Patterns of Creation and Usage of Wikipedia Content

Andrea Capiluppi
DISC – Brunel University
London, UK
andrea.capiluppi@brunel.ac.uk

Ana Claudia Duarte Pimentel
ACE – University of East London
London, UK
u0914698@uel.ac.uk

Cornelia Boldyreff
ACE – University of East London
London, UK
c.boldyreff@uel.ac.uk

***Abstract*—Wikipedia is the largest online service storing user-generated content. Its pages are open to anyone for addition, deletion and modifications, and the effort of contributors is recorded and can be tracked in time.**

Although potentially the Wikipedia web content could exhibit unbounded growth, it is still not clear whether the effort of developers and the output generated are actually following patterns of continuous growth. It is also not clear how the users access such content, and if recurring patterns of usage are detectable showing how the Wikipedia content typically is viewed by interested readers.

Using the category of Wikipedia as macro-agglomerates, this study reveals that Wikipedia categories face a decreasing growth trend over time, after an initial, exponential phase of development. On the other hand the study demonstrates that the number of views to the pages within the categories follow a linear, unbounded growth.

The link between software usefulness and the need for software maintenance over time has been established by Lehman and other; the link between Wikipedia usage and changes to the content, unlike software, appear to follow a two-phase evolution of production followed by consumption.

I. Introduction

The evolution of Wikipedia as a community generated effort has long generated interest by researchers, in terms of both the motivations underpinning the actual contributors [2], [6], and the sustainability of its evolution [8], [4]. An initial model of Wikipedia growth proposed in 2003 predicted exponential growth of content, but when it became clear in 2007 that content growth was no longer exponential, a revised model of logistic, S-type growth was found to more accurately reflect the actual growth of Wikipedia content [9], [7], [5]. In recent research on user-generated web content [1], Wikipedia pages in the category of ”Software Engineering” were found to follow a similar pattern of evolution to active Open Source Projects on SourceForge; they exhibited a slow growth rate followed by faster growth and finally a decrease in growth. Even with these additional results, it is still unclear how wide-spread this pattern is, and whether a declining “production” of content also corresponds to a decreasing of “consumption” by users.

This study analyses the evolution of a selection of Wikipedia pages and their categories, with two objectives. The first objective is to demonstrate that different categories of pages in Wikipedia show a similar pattern of evolution in both the number of contributors (who spend effort in creating or updating the pages) and the resulting output, in terms of number of edits. The second objective is to show that the level of consumption of the Wikipedia web content, measured via the page views by interested readers, is structurally different from the evolution of effort and edits, and that it also evolves differently.

II. Research Design

This research was drawn around the Goal-Question-Metrics approach, that is, using “metrics” to assess some specified “questions” to achieve an overall “goal”. The main *goal* of this research is to extract recurring patterns of evolution of Wikipedia pages when grouped by categories.

The following *questions* were formulated for the evaluation of the above goal:

1) Do Wikipedia categories generally exhibit the same patterns of *number of contributors* over time?
2) Do Wikipedia categories generally exhibit the same patterns of *number of edits* over time?
3) Do Wikipedia categories generally exhibit exhibit the same patterns of *number of views* over time?

The following *metrics* were used in the assessment of the above questions:

1) The **effort** of contributors was evaluated by counting the number of unique (or *distinct*, in a SQL-like terminology) contributors during a specific interval of time. The chosen granularity of time was based on months (as “effort” in man-months [3]). In particular, we evaluated the number of unique (i.e., “distinct”) contributors per month, and also the *cumulated* number of contributors, evaluated by summing up, for month m, all the previous contributors up to month $m-1$ and the unique new contributors in month m.
2) The **work produced** was evaluated by counting the number of edits to the Wikipedia pages during the same intervals of time (i.e., monthly). Each Wikipedia edit is recorded with a plain-text description that is available to download via a dedicated web-page, as detailed below. In particular, for every month m, we evaluated the number of edits that are performed on each category during m, and also the cumulated

number of edits, summing up the edits of month m with all the edits up to month $m-1$.

3) The number of **views** that is times that each page is accessed by readers was measured monthly to indicate in summary whether the analysed pages and relative categories provide value to the users. As an aggregate, we evaluated (for every month m) both the monthly number of views of all the pages in a category, and the cumulated number of the views for the whole category up to month m.

III. Empirical Approach

In order to restrict the categories to be studied for this research to a single domain, we selected the generic "Arts" domain within Wikipedia, and some of its associated categories: Architecture, Arts, Dance, Design, Fashion, Films, Painting, Photography, Sports and Theatre. Each Wikipedia category lists many pages and subcategories, each containing several hundreds of sub-pages. Since in Wikipedia the sub-pages are labeled also with the main category (apart from their subcategory), all the pages from the categories and subcategories were considered to carry out this study.

Table I summarizes the main characteristics of the studied categories, as per the latest month considered (April 2012). For example, the first row shows that the Architecture category is composed of 1,973 pages, it has had 50,000 unique contributors so far, and its pages were added, changed or edited some 550,000 times. Finally, the last column shows the amount of cumulative views that the pages in the category benefited from since the earliest available date to the latest month considered (330 million views).

Category	Pages	Contributors	Edits	Views
Architecture	1,973	50k	550k	330M
Arts	855	26k	215k	265M
Dance	592	12k	153k	89M
Design	1,477	30k	250k	305M
Fashion	554	31k	287k	367M
Films	1,227	25k	280k	380M
Painting	468	26k	240k	54M
Photography	675	23k	187k	295M
Sports	1,098	20k	206k	133M
Theatre	985	17k	147k	140M

Table I
Characteristics of the chosen categories (as of April 2012)

In order to analyse the Wikipedia pages, two Open Source tools were used:

- *Mediawiki Special:Export Interface*: this interface[1] allows a researcher to extract an XML file containing all the revisions of a page (or a category of pages), for offline analysis.
- *Mediawiki Dumper*: this tool[2] produces a SQL dump from the previously downloaded XML dump. Three tables were studied further: the "revision" table containing all the revision histories with timestamps; the "page" table containing the titles and the IDs of each page; and the "text" table containing the whole text revisions as BLOB objects.

A. Toolchain

An overview of the toolchain is visualized below (Figure 1): the coloured items identify code that had to be written to join the inputs or the outputs of the available tools. After choosing a Wikipedia category at the top of the figure, scripts were generated to screen-scrape not only the pages and the sub-categories contained in the chosen category, but also the sub-pages contained in the sub-categories. With the generated list of pages, a request was issued for every month and every page to the JSON server[3] recording the page views. It should be noted that, for all the categories, the earliest data on views available on JSON is Dec 2007. Such raw data had to be collected per page and aggregated, monthly, at the category level.

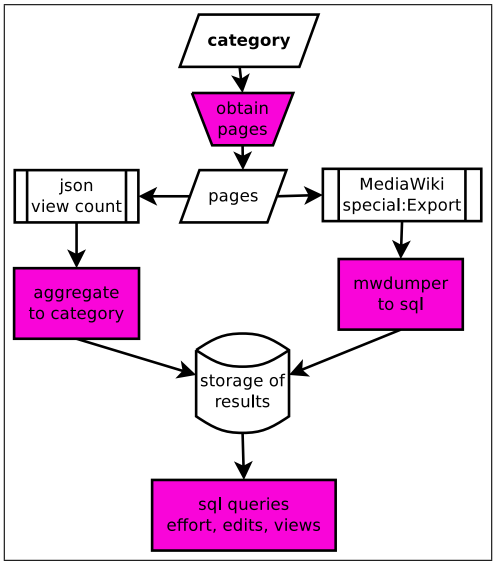

Figure 1. Toolchain used

On the other hand, a POST request was automatically sent to the Special:Export MediaWiki interface, to download the revision histories of the pages and the sub-pages composing the analysed categories. It should also be noted that the MediaWiki interface allows users to download only the first 1,000 revisions for each page[4]. In many cases, the Wikipedia pages have more than 1,000 revisions; for these pages, the latest returned edit and time-stamp were noted, and other POST requests were issued staring from the time-stamp, in

[1] http://www.mediawiki.org/wiki/Manual:Parameters_to_Special:Export

[2] http://svn.wikimedia.org/svnroot/mediawiki/trunk/mwdumper/

[3] http://stats.grok.se/json/

[4] The output revisions are the earliest ones, i.e., starting from revision 1.

order to download all the remaining revisions, in batches of 1,000.

The data of both the views and the edits and contributors was stored in a SQL database and SQL queries formulated to extract the metrics mentioned above.

IV. Results – Edits and Contributors

In this section we outline the results of our experiments, showing the the monthly and cumulative number of edits, and the monthly and cumulative number of unique contributors, for the 10 categories studied. For reasons of space only two categories are reported below, "Dance" and "Fashion" although most of the patterns apply to all the analysed categories.

Figure 2 shows the cumulative number of edits as a dashed line, and the number of monthly edits as a continuous line, for both categories ("Dance" in Figure 2 top graph, and "Fashion" in Figure 2 bottom graph). The monthly number of edits in both reported categories has a very long initial tail, a sharp increase and a peak, after which the trend starts to decline. Similar trends were found in all of the studied categories. On the other hand, the cumulative number of edits seems following an S-shape, logistic curve; after a slow start, a large increase in number of edits has now given way to a slower trend of addition and change of content. For the Wikipedia at large, this result has been shown already [8], but our results show that smaller-level agglomerates share the same pattern.

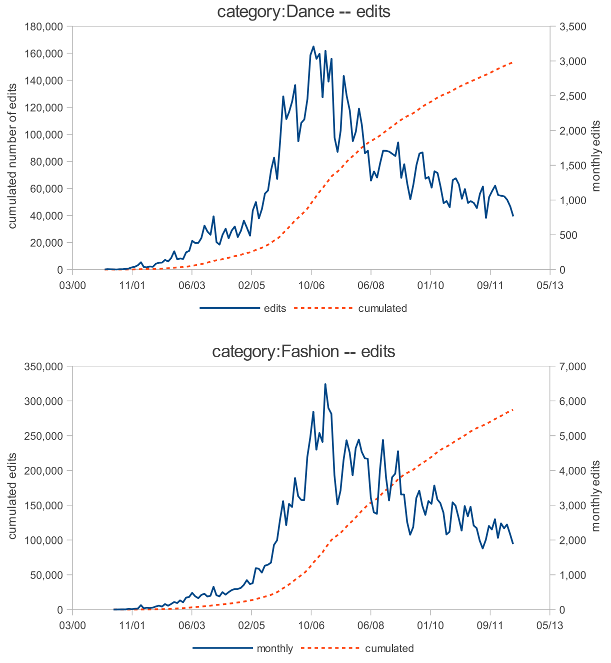

Figure 2. Monthly and cumulated number of edits for two categories

The patterns observed above for the "edits" are also detected when plotting the number of contributors. In Figure 3 the two selected categories are studied in depth again with respect to contributors. A peak can also be seen in the distribution of monthly contributors, which later declines to much smaller values.

When fitting a regression model to these trends, we found that an *exponential* curve (up to the major peak) provides a good fit, but only considering the data up to the peak; in all the categories, and for both the edits and contributors trends, the coefficient R^2 reaches over 90% for the regression lines, as an indicator of the goodness of fit.

The second part of these curves (which normally starts between 2007 and 2008) is instead a descending linear regression, with the R^2 coefficients around $70\% - 80\%$. Whether this second trend is irreversibly declining, and the number of contributors is destined to decline even further, are questions that need to be considered as more data becomes available in the future.

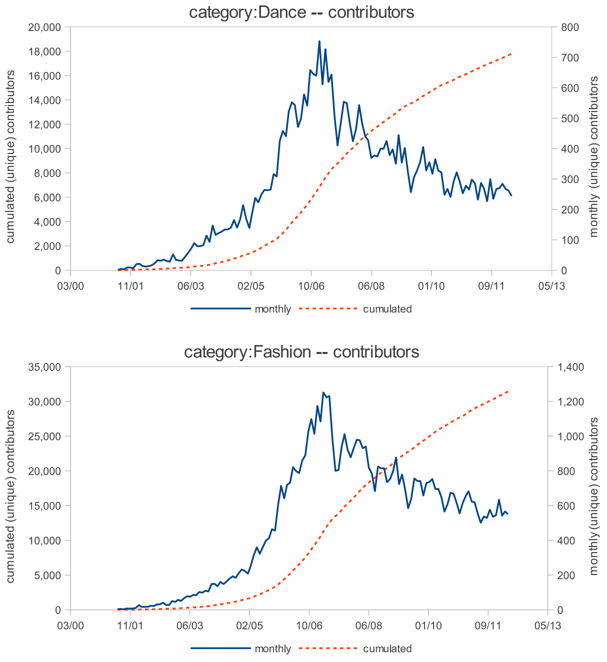

Figure 3. Monthly and cumulated number of contributors

V. Results – Views

The results for the monthly and the cumulated number of views for the two categories described above are displayed in Figure 4. As mentioned above, the available data only starts from Dec 2007 onwards, i.e., around the time of the major peaks in the evolutionary trends of edits and contributors. The plots below show that the number of views in the two categories, and in general for all the 10 categories studied, is increasingly monthly, and it has a very good linear fit at the cumulative level (R^2 over 99% in all the cases).

Summarizing the results on edits, contributors and views, each category shows at the aggregate level a declining phase of contributors' effort and edits, and an increasing trend in terms of views in the same periods. In the next section we discuss this result in the context of production and consumption of user-generated content.

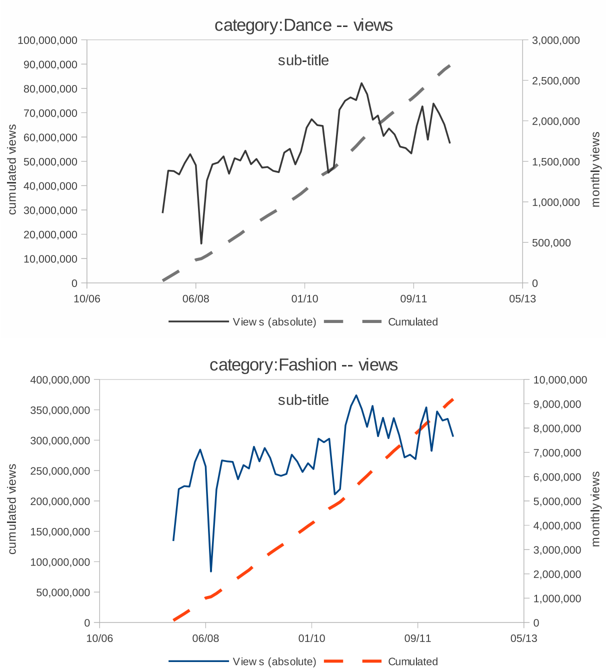

Figure 4. Monthly and cumulated number of page views

VI. Discussion and Implications

Large collections of related Wikipedia pages show similar trends in the number of the overall contributors who provide enhancements to, or create new Wikipedia pages. As also found in other works, the number of contributors, per category, is rapidly decreasing following a logistic curve. This is also mirrored by the number of edits to the pages of the analysed categories; although single pages show a much larger number of edits, and a steady number of contributors, the overall categories see a general decline in the effort provided and the output produced.

In this study we found that the decrease in activity on the Wikipedia pages, as also reported in other works, is only one side of the user-generated content phenomenon; when tracking the number of views, the categories have a different trend from the "consumers" point of view, i.e. the users of such content. After an exponential-driven growth trend of development and refinement (between 2001 and 2007 for most categories), the effort and the work produced in the Wikipedia categories have turned to linear, descending trends. This second descending trend is paired to an ascending linear trend in the number of views, indicating that the pages and categories of Wikipedia have become a reliable source of reference.

Such user-generated content conforms to a two-phase evolution framework: one of production followed by consumption. In the "production" phase, the Wikipedia content is massively generated and optimized; in the "consumption" phase, even if the activity of production declines, the knowledge becomes widely available and accessed by the consumers who can establish, with their increasing views, such knowledge as an valuable and credible source of information.

VII. Conclusion

This paper has presented the results of a quantitative analysis of 10 related categories of Wikipedia pages, developed and evolved online by a large number of contributors. Two types of results have been observed. First, at the aggregated level of categories, it is possible not only to confirm that the Wikipedia categories are experiencing a slow-down in terms of activity and the relative effort; but it is also possible to divide the evolution of such online content into two trunks, a first exponential phase of development, and a second, linearly declining, phase where less contributors and activity are detected. Whether such descending trends will ever come to a null activity has still to be confirmed by continuously monitoring the activity on the pages.

The second result is still unreported in the literature; this study has shown that the categories experience an increasing number of views by interested readers. This trend is specular compared to the "edits" and "contributors" trends; although the production of content has slowed down, the requests for such knowledge are increasing steadily. This provides the foundation for a two-phase framework of user-generated content evolution.

Acknowledgment

The authors would like to acknowledge support received from the University of East London which has funded Ana Claudia Duarte Pimentel's participation in this research under its Undergraduate Research Internship Scheme 2012. The authors would also like to acknowledge the Erasmus Research Interns, Giovanni M. Alluvatti, Giuseppe De Ruvo and Marco Molfetta, who worked with Dr Capiluppi during Spring and Summer 2011 and contributed to the development of some of the tools used in this research.

References

[1] G. M. Alluvatti, A. Capiluppi, G. De Ruvo, and M. Molfetta. User generated (web) content: trash or treasure. In *Proc of* 12^{th} *Intl Workshop on Principles of Software Evolution and* 7^{th} *Annual ERCIM Workshop on Software Evolution*, IWPSE-EVOL '11, pages 81–90, New York, NY, USA, 2011. ACM.

[2] R. Almeida, B. Mozafari, and J. Cho. On the evolution of wikipedia. In *International Conference on Weblogs and Social Media*, 2007.

[3] F. P. Brooks, Jr. *The mythical man-month (anniversary ed.)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

[4] C. Haythornthwaite. Crowds and Communities: Light and Heavyweight Models of Peer Production. In *System Sciences, 2009. 42nd Hawaii International Conference*, pages 1–10, 2009.

[5] S. T. K. Lam and J. Riedl. The past, present, and future of wikipedia. *Computer*, 44:87–90, March 2011.

[6] O. Nov. What motivates Wikipedians? *Commun. ACM*, 50(11):60–64, Nov. 2007.

[7] F. Ortega. *Wikipedia: A quantitative analysis*. PhD thesis, Universidad Rey Juan Carlos – Escuela Técnica Superior De Ingeniería De Telecomunicación, 2009.

[8] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The singularity is not near: slowing growth of Wikipedia. In *WikiSym ’09: Proc of the 5^{th} Intl Symposium on Wikis and Open Collaboration*, pages 1–10, New York, NY, USA, 2009. ACM.

[9] J. Voss. Measuring Wikipedia. In *Proc of 10^{th} International Conference of the International Society for Scientometrics and Informetrics*, Stockholm (Sweden), July 2005.