Future Generation Computer Systems 31 (2014) 120–133

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Visualizing large-scale human collaboration in Wikipedia

Robert P. Biuk-Aghai *, Cheong-Iao Pang, Yain-Whar Si

Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Av. Padre Tomas Pereira, Taipa, Macau S.A.R., China

HIGHLIGHTS

- A novel method for analysis and visualization of large wikis such as Wikipedia.
- Visualization of a wiki in a form similar to a geographic map.
- Analyzed and visualized English, German, Chinese, Swedish and Danish Wikipedia.
- Significant co-author count differences between different language Wikipedias.
- Superior over text data in usability, accuracy, speed and user preference.

ARTICLE INFO

Article history:
Received 9 January 2012
Received in revised form
10 November 2012
Accepted 6 April 2013
Available online 23 April 2013

Keywords:
Visualization of collaborative processes & applications
Wikipedia
Information visualization
Category
Co-authoring

ABSTRACT

Volunteer-driven large-scale human-to-human collaboration has become common in the Web 2.0 era. Wikipedia is one of the foremost examples of such large-scale collaboration, involving millions of authors writing millions of articles on a wide range of subjects. The collaboration on some popular articles numbers hundreds or even thousands of co-authors. We have analyzed the co-authoring across entire Wikipedias in different languages and have found it to follow a geometric distribution in all the language editions we studied. In order to better understand the distribution of co-author counts across different topics, we have aggregated content by category and visualized it in a form resembling a geographic map. The visualizations produced show that there are significant differences of co-author counts across different topics in all the Wikipedia language editions we visualized. In this article we describe our analysis and visualization method and present the results of applying our method to the English, German, Chinese, Swedish and Danish Wikipedias. We have evaluated our visualization against textual data and found it to be superior in usability, accuracy, speed and user preference.

1. Introduction

The emergence of Web 2.0 technologies in recent years has made human-to-human collaboration on unprecedented scales not only possible but a reality. One of the best-known examples of world-wide large-scale collaboration is Wikipedia, "the free encyclopedia that anyone can edit" (Wikipedia's own slogan) [1]. Wikipedia has great value that has not yet been fully researched. Past research on Wikipedia has focused on both a *micro-level* (e.g. [2,3]) and a *macro-level* of analysis (e.g. [4–7]). A micro-level of analysis typically focuses on a single article, whereas a macro-level of analysis studies the wiki as a whole, exploring relationships and the evolution of the entire content collection, among others. Our research falls in the latter class and aims to obtain an overview of Wikipedia and identify popular topic areas. By applying this to different language Wikipedias we wish to discover differences among those language editions, and by implication to discover differences of interest in those topic areas among the user communities of those language groups. However, our aim in this research is for our methods and tools to be general enough to be applied to other wikis besides Wikipedia, for example intra-organizational wikis.

The technology underlying Wikipedia is relatively simple: a wiki engine (MediaWiki) implemented in PHP on a web server which most users access through a web browser, and primarily making use of three main functions: searching for, reading and editing articles. Other functions, used to a much lesser extent by common users, are asynchronous discussion of articles, viewing the revision history of an article, comparing revisions to find out what has changed between them, undoing specific revisions, and a few others. Wikipedia administrators have additional privileges, allowing them to protect articles (making them read-only), moving (renaming) articles, deleting articles entirely, blocking users, and other administrative/maintenance functions.

The Wikipedia user base is large and broad: the English Wikipedia edition alone counted about 17.8 million registered

* Corresponding author. Tel.: +853 83974375.
E-mail addresses: robertb@umac.mo (R.P. Biuk-Aghai), inbox@patrickpang.net (C.-I. Pang), fstasp@umac.mo (Y.-W. Si).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.04.001

Table 1
Wikipedia user statistics, as at 8 Nov 2012 (active user % is relative to all registered users, admin user % is relative to active users).

Language	Users				
	Registered	Active		Admins	
		Total	(%)	Total	(%)
English	17,813,716	132,800	0.7	1462	1.1
German	1,535,302	21,649	1.4	267	1.2
Chinese	1,316,773	6,994	0.5	78	1.1
Swedish	299,093	3,136	1.0	88	2.8
Danish	171,699	1,155	0.7	37	3.2

users in November 2012, out of which 132,800 (0.7%) are considered "active" users (meaning that they have performed some action within the past 30 days). A small portion of these registered users are site administrators, under 1500 (about 1% of active users) in the case of the English Wikipedia. An overview of user statistics for a few selected Wikipedia language editions that we have studied is shown in Table 1. We selected these Wikipedia language editions mainly for the practical reason that we understand these languages (which is required for interpreting the visualized result), but also to give us a selection of very large (English), medium-sized (German, Chinese) and small (Swedish, Danish) Wikipedias.

Wikipedia content is user-contributed, meaning that end-users can add to, modify and delete content in Wikipedia articles. They can also write entirely new articles and link these to other articles. To better organize content Wikipedia has a hierarchical category system, and any given article can be marked as belonging to any number of categories. For instance in the English Wikipedia (as of January 2012), article "Wiki" is assigned to category "Wikis" (plus five other categories), which in turn has parent category "World Wide Web" (plus four other parent categories), which in turn has parent category "Digital Media" (plus six other parent categories), and so on. The same as with articles, categories are also user-contributed: users can create new categories, assign categories to parent categories, assign articles to categories, and change existing article-to-category and category-to-category assignments. The result is an organically evolving category system that reflects the current needs of the user-contributor community. One of the implications of such an open editing process is that it may result in different granularity of the category hierarchy. Table 2 shows the numbers of articles and categories of the five Wikipedia language editions we have analyzed (these counts include all articles and categories, including non-content ones that we later remove). The absolute numbers of articles and categories differs significantly in these different language editions, but so does the average number of articles per category (the right-most column in Table 2) which indicates the granularity of the category hierarchy. In four of the five analyzed Wikipedia language editions the number of articles per category ranges between about 4 and 8, but in the German Wikipedia there are on average 17.7 articles per category, suggesting a much coarser category hierarchy granularity. As documented on Wikipedia itself, the German edition of Wikipedia differs from other editions: "Compared to the English Wikipedia, the German edition tends to be more selective in its coverage" and "Categories are usually introduced only for a minimum of ten entries and are not always subdivided even for larger numbers of items,"[1] which explains this difference in the articles per category statistics. In fact, the absolute number of categories in the German Wikipedia is even smaller than that in each of the Chinese and Swedish Wikipedias although the number of articles is significantly larger. Different language communities clearly have different standards as to how fine-grained they believe their category hierarchies should be.

Table 2
Sizes of Wikipedia language editions studied (database dump of January 2011).

Language	No. of articles	No. of categories	Art./Cat.
English	3,411,491	602,141	5.7
German	1,217,553	68,677	17.7
Chinese	352,562	82,639	4.3
Swedish	393,504	82,039	4.8
Danish	147,576	19,193	7.7

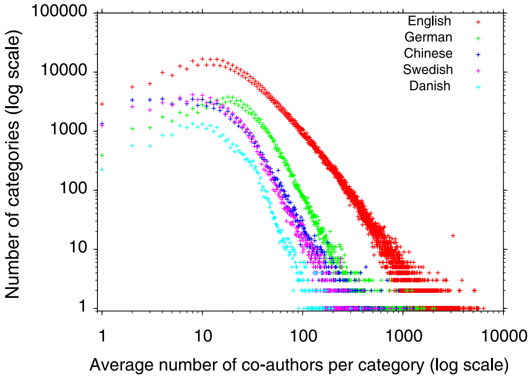

Fig. 1. Distribution of the average number of co-authors per category in English, German, Chinese, Swedish and Danish Wikipedia.

Wikipedia is not only user-contributed, but as a direct result of its openness the number of contributors that get involved in editing a given article can also be very large. We have analyzed this number of co-authors and for each category calculated the average number of distinct co-authors of all articles assigned to that category. This average count of co-authors per category varies dramatically between categories. For example, in the English Wikipedia there are 15 categories, each of which has an average number of co-authors of over 5000. On the other hand there are over 100,000 categories, each of which has an average number of co-authors of 10 or fewer. The distribution of average number of co-authors per category in the five Wikipedia language editions we analyzed is plotted in Fig. 1. Interestingly, despite all the differences in scale and category hierarchy granularity among the different language Wikipedias, their curves have essentially the same shape. We determined goodness of fit using the Anderson–Darling test and found the data from all five language editions to follow a geometric distribution, with *p* ranging between 0.03 and 0.05 in the different languages.

However, this distribution of the average number of co-authors per category does not reveal *where* the differences lie—which categories attract the most co-authors to their articles, and which the fewest. This may also differ between different Wikipedia language editions, as the top-10 list of categories with highest co-author count shown in Table 3 indicates politics and religion feature strongly in the English Wikipedia, whereas in the German Wikipedia it is art and society that feature strongly, with some sports and television appearing in both top-10 lists. We also do not know if similar co-author counts cluster together by topic, i.e. whether categories that belong to the same parent category also have similarly high co-author counts. This information is difficult to obtain as topic clusters are hard to determine given the large number of parent categories that a given category may belong to.

We have devised a method for analyzing the category hierarchy to determine which major parent category a given category should belong to. This allows us to aggregate co-author counts from individual categories recursively up to their ancestor until the top of the category hierarchy. Doing so reveals which categories at the highest level are the most collaborative, and which the least. We have then used the output of this analysis to visualize the

[1] http://en.wikipedia.org/wiki/German_Wikipedia.