An Evaluation of Wikipedia's Statistical Health Indicators

Andrew Lih, Visiting professor
University of Southern California
Annenberg School for Communication and Journalism
3502 Watt Way
Los Angeles, CA 90291
alih@usc.edu

An Evaluation of Wikipedia's Statistical Health Indicators

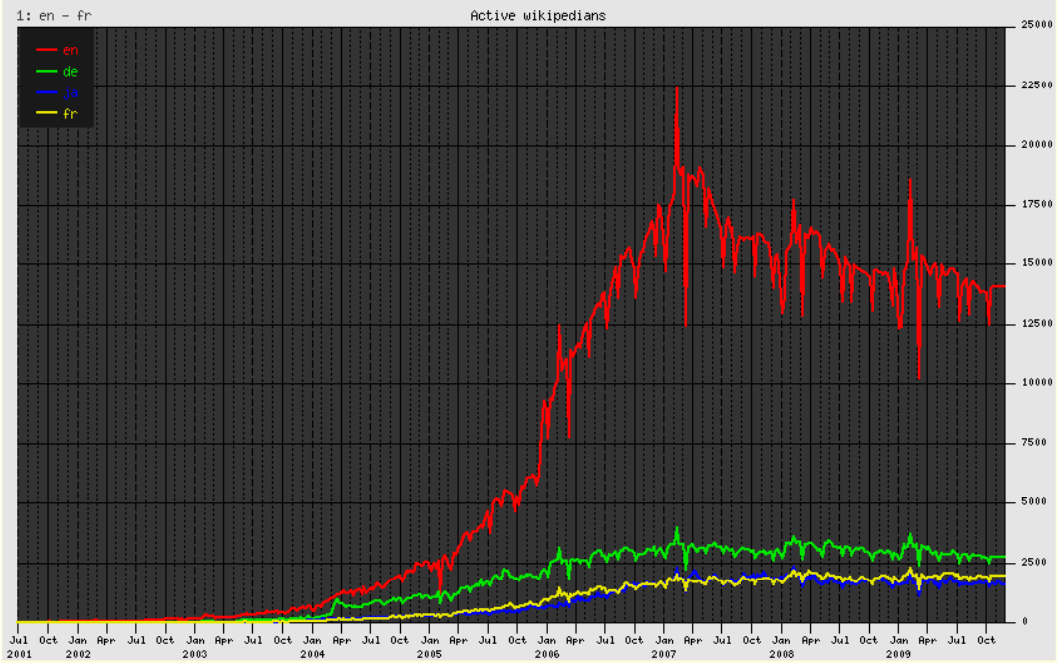
Abstract
Wikipedia is an Internet-based, collaboratively edited encyclopedia that has risen since 2003 to become the fifth most visited Web site in the world (Comscore, 2009). With over 3 million English language articles, it has become an indispensable part of the Internet as both a broad-based reference and continuously updated news source.

Wikipedia had been experiencing exponential growth since its founding, but statistics in the first quarter of 2007 indicated article production started to slowdown for the first time in its history. Analysis of this phenomenon was complicated by the lack of accurate statistics about Wikipedia's database as it became too large to make a backup "dump" in a format that could be analyzed by researchers. With the recent availability of full database downloads there has been a new round of research undertaken.

This paper examines Wikipedia's English language article production and user community statistics, and does a comparative study of research efforts by Palo Alto Research Center, academic groups, and the community itself into the decline in Wikipedia's metrics. We discuss the proper methodologies for measuring peer-production and the added complications of calculating "departure" of members from online volunteer communities through survival analysis. We consider what these numbers may mean for the Wikipedia community, and how this affects maintainability of the articles over time.

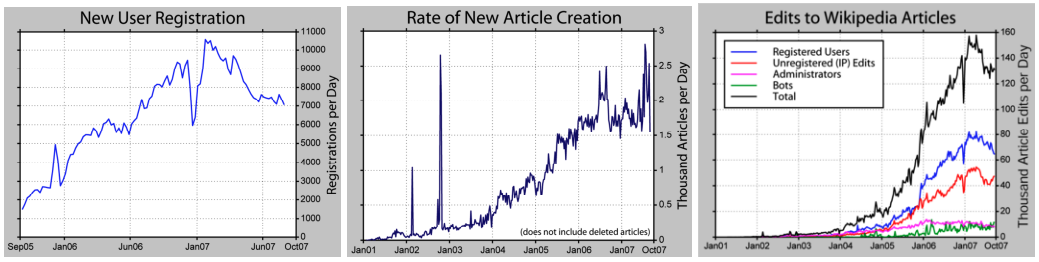
Introduction
Wikipedia has been the largest and most successful example of participatory journalism to date. Its articles are embedded in Google results for nearly all topics, as people use it as a reference work and as a continuous log of world events. Euphoric expectations regarding its exponential growth became tempered in 2007 as users started to see clues of a drop-off in new article creation in the English and German editions. It was unclear whether a natural saturation point had been reached in knowledge production, or if a sudden decline in community health was to blame.

Research Indicators
The Wikipedia community first observed article growth slowdown in the latter half of 2007 as statistics related to new user registrations and new article creation showed a flattening in the first time in Wikipedia's history. Full "dumps" of Wikipedia's database are needed to do a full accounting of user arrival and departure, and for detailed accounting of editing activity. Unfortunately the database dump technical process started failing consistently in the fall of 2006, so that complete log analysis was not possible. For the better part of two years, there were no successful full dumps of the English language Wikipedia because computing resources were inadequate to complete the process, which greatly complicated the ability for researchers to analyze Wikipedia's activity.



Historical chart of active Wikipedians, those editing 5 times or more in a week. English edits in red. (Source: http://stats.wikimedia.org/EN/PlotsPngWikipediansEditsGt5.htm)

Despite this complication, Wikipedia user Robert Rohde created statistics based on a random sample of 118793 articles, or roughly 6% of all articles in the encyclopedia at the time. The results were clear – a marked decrease in production across the board in article production and in new users in the community were reported in October 2007 which showed that the first quarter of 2007 was the start of a sharp downturn.



Statistics generated from a random sample of articles in October 2007 showing a slowdown in article production and a decline in new user registration. (http://en.wikipedia.org/wiki/User:Dragons_flight/Log_analysis)

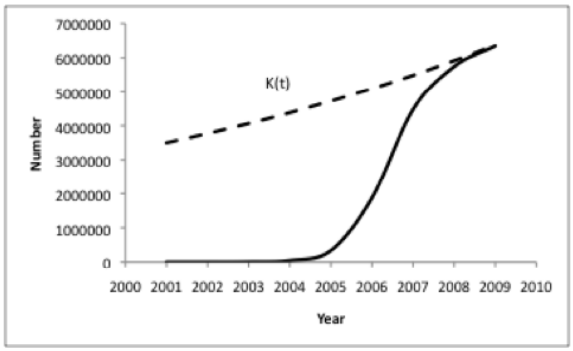
Rohde's work spawned a number of formal research studies into Wikipedia's apparent slowdown by academics and research groups.

One of the most significant efforts to look into the Wikipedia user community through full log analysis was by researchers at Palo Alto Research Center's Augmented Social Cognition Research Group.

> Bongwon Suh, Gregorio Convertino, Ed H. Chi, Peter Pirolli. *The Singularity is Not Near: Slowing Growth of Wikipedia*. In Proc. of WikiSym 2009, Oct,

2009

Using a complete dump of English Wikipedia in 2009, Chi and Suh observed that Wikipedia's growth had indeed slowed and showed that reverted edits had also increased over time in reaction to contributions from new users. Instead of a power-law growth function, they proposed a hypothetical Lotka-Volterra population growth model bound by a limit K(t) that itself grows as a function of time. This suggests a model of Wikipedia growth that adheres to a monotonically increasing level of human knowledge, with the exponential growth in its early days as a "catchup" phase to meet this constant slope.

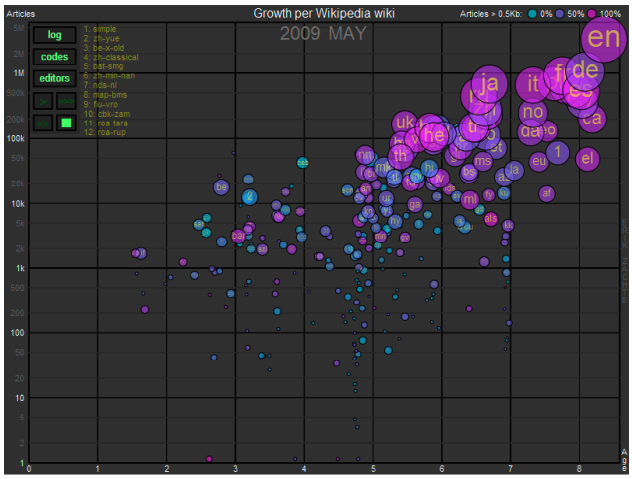


Felipe Ortega of LibreSoft Research undertook quantitative research in his Ph.D. thesis that looked at the problem of Wikipedia's growth from the approach of survival analysis to analyze when members of the community abandoned the role of editing. The use of survival analysis is complicated by the fact that "departures" from the community are notoriously hard to measure, and his categorization of editor "birth" and "death" have been disputed. We examine the merits of this style of measuring online community participation in Wikipedia, and how this compares with approaches of other research to measure participant longevity.

> "Wikipedia: A Quantiative Analysis," Felipe Ortega, PhD Thesis, defended it at Universidad Rey Juan Carlos, April 1st, 2009.
>
> "Quantitative Analysis of the Wikipedia Community of Users". Felipe Ortega, Jesus M. Gonzalez-Barahona. Proceedings of the 2007 International Symposium on Wikis, pages 75 - 86; Montreal, Quebec, Canada, October 21-25, 2007. ACM 2007, ISBN 978-1-59593-861-9.

Finally, we look at efforts from the Wikimedia Foundation's own statistics efforts from chief data analyst Erik Zachte. The 2009 analysis by the Foundation shows the same 2007 drop, but conclude a constant core of "very active" Wikipedians, or those who edit at least 100 times a month, "has remained stable since then."

"Every month, some people stop writing, and every month, they are replaced by new people," claims the Wikimedia Foundation. (http://blog.wikimedia.org/2009/11/26/wikipedias-volunteer-story/).



Modeling Wikipedia's growth by Erik Zachte and the Wikimedia Foundation (http://stats.wikimedia.org)

Conclusions

The paper will compare and contrast these different approaches and evaluate the merits of these methods for measuring community health. There is a relative lack of experience in measuring online volunteer peer-production communities given this is a new field of study. Are the findings by various teams logically concluded given their methodologies, and what are the best practices in each study that Wikipedia researchers can carry over to their future work?