrieve public data from software repositories and the various ways that these data can be analyzed, especially from the point of view of software maintenance and evolution, can be fond in Gregorio Robles' dissertation (Robles, 2006). This work puts special attention to developer-related (or social) aspects as these give valuable information about the community that is developing a software.

Specifically on the Wikipedia, we can find also some previous studies. (Buriol et al., 2006) quantifies the growth of Wikipedia as a graph. The authors find many similarities among several language versions of Wikipedia, as well as with the structure of the World Wide Web. This should be no surprise, because in some way, wikis are simply another flavor of websites where content may be linked from other contents (by using HTML hyperlinks). Jakob Voss (Voss, 2005) introduced some interesting preliminary results about the evolution of contents and authors, mainly focusing on the German version of the Wikipedia: the number of distinct authors per article follows a power-law while the number of distinct articles per author follows Lotka's Law. Buriol et al. (Buriol et al., 2006) showed that growth in number of articles and users were consistent with Voss' results, but this time in the English version.

Finally, Viegas et al. (Viegas et al., 2004) found an alternative approach for studying contribution patterns to Wikipedia articles. They have developed a software tool, History Flow, that can navigate through the complete history of an article. This way, it is possible to identify periods of intense growth in the content of articles, acts of vandalism and other interesting patterns in users contributions.

3 METHODOLOGY

In this section, we present the methodology for a quantitative analysis of different language versions in Wikipedia. Firstly, we introduce some of the most relevant features of the database Wikipedia uses to store all its contents and edit information. Then, we briefly present the automatic system that the Wikimedia Foundation uses to create database dumps for all of its projects, and specifically for Wikipedia. Finally, we describe WikiXRay[4], our own tool developed for generating quantitative analysis of different language versions in Wikipedia automatically.

3.1 Wikipedia Database Layout

The MediaWiki software is currently strongly tied to the MySQL database software. Many functions and data formats are not compatible with other database engines. The logical model of the database that stores Wikipedia contents has suffered a deep transformation since version 1.5 of the MediaWiki software. The most up-to-date database schema always resides in the *tables.sql* file in the Subversion repository.

The tables of the database logical model that are relevant to our purposes are:

- *Page*: One of the core tables of the database. In this table, each page is identified by its title, and provides some additional metadata about it. The

[4]http://meta.wikimedia.org/wiki/WikiXRay

name of each page refers to the namespace to which it belongs to.

- *Revision*: Every time a user edits a page, a new row is created in this table. This row includes the title of the page, a brief textual summary of the change performed, the user name of the article editor (or its IP address the case of an unregistered user) and a timestamp. The current timestamp support is somewhat basic as it is implemented using plain strings.
- *Text*: The text for every article revision is stored in this table. The text may be stored in plain UTF-8 compressed with gzip or in a specialized PHP object.

Database dumps do not only contain the articles, but other relevant pages that are used by the user community of that language on a daily basis. To classify these pages, MediaWiki groups pages in logical domains known as namespaces. A tag included in the page title indicates the namespace a page belongs to. Articles are grouped in the `Main` namespace. Other relevant namespaces are `User` for the homepage of every registered user, `Meta` for pages with information about the project itself, and `Talk`, `User_talk` and `Meta_talk` for discussion pages related to articles, users and the project respectively. Most of our research work is focused on articles in the `Main` namespace.

3.2 Database Dumps

There are database dumps available for all Wikipedia versions through the web[5]. Some major improvements have recently been included in the database dumps administration, the most relevant the automation of the whole dump process, including real-time information about the current state of each dump. Other new features include the automatic creation of HTML copies of every article stored, and the upcoming system for creating DVD distributions for different language versions. The tool employed to perform database dumps is the Java-based mwdumper, also available in the Wikipedia SVN repository[6]. This tool creates and recovers database dumps using an XML format. Compression is achieved with bzip2 and SevenZip.

In our study we have retrieved a simplified version of the dumps which provides data only for the page and revision tables of each language. An additional dump with the page table alone had also to be downloaded, because we needed information about the length in bytes of every single page. The simplified dump does not include that information.

3.3 Quantitative Analysis Methodology: Wikixray

The methodology we have conceived to analyze the Wikipedia is composed of following steps: First, we collect the database dumps for the top-ten Wikipedia languages (in number of articles, according to the list publicly available from Wikipedia main page). We have therefore developed a Python tool, called WikiXRay, to process the database dumps, automatically collecting relevant information, and processing this information and proceed to an in-depth statistical analysis.

Quantitative results for each language can be obtained from two different points of view:

- *Community of authors*: This is the first important parameter that affects the growth of the database. Relevant aspects include studying the total number of editors and contributions, number of contributions for each author over a certain period of time (i.e. contributions per month or per week) and correlating results with Wikipedia's own statistics.
- *Evolution of articles*: We analyze the growth of the database from a different perspective, focusing on the evolution of the size of the articles over time, the distribution of articles sizes in general and how the evolution of articles correlates to the contributions made by users.

4 CASE STUDY: THE TOP-TEN WIKIPEDIA LANGUAGES

As case study for our methodology, we have considered convenient to analyze the database dumps of the top-ten largest language versions of the Wikipedia. At the time of writing, the most popular language corresponds to English, followed in this order by German, French, Polish, Japanese, Dutch, Italian, Portuguese, Swedish, and Spanish. Due to spaces limitations, we will not be able to include in this paper all the results obtained, but will show the most relevant ones[7]. Figure 1 is a graphic that shows the evolution over time of the number of contributions to articles for the top-ten languages. A contribution is considered any edition made by an user to an article. A logarithmic scale in

[5] http://download.wikimedia.org

[6] http://svn.wikimedia.org/

[7] http://meta.wikimedia.org/wiki/WikiXRay offers additional graphic results.

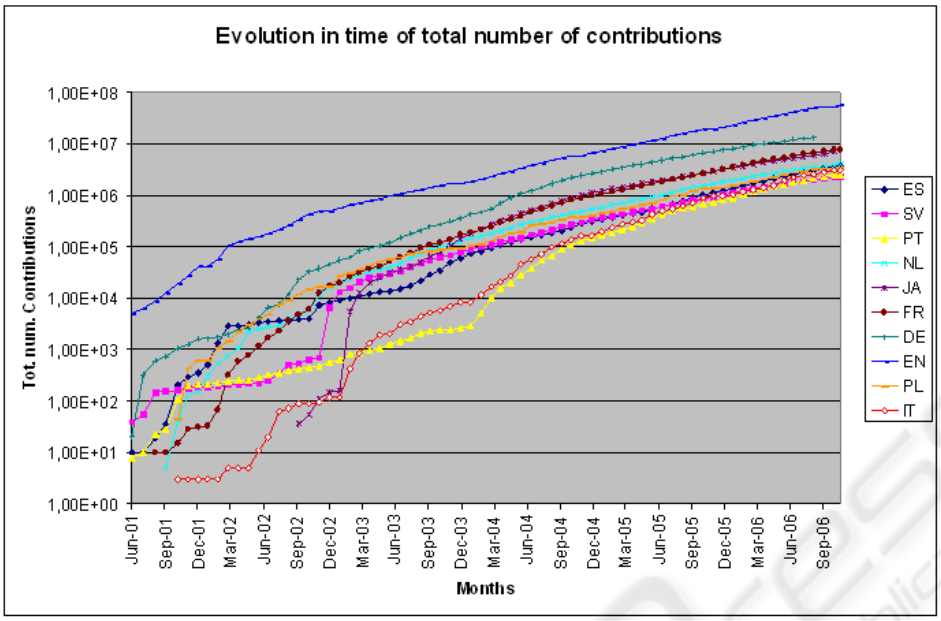

Figure 1: Evolution over time of the total number of contributions.

the vertical axis to plot the graphics has been used, resulting in a clear common behavior at least since December 2004 for all languages.

For some language versions we can establish a strong correlation between relevant past events and abrupt increases of their growth rates for total contributions. For example, the Japanese Wikipedia experimented a quite remarkable growth of two orders of magnitude in its total number of contributions from February to March, 2003. In January 31, 2003, the Japanese online magazine Wired News covered Wikipedia. This has been reported as the first time Wikipedia was covered in the Japanese media[8]. So, we can infer a direct relationship between Wikipedia popularity and the number of contributions it receives.

4.1 The Community of Authors

One of the parameters we are interested in is the level of inequality that can be found for contributions. As already mentioned, previous research on the libre software phenomenon has shown that a relative small number of developers concentrate a large part of the contributions. Analyzing inequality will allow us to see if both phenomenons present similar patterns.

We will measure inequality by means of the Gini coefficient. This coefficient, introduced by Conrado Gini (Gini, 1936) to measure welcome inequality in economics, shows how unequal something is distributed among a group of people. To calculate the Gini coefficient we have first to obtain the Lorenz curve, a graphical representation of the cumulative distribution function of a probability distribution. Perfect distribution among authors is hence given by a 45 degree line. The Gini coefficient is given by the area between the two curves, providing how far the actual distribution is from the perfect equality. Figure 2 presents the Lorenz curve for all the languages under study. All of them present similar behaviors, with aproximately 90% of the users responsible all together for less than 10% of the contributions, (Gini coefficients ranging from 0.9246 in the Japanese version to 0.9665 in the Swedish version). Hence, we can state that as in the case of libre software, we also find a small amount of very active contributors.

4.2 Articles

An important parameter of articles is their size, as it gives the amount of content included in them. We have therefore plotted histograms for article sizes for all languages under study. This way we will be able of inferring different types of articles.

Figure 3 shows the histogram of the article size in the English and Polish Wikipedias. We take the decimal logarithm of the article size in bytes for this representation, which facilitates the identification of patterns. The solid black line plotted over the histograms represents the probability density function of the article size, giving about the same information but with better resolution.

After inspection, we can group articles attending

[8]http://en.wikipedia.org/wiki/Japanese_wikipedia

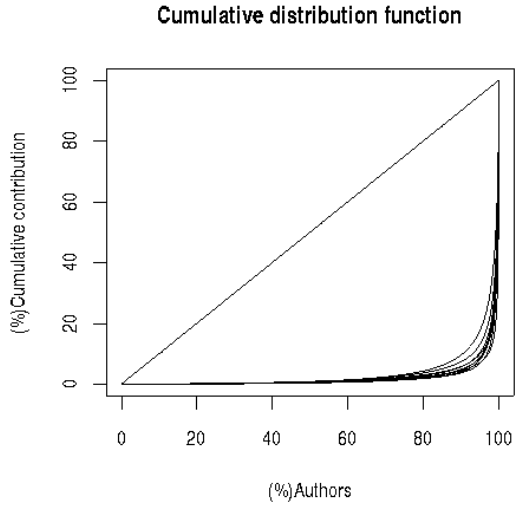

Figure 2: Lorenz curves for contributions of authors to the top-ten languages.

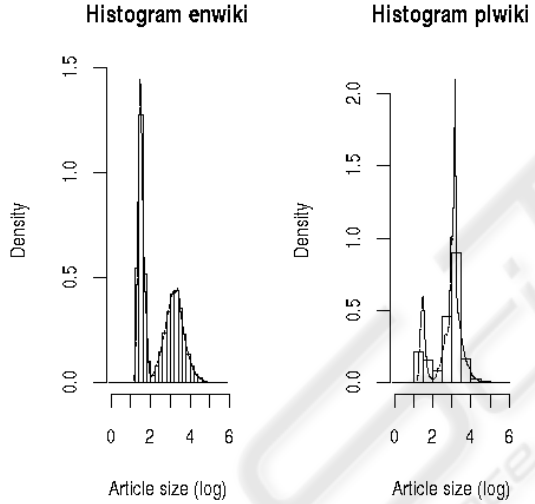

Figure 3: Histogram for sizes of articles in the English (left) and Polish (right) languages.

to their size in two groups:

- *Tiny articles*: The left side of the histograms shows a subpopulation conformed by those articles whose size varies from 10 bytes to 100 bytes. Some of those belong to a special category of articles known as 'stubs' in the Wikipedia jargon. Stubs are templates automatically created when a user requests a new article about some topic not previously covered. This way, the software makes it easier for any upcoming user interested in that topic to further contribute to the article. However, most of the articles in this subpopulation fall into another important category in Wikipedia: 'redirects'. One of the biggest problems for any encyclopedia is how to select an accurate entry name for each article, because many topics present alternative names users can also search for. Redirects are the perfect answer to deal with multiple names for the same article. They are special articles with no content at all, but a link that points to the main article for that topic. So, when users search for alternative names, the find the equivalent page that *redirect* them to the main article for that topic. Redirects also allow contents to be centralized in certain articles, thus saving storing capacity.
- *Standard articles*: On the other hand, we can identify a second subpopulation on the right side of the histograms, corresponding to those articles whose size grows beyond 500 bytes, that is, articles that have a certain amount of content. Further research should be conducted to explain whether the community is more interested in those topics, or those articles have been on-line for a longer period of time, increasing the probability of receiving contributions.

We can extract interesting conclusions from the shape of the density function, as each subgroup of articles exhibits a Gaussian distribution. Its mean can be use to characterize the contributions of each user community to standard articles, and the average redirect size (for the tiny articles). We have calculated the ratio between the normalized mass of the density function for tiny and standard articles for all languages. Results, presented in Table 1, show some communities that are not very interested in creating redirects offering alternative entry names for their articles (for example Polish, Italian and Portuguese Wikipedias), while other ones (for example the English Wikipedia) generate more redirects balancing the probability mass of both subpopulations. Therefore, simply counting the number of different articles of a certain language version does not give us a clear picture about the size or quality of its articles. Some language versions may create a lot of redirects or stubs, while others may concentrate on adding real contents to existing articles. Further research about these results may lead to identify interesting content creation patterns in different communities of authors.

Other factors, such as robots that automatically create a bunch of new stubs from time to time, need to be taken under consideration. For example, the ratio for the English language is noticeable; a closer inspection has thrown as result that this is due to many automatically created articles e.g. with information from the U.S. census. Also noticeable is the case of the Polish Wikipedia. In July 2005 a new

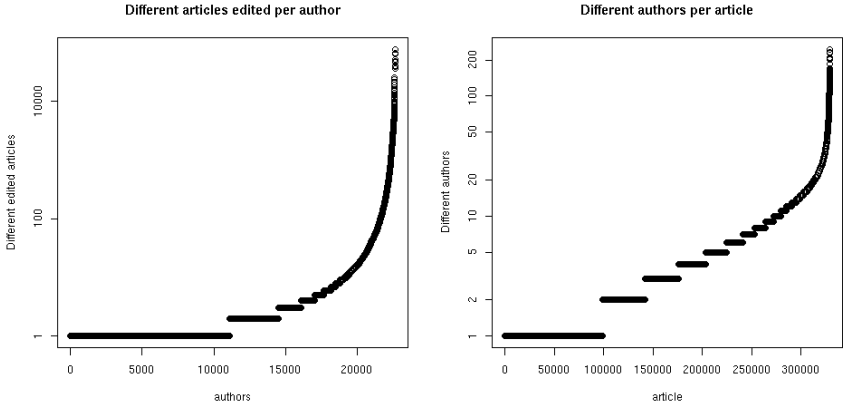

Figure 4: Number of different articles edited per author (top) and number of different authors per article (bottom) for the Dutch Wikipedia.

task was created for the `tsca.bot`, one of the bots of the Polish Wikipedia. It was programmed to automatically upload statistics from official government pages about French, Polish and Italian municipalities. This new feature introduced more than 40,000 new articles in the following months. That allowed the Polish Wikipedia to overtake the Swedish, Italian and Japanese language versions and become the 4th largest Wikipedia by total number of articles[9].

Table 1: Probability mass value for tiny articles and standard articles.

Lang	Tiny Mass	Std Mass	Mass Ratio
English	0.51	0.49	0.96
German	0.38	0.62	1.63
French	0.3	0.7	2.33
Polish	0.18	0.82	4.55
Japanese	0.37	0.63	1.7
Dutch	0.29	0.71	2.44
Italian	0.23	0.77	3.34
Portuguese	0.24	0.76	3.16
Swedish	0.34	0.66	1.94
Spanish	0.33	0.67	2.03

Figure 4 gives the number of articles edited by author (only registered authors are considered) and the number of authors per article for the Dutch version of Wikipedia. The article with the highest number of contributors accounts for over 10,000 registered users; around 50 have been the work of over 5000 authors. On the other hand, we can find authors that have contributed to more than 10,000 articles. Future research should focus on these results and explore if these type of authors concentrate in specific tasks, as for instance correcting errata or adapting the style to meet Wikipedia conventions.

On the other side, more than 250,000 articles of the Dutch version were contributed by less than 10 authors, corroborating that, in general, a majority of authors tend to focus their contributions only on a few articles.

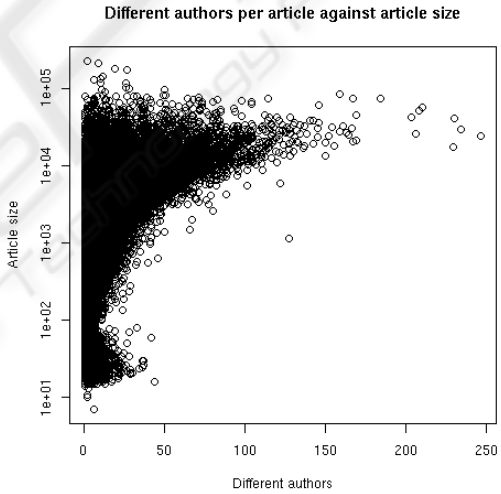

Figure 5: Number of authors against article size (in bytes) for the Dutch Wikipedia.

Finally, in Figure 5 we represent the number of authors per article against the article size in the Dutch Wikipedia. We see that very few articles have been worked on by more than 150 authors. In general, article size correlates positively to the number of authors who have edited it, with smaller sizes for those articles with less contributors and larger sizes for those with wider number of authors. Despite these facts, it is also interesting to notice that the 10 largest articles reflect the work of less than 30 authors, and that the largest one has received contributions from less than 10 authors. So, the number of different authors should not be considered as the unique parameter affecting articles size.

[9]http://en.wikipedia.org/wiki/Polish_Wikipedia

5 CONCLUSIONS AND FUTURE RESEARCH

Some useful conclusions can be extracted from this research. We have shown that the top-ten language versions of Wikipedia present interesting similarities regarding the evolution of the contributions to articles over time, as well as the growth rate in the sum of article sizes. The Gini coefficients found for the studied languages present (as expected) big inequalities in the contributions by authors, with a small percentage being responsible for a large share of the contributions. However, the Gini values found for the languages could help to characterize the underlying author communities.

We have also identified certain patterns that could be used to characterize Wikipedia articles attending to the length (or size) of the articles. Two main subgroups (tiny articles and standard articles) represent the peculiarities of contributions behaviors in each language community. The ratio between them shows the interest of the corresponding communities in linking or opening new topics versus completing and improving existing ones.

Finally, we have found that there is no simple correlation between the number of authors that contribute to a certain article and the total size reached by that article. This leads us to think about additional factors that could affect the production process, including the nature of the topic and the level of popularity of that topic in the author community.

The methodology we have proposed provides an integral quantitative analysis framework for the whole Wikipedia project, a very ambitious goal that we confront for the near future.

REFERENCES

Amor, J. J., Gonzalez-Barahona, J. M., Robles, G., and Herraiz, I. (2005a). Measuring libre software using debian 3.1 (sarge) as a case study: preliminary results. In *Upgrade Magazine*.

Amor, J. J., Robles, G., and Gonzalez-Barahona, J. M. (2005b). Measuring woody: The size of debian 3.0. In *Technical Report. Grupo de Sistemas y Comunicaciones, Universidad Rey Juan Carlos. Madrid, Spain.* Grupo de Sistemas y Comunicaciones, Universidad Rey Juan Carlos. Madrid, Spain.

Buriol, L. S., Castillo, C., Donato, D., and Millozzi, S. (2006). Temporal evolution of the wikigraph. In *Proceedings of the Web Intelligence Conference, Hong Kong*. IEEE CS Press.

Ghosh, R. A. and Prakash, V. V. (2000). The orbiten free software survey. In *First Monday*.

Gigles, J. (2005). Internet encyclopedias go head to head. In *Nature Magazine*.

Gini, C. (1936). On the measure of concentration with especial reference to income and wealth. In *Cowless Comission*.

Godfrey, M. and Tu, Q. (2000). Evolution in open source software: A case study. In *Proceedings of the International Conference on Software Maintenance (pp. 131-142). San Jos, California*.

Gonzalez-Barahona, J. M., Ortuno-Perez, M., de-las Heras-Quiros, P., Gonzalez, J. C., and Olivera, V. M. (2001). Counting potatoes: the size of debian 2.2. In *Upgrade Magazine, II(6) (pp. 60-66)*.

Gonzalez-Barahona, J. M., Robles, G., Ortuno-Perez, M., Rodero-Merino, L., Centeno-Gonzalez, J., Matellan-Olivera, V., Castro-Barbero, E., and de-las Heras-Quiros, P. (2004). *Analyzing the anatomy of GNU/Linux distributions: methodology and case studies (Red Hat and Debian). Free/Open Software Development. Stefan Koch, editor, (pp. 27-58)*. Idea Group Publishing, Hershey, Pennsylvania, USA.

Koch, S. and Schneider, G. (2002). Effort, cooperation and coordination in an open source software project: Gnome. In *Information Systems Journal, 12(1) pp. 27-42*.

Lehman, M. M., Ramil, J. F., and Sandler, U. (1997). Metrics and laws of software evolution the nineties view. In *METRICS 97: Proceedings of the 4th International Symposium on Software Metrics, page 20*.

Mockus, A., Fielding, R. T., and Herbsleb, J. D. (2002). Two case studies of open source software development: Apache and mozilla. In *ACM Transactions on Software Engineering and Methodology, 11(3) (pp. 309-346)*.

Raymond, E. S. (1998). The cathedral and the bazaar. In *First Monday, 3(3)*.

Robles, G. (2006). *Empirical software engineering research on libre software: Data sources, methodologies and results. Doctoral Thesis*. Universidad Rey Juan Carlos, Mostoles, Spain.

Viegas, F. B., Wattengberg, M., and Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems, pp.575-582. Viena, Austria*.

Voss, J. (2005). Measuring wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Infometrics 2005, Stockholm*.