Exploring Authorship in Wikipedia to Modify Lotka's Law: Bibliometric Patterns in the World's Largest Encyclopedia

M. Cameron Jones, Karen Medina

mjones2@uiuc.edu, kmedina@uiuc.edu
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign,
501 E. Daniel St. Champaign, IL 61820 (USA)

Introduction

Lotka's model characterizes the distribution of authorship in a given field. It states that, "the number (of authors) making n contributions is about $1/n^2$ of those making one; and the proportion of all contributors, that make a single contribution, is about 60 percent" (Lotka 1926, cited in Potter 1988). Lotka's model has been applied in a variety of modes of academic work including the distribution of contributions to open-source software development (Newby, et al., 2003) but a study of the applicability of Lotka's model to other modes of collaborative writing has yet to be conducted.

Similar to open source software development, the Wikipedia represents a new mode of collaborative academic production, one in which large numbers of distributed individuals contribute knowledge to a central, collaboratively written document. The Wikipedia has over 16,000 registered users, collaboratively creating, editing and revising over 446,000 articles. By comparison, Encyclopaedia Britannica contains approximately 65,000 articles authored by some 4,000 contributors. What makes Wikipedia unique is the degree to which a single article may be collaboratively written. For example, the Wikipedia entry for Chocolate has been modified by no fewer than 128 non-anonymous users. The degree to which entries like Chocolate are unique or representative in the universe of the Wikipedia has yet to be formally measured or understood.

Wikipedia provides a unique opportunity to study the question of what to do with multiple authors. In the month of December, 2004, of the 16,000 users, 5,908 had made more than 5 contributions each and 913 made more than 100 contributions each. This is promising evidence that a Lotka distribution may be applicable to the authorship of Wikipedia articles. Through the history function of Wikipedia, the contribution of each author is recorded. In other bibliometric situations monitoring the contributions of individual authors is extremely difficult. While multiple-authorship is present in traditional scholarly writings, it is hard to know what the individual contributions to the work are. As a result, it is difficult to decide how to attribute a given document to each author. Wikipedia gives us an opportunity to study this where other studies have not.

A New Taxonomy of Authorship

In this poster we present the new taxonomy of authorship, which Wikipedia allows us to explore. Our taxonomy utilizes several definitions of authorship derived from analogues in the print world, and then builds on them to include other aspects of authorship, which the Wikipedia enables us to measure.

We start with current definitions of authorship counts. Nicholls's review (1987) summarizes three possible means of measuring co-authorship in print media: 1) complete count, 2) adjusted count, and 3) straight count. "Complete count" gives each collaborating author equal, whole authorship credit. "Adjusted count" is a way of proportioning the authorship equally among the authors (if there are 4 authors, then each gets 1/4 of a credit point). "Straight count" attributes authorship to the first author listed on the article. The first two methods of counting can be mapped easily to a Wikipedia article. The third method is more difficult. First author in a Wikipedia article is not explicit. If we can determine which author contributed the most content to the article, this author may be considered the principal author and thereby receive full credit in the straight count method. These methods of counting authorship are fairly well agreed upon.

Other methods have been suggested which divide up the adjusted count. For instance, (Diodato, 1994) divides authorship between the co-authors by dividing the total number of pages by the number of co-authors. (Trenchard, 1992) assigns the adjusted count by a hierarchical weighting where the first author gets weighted more, and the final count for all the authors still depends on the number of pages in the paper.

With these methods of measuring co-authorship, we are armed to dive into other forms of authorship which Wikipedia opens up for us. The Wikipedia allows us to see the personal contributions each author makes to a given article by way of the history function. Some of these contributions may be substantial while others may be cosmetic or even destructive.

For substantial contributions, we may count the number of bytes that are changed in the article. We could consider any number of bytes as a significant

contribution, or we could set a certain threshold below which the contribution is considered insignificant.

When an article is started, it is given a title and has no content. At this point it is a stub. The author who initialized the stub could be considered a contributing author, even though no actual content was created. The stub creator put thought into action without which the article may never have come into existence. But this authorship may be counted a special way, perhaps as a supervisory author.

Appending an article (adding more bytes to the end) is a special case of adding bytes to an article because it is not correcting, clarifying, or editing existing bytes but is more creatively adding. Appending is considered a significant contribution, but again, just adding a period to the end of an article may not be considered a significant contribution.

Others edits made to an article involve varying combinations of adding and deleting existing bytes from the article. If there is no net change in the number of bytes, this may still be a significant contribution if the article is substantially more understandable. Net changes which involve only whitespace may not be considered significant. Note that the Wikipedia itself collapses certain types of white spaces, especially those added at the end of an article.

Correcting the spelling or grammar of an article is not considered a significant contribution. While this is very helpful for improving the understandability of the article, the role is more that of an editor or a friendly colleague than that of a major contributor, and will not be counted as a contributing author.

Adding a clarifying example in the middle of an article is a substantial contribution.

Moving a paragraph from one place to another is again considered part of the friendly colleague role.

Deleting an entire section or rolling back to an earlier version may be considered a significant contribution if these changes remain intact for a significant period of time.

Random, nonsensical, or vicious edits are not considered significant contributions.

While some of these changes are very easy for a machine to detect, others such as vicious edits are not. If the edits remain intact for a substantial amount of time, or perhaps after a certain number of visits or other edits, then the edits will be considered well-meaning.

Our new taxonomy will be used to analyze the Wikipedia authorship, and in turn will allow us to make adjustments and recommendations to Lotka's model using our refined definitions of "authorship," and "significant contribution"; and some suggestions for the identity conditions of a digital document (see Renear, 2003). All of these will apply toward any and all bibliometrics involving mutable content with multiple authors.

Acknowledgments

The authors would like to thank Allen Renear, Michael Twidale, Jin Ha Lee and Ingbert Floyd for their support of and input into this project

References

Diodato, V. (1994). Dictionary of Bibliometrics. New York: The Haworth Press.

Encyclopedia Britannica. Available from http://www.britannica.com (accessed February 28, 2005).

Newby, G. B., Greenberg, J., & Jones, P. (2003). Open Source Software Development and Lotka's Law: Bibliometric Patterns in Programming. Journal of the American Society for Information Science and Technology. 54(2):169-178.

Nicholls, Paul Travis. (1989). Bibliometric modeling processes and the emprical validity of Lotka's Law. Journal of the American Society for Information Science. 40(6), pp. 379-385.

Potter, William Gray. (1988). 'Of Making Many Books There is No End': Bibliometrics and Libraries. The Journal of Academic Librarianship 14, p. 238.

Renear, Alan H. and Dubin, David S. (2003). Towards identity conditions for digital documents. In S. Sutton, editor, Proceedings of the 2003 Dublin Core Conference, Seattle, WA, October 2003. University of Washington.

Trenchard, P. M. 1992. Hieararchical bibliometry: a new objective measure of individual scientific performance to replace publication counts and to complement citation measures. Journal of Information Science. 18, p. 69-75.

Wikipedia Statistics. Available from http://en.wikipedia.org/wikistats/EN/Sitemap.htm (accessed February 28, 2005).