Exploring Underproduction in Wikipedia

Andreea D. Gorbatai
Cotting 324A
Harvard Business School
Boston, MA
agorbatai@hbs.edu

ABSTRACT

Researchers have used Wikipedia data to identify a wide range of antecedents to success in collective production. But we have not yet inquired whether collective production creates those public goods which bring most value-add from a social perspective. In this poster I explore two key circumstances in which collective production can fail to respond to social need: when goods fail to attain high quality despite (1) high demand or (2) explicit designation by producers as highly important. In the context of Wikipedia, I propose first to examine articles that remain low quality, or underproduced, despite the fact they are viewed often; and second, to examine articles that remain low quality despite the fact that they were identified as important by Wikipedia contributors. This research highlights the fact that collective production needs to be examined not only by itself but also in the context of a market for goods in order to ascertain the benefits of this production form. The final version of this study will integrate data on underproduced articles with data on knowledge categories to uncover systematic patterns of underproduction at the category level and predict which categories are most in need of quality improvement. Additionally I will use in-depth qualitative methods to examine the mechanisms through which underproduction occurs in select knowledge categories to distill practical recommendations for collective production improvement.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces**]: Group and Organization Interfaces – Computer-supported cooperative work

General Terms

Measurement, Performance.

Keywords

Social goods, Collective production, Underproduction.

1. INTRODUCTION

Wikipedia has been hailed as an exemplary of collective production. Started in 2001, Wikipedia manifested a spectacular growth in terms of number of articles, contributors, languages covered (over 250 languages as of March 2011) and readership. A mere five years after its launch, Wikipedia was ranked in the top ten most visited sites in the world, and has consistently preserved this position ever since. Together with several other early starters such as Amazon Mechanical Turk and Facebook, Wikipedia' success inspired a new wave of collective and distributed work applications aimed at creating productive social network platforms and online marketplaces.

Over the past few years, several researchers have drawn attention to the fact that the growth of the English Wikipedia has slowed down [1], partly because it has reached comparable comprehensiveness to traditional encyclopedias. However, Wikipedia coverage and accuracy still varies widely among different knowledge domains. Several researchers including Halavais and Lackaff [2] have concluded that "Wikipedia's topical coverage is driven by the interests of its users, and as a result, the reliability and completeness of Wikipedia is likely to be different depending on the subject-area of the article."

The uneven coverage can be traced down to several potential causes. First, surveys suggest that the socio-demographic characteristics of participants may be skewed towards over-representing one gender more than the other, as well as certain age groups, geographic locations and occupations [3]. Second, article quality and topic coverage may be affected by availability of information. The more difficult and costly it is to locate and access information about a topic, both financially and time wise, the less likely the topic will be represented. Lastly, participants may have access to certain information about a topic of interest but lack the knowledge that the information is particularly relevant or needed.

This poster proposes two types of article underproduction: low-quality articles which are in high demand from readers, and low-quality articles which are rated by contributors as of high importance for at least one knowledge domain. I first provide a brief review of existing research on Wikipedia coverage. Then, using a dataset that includes information about article views, length, quality, importance, as well as main knowledge category labels for all the English Wikipedia articles as of May 2009, I describe the types and extent of collective production failures and summarily address the implications of these findings.

2. WIKIPEDIA COVERAGE

A number of scholars from diverse disciplines have examined Wikipedia coverage. For example, one study examined the network structure of Wikipedia articles from the perspective of semantic coverage, as compared against Encyclopedia Britannica and Encarta.com. Others examined samples of Wikipedia articles from a given knowledge field, such as history or psychology [4]. An in-depth study by Halavais and Lackaff [2] (1) contrasted the distribution of books in print against a random sample of 3,000 Wikipedia articles; and (2) compared the distribution of topics in three established, field-specific academic encyclopedias (poetry, linguistics and physics) with topic distribution within corresponding Wikipedia categories. I extend this research by examining quality versus both actual need (as measured by times an article was accessed/ viewed) and idealized need (as measured by contributor-designated importance) instead of using print publications as a reference for topic coverage.

3. DATA AND RESULTS

For this study I have merged three separate sources containing data from the English Wikipedia. The first dataset contained information regarding the quality and importance ratings of 2,752,543 articles (as of May 2010). The second dataset contained bi-monthly logs of views for 1,418,759 articles, constructed from

Copyright is held by the author/owner(s).

WikiSym '11, Oct 03-05 2011, Mountain View, CA, USA
ACM 978-1-4503-0909-7/11/10.

server logs between November 1st, 2008 and February 28, 2009 provided by Wikimedia contributor Midom. A third dataset on main knowledge categories for 1,422,050 articles (to be used in further analyses) was obtained by the author in March 2011.

Wikipedia article quality involves one of seven ratings. Articles rated FA, A, or GA are high-quality articles that are considered to provide a "useful reading experience" to nearly all readers. By contrast, Start and Stub articles are considered of low quality. In the middle of the quality spectrum, B and C articles fall short of being useful to all readers but are expected to be at least "useful to readers looking for a well-structured, reasonably detailed overview." For the purposes of this analysis, FA-, A- and GA-quality articles and B- and C-quality articles have been respectively grouped together to form - along with Start and Stub articles - four quality categories ranging from high to low.

A simple cross tabulation of article quality by number of article views (*Table 1*) revealed that articles with under 148 average views per two-week period have a similar distribution across quality categories as articles with 148 to 22,026 views: about 60.3-74.8% were Stub quality, 21.7-33.5% were Start quality, and only 0.2-0.4% were high-quality (FA, GA, or A) articles. In contrast, pages viewed more than 22,026 times in a two-week period reflected a quality distribution in which 7.3% were Stub quality, 3.5% were high quality, and the remaining were approximately equally distributed between the Start and B/C quality. Hence almost half of the very frequently accessed articles are of at least good quality, but high-quality articles are hardly better represented among articles in high demand compared to those of medium interest.

Table 1. Cumulative article views by quality

Page Views	Article quality			
(ln)	Stub	Start	B/C	FA/A/GA
1-4	74.8%	21.7%	3.2%	0.2%
5-9	60.3%	33.5%	5.8%	0.4%
10+	7.3%	46.5%	42.8%	3.5%

The second type of underproduction occurs when highly important and core concepts are not adequately explained. Wikipedia articles are rated in terms of their importance by WikiProject participants as top, high, mid, and low,[1] where a low-importance article is of specialist interest, while a high- or top-importance article is a "must-have" for a print encyclopedia or contributes significant depth of knowledge to a domain. I find that although FA/A/GA-quality articles represent only 0.64% of total articles, they represent 1.76% of top-importance articles, and 1.42% of high-importance articles. Similarly, B/C-quality articles represent 7.68% of the total articles but over 36% of top-importance articles and almost 25% of high-importance articles (*Table 2*). This suggests that a large proportion of important articles are high quality: about 26% of high-importance articles, and about 38% of top-importance ones are of C-quality or better.

On the other hand, there are many more low-importance articles than of top or high importance, and hence the percentages of high-quality articles among these are not directly comparable. Examination of the percentage of high-quality articles that are top or high importance reveals another story: only 16.5% of high-quality articles fall within this category, and about one quarter of B/C-quality articles are of top or high importance. Thus, the majority of medium- and high-quality articles found in Wikipedia are articles which address minutiae for specialized audiences, whereas more than 74% of high-importance articles and over 60% of top-importance articles are of no use to most consumers of Wikipedia content because they are of Start or Stub quality. These findings suggest that despite improved coverage of knowledge in Wikipedia, the collective production process commonly fails to improve the quality of the most important articles.

Table 2. Article quality and importance analysis.

Importance	Low	Med.	High	Top	% quality
Stub	65.50	39.78	25.22	14.24	56.68
Start	30.03	46.18	48.68	47.35	35.01
B/C	4.01	13.09	24.68	36.65	7.68
GA/A/FA	0.45	0.95	1.42	1.76	0.64
%Importance	70.19	22.69	5.99	1.13	100

*Note: N=507,706 due to unassessed "importance" articles.

4. CONCLUSION

Preliminary findings intimate that, despite Wikipedia's success in providing free access to human knowledge, contributors fall short of producing high-quality work for some of the most read or important articles. Using article topic category data, I propose to uncover which categories are more prone to underproduction, either compared against article views or designated importance, and to engage in Talk page analyses and interviews with Wikipedia editors to explore the causes of underproduction. This research is theoretically important because it proposes two types of underproduction and will examine the social processes that underlie the differential development of articles. From a practical perspective, this paper will inform potential improvements to Wikipedia and other collective production systems so that needed or important goods are provided and are of sufficient quality.

5. ACKNOWLEDGMENTS

The author thanks Wikipedia contributors (aliases) Magnus Manske, Midom (Domas Mituzas), and MCMcBride for facilitating access to data, and Asa Tapley for his patient feedback.

6. REFERENCES

[1] Ortega, F. Wikipedia: A Quantitative Analysis. Unpublished doctoral dissertation at University Rey Juan Carlos, Madrid, Spain, 2009.

[2] Halavais, A. and Lackaff, D. An analysis of topical coverage of Wikipedia. Journal of Computer Mediated Communication, 13, 2 2008), 429-440.

[3] Glott, R. and Ghosh, R. Analysis of Wikipedia survey. Topic: Age and Gender Differences. UNUMerit, City, 2010.

[4] Schweitzer, N. Wikipedia and Psychology: Coverage of Concepts and Its Use by Undergraduate Students. Teaching of Psychology, 35, 2 2008), 5

[1] A WikiProject is a project in Wikipedia through which a set of contributors set to manage a specific topic or family of topics within Wikipedia. It is composed of a collection of pages and a group of editors who use those pages to collaborate on encyclopedic work.