Distributed Problem Solving in Wikipedia

Matthijs den Besten, Max Loubser, and Jean-Michel Dalle

Wikipedia, the free online encyclopaedia put together by volunteers, is a prime example of a distributed problem-solving network if only because it is so highly visible on the Internet nowadays and so widely used, revered, and abused. That being said, it should be noted that Wikipedia is not one network but many. Moreover, the construction of an encyclopaedia does not constitute just one problem but a manifold. It is in terms of volume and diversity that Wikipedia can be said to have stumbled upon a better solution than its contenders and it is in the understanding of how to keep a behemoth like Wikipedia afloat that most value can be accrued.

Extracting Value: Follow the Leader

One of the persons who always seems to have had a keen sense of ways to benefit from Wikipedia is its founder Jimmy Wales. Recall that before setting up Wikipedia with help of Larry Sanger and others, Mr Wales had made his living with a company that sold advertisements around content that was harvested from the web, benefiting from efforts like the open directory project. Even though Wikipedia was set up as a not-for-profit organization and no advertisements are displayed on its pages, this does not prevent others from benefiting from the efforts of the volunteers and making money by selling advertisements around content that has been reworked. In return, some, Answers.com comes to mind, use parts of their proceeds to support the Wikipedia foundation and support its sustainability. Whatever the ulterior motives of the people involved in the construction of Wikipedia, ultimately what counts more is that Wikipedia has proven to be of value for a large proportion of Internet consumers, be it for reference or for entertainment. It is in this area that less distributed problem-solving organizations such as the Encyclopedia Brittanica and Yahoo! Answers have proved to be far less successful so far. Not least, it is as founder and guru of Wikipedia that Mr Wales makes his living nowadays. He has been frequenting the conference circuit and recently, together with other Wikipedia luminaries, he set up a company called Wikia, which makes its money by hosting Wikipedia-like sites for companies like game-producers who would like to nurture their fan-community.

Variety of Problems – Diversity of Solutions

When we talk about Wikipedia, it is important to keep in mind that our object of study consists in fact of a variety of collections. In addition to the main English Wikipedia, separate collections have sprung up for virtually every language on the face of the

earth and for many of its dialects. Moreover, in some cases, a separate collection is maintained of articles that have to respond to a different set of criteria than articles in the main collection. For example, Simple Wikipedia aims to be a collection of articles that are easy to read. Another thing to keep in mind is that the construction of an encyclopaedia does not finish once a sufficient number of articles has been produced. In addition to authoring individual articles, the construction of an encyclopaedia also requires a lot of editing, integration efforts and direction. And these are not always easy tasks to perform as witnessed by the ongoing debate that as been raging between “inclusionists” who welcome articles on obscure topics and “deletionists” who would like to exclude them. Last, but not least, the life cycle of an article goes through many phases: Many articles start as “stubs”, small articles with little content; as content appears, they may become “unsimple”, or “controversial”; and as these barriers are overcome some articles reach the status of “featured article”. The transition from on phase to the other in itself can be regarded as the resolution of a sub-problem.

Corresponding to this variety within Wikipedia, there is a diversity in the methods that are applied to solve problems that are encountered. Roughly speaking, we encounter forking, shielding, and broadcasting. Forking happens when people decide to start a new collection for which they are then able to impose different quality thresholds or subject boundaries. Examples are Simple Wikipedia, but also French Wikipedia, as articles written in French would not be accepted in the main Wikipedia, and a collection of articles devoted to Star Trek characters. Shielding happens when people decide that permissions to edit articles should be restricted to a sub-set of people. This is for instance the case for the main Wikipedia page, which can only be edited by people with administrator rights. Lastly, broadcasting happens when people decide to put a label on an article indicating that there are certain things that need to be improved, as is the case with labels such as “stub”, “unsimple”, and “controversial”, or, alternatively, indicating their approval of what the article has become (“featured article”).

Measures & Metrics

What makes the study of problem solving in Wikipedia particularly appealing is the volume and the depth of the data that are made available. For every article in Wikipedia, we can know exactly which user made which changes at what time. We can measure overall volume and the diversity of the collections, and we can also look at the turnover among editors and the rate of change within articles. Table 1 summarizes a variety of measures that have been employed in studies of Wikipedia sofar. Undoubtedly, many more will follow.

Table 1 Wikipedia measures by publication

Study	Measures
Lih 2003	Edits per article
	Unique editors per article
	Average article size over time
Viegas, Wattenberg, and	Article length over time

Dave 2004	
	Mass deletions
	Survival time of edits
Stvilia, Twidale, Smith, Gasser 2005	Num. of Anonymous User Edits
	Total Num. of Edits
	Num. of Registered User Edits
	Num.of Unique Editors
	Article length (in # of characters)
	Currency (a)
	Num. of Internal Links
	Num. of Reverts
	History Num. of External Links
	Article Median Revert Time
	Num. of Internal Broken Links
	Connectivity (b)
	Num. of Images
	Article Age
	Diversity (c)
	Information Noise(content) (d)
	Flesch readability
	Kincaid readability
	Article Admin. Edit Share
den Besten and Dalle 2007	Flesch readability
Kittur, Suh, Pendleton, and Chi 2007	Indirect work (f)
	Direct work
	Reverts

	Vandalism fixes
	Revisions
	Page length
	Unique editors
	Unique editors / revisions
	Links from other articles
	Links to other articles
	Anonymous edits (#, %)
	Administrator edits (#, %)
	Minor edits (#, %)
	Reverts (#, by unique editors)
Kittur, Chi, Pendelton, Suh, and Mytkowicz 2007	Percentage of total edits made by admins
	Number of edits per month made by admins
	Percentage of total edits made by bots
	Average number of edits per user per month
	Number of words added and removed per edit
Anthony, Smith, and Williamson 2007	Retention rate of contributions
	Number of contributions per author
	Registered/unregistered status of author
	Article size
	Contribution size

Simple Wikipedia – Preliminary Results

We did a pilot study on Simple Wikipedia. Simple Wikipedia is a spin-off of Wikipedia that was initiated in 2003 because people felt that many articles in Wikipedia were too hard to read, due to jargon, formality, or for other reasons – especially for children and non-native speakers of English. Simple Wikipedia, was the hope, would be the place where people go to look for an easily readable descriptions of topics. At the same time, contributors to Simple Wikipedia would commit to the ideals of this sub-project or at least adhere to an editorial policy that calls for greater readability when contributing descriptions of topics.

Simple Wikipedia is an ideal candidate for a study of the performance of distributed problem solving networks in the way that we envisage such a study. For, Simple Wikipedia is a project for which we can determine relatively easily how well it adheres to its goals, zoom in on efforts that are made to address specific problems, and assess the result of these efforts. First of all, with less then 20 000 articles in its collection Simple Wikipedia is a relatively small encyclopedia project. Consequently, no extraordinary computational resources are needed to extract the project archive and analyze its contents. Besides, Simple Wikipedia is a project centered around a very specific goal, readability. What's more, readability is something that can be measured. So, here we have a project that we can assess on its own terms. Moreover, Simple Wikipedia, as a separate project, was able to experiment and implement with its own editorial policies and managerial policies specifically to help attain its goal of simplicity. Most importantly, Simple Wikipedia has come to rely on the tag "unsimple" – more recently called "complex" – to single out articles, which do not meet its standards and need to be improved.

The archive of Simple Wikipedia is available from the web at downloads.wikimedia.org. Our analysis here is based on the archive of July 2007, which contains the revision history of over 25 000 articles and around 27 000 pages of a different type such as discussion pages and user-pages where regular contributors present themselves. For each edit on an article, the archive lists the ip-address or user-name and user-id of the editor, the time of edit, comments made by the editor, and the actual text of resulting from the edit. The text is marked up with tags to identify structural elements like sections and sub-sections and tags of a different type to identify labels – also called templates – that are applied to the text. For instance, an article that is considered to be hard to read, will contain the string "{{unsimple}}" in the raw text. Slightly harder to determine is the status of an editor. We can easily distinguish between editors who are known to the system and editors who contribute anonymously as the latter are identified by their user-id while for the former only the ip-address is listed. Bots, scripts that carry out small repetitive edits such as spell checks and interlinking of articles, usually have a user-name that ends with "bot". In addition, the user-id of these bots is listed as a user belonging to a special group in the auxiliary user-group table, as is the user-id of the users with special rights known as administrators. The readability of an article is determined by computing the Flesch readability score of the article's text with help of the GNU Style package. This score is a function of the number of syllables per word and the number of words per sentence (Flesch, 1979). More precisely, the formula 'score = 206.835 – 84.6*syllables/words – 1.1015*words/sentences' yields a number that is usually between 0 and 100 and between 60 and 70 for standard English texts. This Flesch reading easy formula, which has been elaborated on the basis of school texts by Flesch in 1948, has been very popular, especially in the US, as a measure of plain English. Its popularity rests on the fact that the formula is easy to compute, yet often accurate. Even work-processing programs like Word often provide the score as part of their statistics.

Error! Objects cannot be created from editing field codes.

Figure 1 Performance at Project Level – Simple Wikipedia

On the basis of the archive, it is possible to reconstruct the history of Simple Wikipedia. For each article in the archive we know when it was first introduced and for every month that Simple Wikipedia existed, we can find the versions of the articles in the archive that existed in that month and we can count their number and properties like their overall readability. Figure 1 shows the result of such a reconstruction for Simple Wikipedia from January 2003 until December 2007. The figure shows an upward sloping line indicating the total number of articles in the collection for each month and a downward sloping line indicating the overall readability of the articles in the collection for each month. Note that the continuous growth in the number of articles in Simple Wikipedia is at least partially due to the fact that articles that have been removed from the collection and are not currently available anymore, do not appear in the most recent archive either. Even so, the growth in the number of articles over time is impressive. This rate of growth may be a factor that explains why the readability indicated by the Flesch readability score shows a gradual decline: As the size of articles in Simple Wikipedia grew, it became more unwieldy and editors faced an ever harder task to maintain the standards of readability, is one interpretation that suggests itself. Looking closely, we can distinguish a phase of substantial decline in readability in the first half of 2004 followed by a more stable phase in the second half of 2004. It was in the second half of 2004 that the practice of tagging articles with the label “unsimple” first appeared. It might be that the slow down in the decline in readability could be attributed to this. With a readability index of over 70, Simple Wikipedia still scores very well given that standard English falls between 60 and 70. Still, the fact that readability continues to decline should be worrying.

Selected Readings

[1] D. Antony, S. W. Smith, and T. Williamson. Explaining quality in internet collective goods: Zealots and good Samaritans in the case of wikipedia. Technical report, Darmouth College, Hannover, NH, November 2005.

[2] M. den Besten, A. Rossi, L. Gaio, M. Loubser, and J.-M. Dalle. Mining for practices in community collections: finds from simple wikipedia. Submitted to *Open Source Systems 2008*, Milan, 2008.

[3] W. Emigh and S. C. Herring. Collaborative authoring on the web: A genre analysis of online encyclopedias. In *Proceedings of the Thirty-Eighth Hawai’i International Conference on System Sciences (HICSS-38)*, 2005.

[4] J. Mateos Garcia and W. E. Steinmueller. Applying the open source development model to knowledge work. INK Open Source Research Working Paper 2, SPRU - Science and Technolgy Policy Research, University of Sussex, UK, January 2003.

[5] L. Sanger. The early history of Nupedia and Wikipedia: A memoir. In *Open Sources 2.0*. O’Reilly, 2005.

[6] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology*, 2008.