The Article/Editor Ranking : Bringing Order to Wikipedia

Maximilian Klein
OCLC Research
777 Mariners Island Blvd
San Mateo, CA, 94404
kleinm@oclc.org

Thomas Maillart
School of Information
University of California,
Berkeley, 102 South Hall
Berkeley, CA 94720
thomas.maillart
@ischool.berkeley.edu

John Chuang
School of Information
University of California,
Berkeley, 102 South Hall
Berkeley, CA 94720
chuang
@ischool.berkeley.edu

ABSTRACT

We introduce a new method to jointly rank the quality of articles and the expertise of editors in Categories of Wikipedia, based on the bi-partite network information of who has contributed at least once to an article on the one hand, and what are the articles that have edited at least once by a given author on the other hand. We show that this "reflexive" ranking method exhibits high correlations with usual article quality and user expertise metrics, which account for quality on Wikipedia (that we assume to be a grand truth here). In particular, we find that the quality of an article can be captured very well by our method right after a few edits, while the expertise of editors is captured increasingly better over time. Our results suggest that it is easier to predict the quality of an article from the editors who touched it, rather than editor expertise from articles they have edited.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

to be completed

Keywords

to be completed, if necessary

1. INTRODUCTION

From product reviews to online collaboration platforms, the World Wide Web is increasingly populated with knowledge contributed by the crowds. One positive aspect is the immediate sharing of information, which in turn helps others take more informed decisions about the quality of a product, the cleanliness of a restaurant, whether it is worth buying a book. But the reliability of the contribution is limited by the level of expertise of the person who made the contribution. In many case, repeated contributions of the same type

Figure 1: Matrix

by several individuals (e.g., reviews) average out idiosyncracies (assuming that individual do not influence each others). Many crowd sourcing mechanisms have been designed in this way []. The reliability/quality of a contribution is however more critical when the same contribution is not necessarily repeated, or repeatable. Open source collaboration projects face such an issue: rewriting several times Wikipedia or the whole open source software codebase would be simply be impossible. Yet these open collaboration projects can achieve remarkable quality [**?**]. This is achieved through *peer-production* a labor organization, which mainly relies on two main ingredients : (i) task self-selection and (ii) peer-review : individuals choose the tasks they believe they are more qualified for, and they review the work of others also according to their skills [**?**]. These rules have been initially used for open source software development [**?**], for which it is possible to ultimately test the quality of the work by executing the code. However, for natural language knowledge, like Wikipedia, there is no "machine" to execute the code. Therefore the quality of knowledge remains subjective, and it is mainly tied to the expertise of the person in the domain of contribution, and to the number of persons (and their respective expertise) who have contributed to a piece of knowledge (e.g., an article). The level of expertise in one domain can in turn be approximated by the number of related articles contributed by the same individual, and so on.

On Wikipedia, is there a way to assess relative quality of both articles and editors by only considering whose editor has touched which article ? This is the problem we are addressing in this paper.

The rest of this paper is organized as follows. With begin with discussion of related work, followed by a description of the method. We then present the results and conclude.

2. RELATED WORK

The model we present here is deeply influenced by a recent stream of research in economics that aims at explaining the GDP of countries based on the nature of goods they produce and export. The first model was proposed by [**?**, **?**], and reworked by

A network analysis of countriesÕ export flows: firm grounds for the building blocks of the economy [**?**]; A new metrics for countries' fitness and products' complexity [**?**]; Measuring the Intangibles: A Metrics for the Economic Complexity of Countries and Products [**?**]; Economic complexity: Conceptual grounding of a new metrics for global competitiveness [**?**]; CompetitorsÕ communities and taxonomy of products according to export fluxes [**?**]

The model proposed by Caldarelli et al. [**?**] is a sort of two-dimensional PageRank algorithm [**?**]. Instead of jumping from web page to webpage, the random walker goes from country to products, and from products to countries.

We then applied the most general implementation of the **FQ** algorithm as developed for modelling the economy and competitiveness of countries. The **FQ** is a nonlinear generalization of the Hidalgo Hausman "Reflections Method". [**?**]. The algorithm has both a stochastic, iterative implementation, and an analytic solution. We demonstrate the iterative solution, to gain some intuition for the algorithm.

$$\begin{cases} w_c^* = A(\sum_{p=1}^{N_p} M_{cp} k_p^{-\alpha}) k_c^{-\beta} \\ w_p^* = B(\sum_{c=1}^{N_c} M_{cp} k_c^{-\beta}) k_p^{-\alpha} \end{cases} \quad (1)$$

At each iterative step we simultaneously rank editor "fitness", and article "ubiquity". In the linear model, the first iteration of "fitness" is the sum of articles to which that editor has contributed, and the "ubiquity" is the sum of editor who have contributed to that article. In the second iteration, say a user is as fit as the average the ubiquities of the of the pages edited. But this is all things being equal.

In the economic domain, the best products are those that are made by the fewest countries. Therefore in our average we want to give more weight to those best producing countries. This measure of good contributors being more important to success, is measured by alpha. A higher alpha means that a good product needs to be exported by the best countries. In Caldarelli, to correlate best with GDP rankings alpha = 1.5 Our result we find the opposite - negative values of alpha. in the not competitive but collaborative wikipedia, where the best articles are produced by the highest number of editors

3. METHOD

While it relies on the same principles as in [**?**], our proposed model is conceptually different : our "products" are Wikipedia articles, which are not the basis for competitions but rather for cooperation. Editors enrich the articles together to make best articles. However, like countries, editors have limited capabilities and limited resources (e.g., time), which force them make choices on their contributions.

The matrix **M** shown in Figure 1 shows when an article has changed at some point by a given editor. The matrix is ordered on both dimensions by decreasing order of editors who have changed more articles (vertical axis) and by decreasing order of articles that have been changed by most editors (horizontal axis) for a category of Wikipedia articles (here Feminist Writers). Although it is a rough count, the matrix tells already about the experience of an editor in

Figure 2: Convergence

the given category, and the attention an article has gotten from editors, which is an implicit quality measure according to the second principle of peer-production : peer-review. This count is the zero order of the Article/Editor ranking algorithm, and thus the initiation step is given by

$$\begin{cases} w_c^* = \\ w_p^* = \end{cases} \quad (2)$$

Now, let's consider the second step : if an article has been changed by editors who edited more articles, then the quality of the article should be higher. Similarly, if an editor has edited articles that have been edited by more editors, then the expertise of the editor should be higher (the only reason for making this claim is the collaborative nature of Wikipedia, and learning by imitation). Accordingly, the third step is the following : if an article has been changed by editors who edited more articles that have edited by more editors, then the quality of the article should be higher. Similarly, if an editor has edited articles that have been edited by more editors that have edited more articles, then the expertise of the editor should be higher. The algorithm goes on recursively, incorporating the quality (resp. expertise) information of the article (resp. editor) at the previous step.

tell about the stochastic process here [**?**]

say that it is a ranking algorithm. We care only about the ranking of articles and editors

The algorithm at step n then writes,

$$\begin{cases} w_c^* = \\ w_p^* = \end{cases} \quad (3)$$

As shown on Figure 2, the algorithm converges in a non-trivial way **(explanation for this?)**. In the iterative solution we see how certain editors start low, but then climb in rankings. This means that they are editing few articles, but those articles are of higher quality. Likewise certain articles climb over iterations, they are edited by relatively few editors, but those editors are fitter.

and an analytical solution can be found :

$$\begin{cases} w_c^* = A(\sum_{p=1}^{N_p} M_{cp} k_p^{-\alpha}) k_c^{-\beta} \\ w_p^* = B(\sum_{c=1}^{N_c} M_{cp} k_c^{-\beta}) k_p^{-\alpha} \end{cases} \quad (4)$$

that we use onwards.

4. DATA

The main input of the Article/Editors ranking algorithm is the matrix M which determines, for a category on Wikipedia,

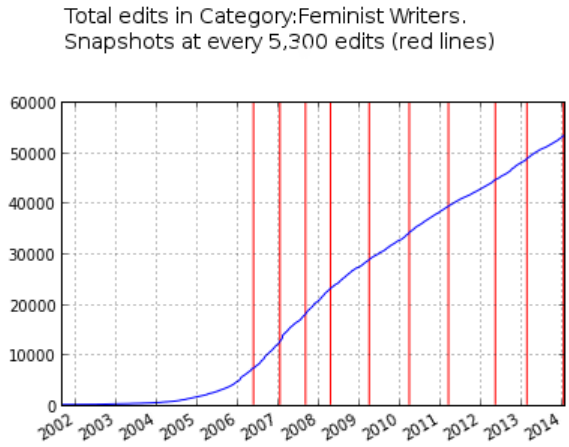

Figure 3: cumulative snapshots Feminist Writers eps

which articles have modified at least once by each editor. We collected contribution data for articles in 10 categories (c.f. table 4 for summary statistics on the categories). In addition, we made 10 snapshots of equal number of edits to account for the evolution over time of each category (see Figure **??**).

Category	Users	Articles	Edits
2013 films	5215	1896	150956
American male novelists	9946	2460	224783
American women novelists	5968	1936	138716
Bicycle parts	210	70	4981
Computability theory	272	92	7117
Counterculture festivals	578	66	10515
Economic theories	1145	212	28658
Feminist writers	1357	233	25738
Military history of the US	854	180	20172
Nobel Peace Prize laureates	4165	104	91522
Sexual acts	2190	93	45901
Yoga	730	123	25315

For each snapshot, we constructed the matrix $\mathbf{M_{e,a}}$ of contributors versus edited articles, similar to the country versus products matrix of the Economics Domain. For each snapshot, the values in $\mathbf{M_{e,a}}$ are defined as the number of edits made by editor e on article a in the category occurring in the snapshot time. Note that the final snapshot represents the entire history of the category up to the present date.

The matrix $\hat{\mathbf{M}}_{\mathbf{e,a}}$ is a binary representation of $\mathbf{M_{e,a}}$ where each nonzero entry is replaced with 1. This represents if editors have touched which articles rather than how much they have touched each article. In the economics domain, the distinction of making a binary matrix out of the data is interpreted an alternative metric to GDP per capita, rather than GDP. Here we could also see the distinction as normalized editor fitness. It has a typical triangular structure as shown on Figure **??**. The matrix $\hat{\mathbf{M}}_{\mathbf{e,a}}$ constitutes the basic input for implementing the Biased Markov Chain Approach, which we will call the $\mathbf{w}^*$ algorithm, which is an analytic solution to the iterative $\mathbf{W}$ algorithm. [**?**]

The exogenous metric for editors v_e we take is $labourhours$. For each editor the contribution history upto the snapshot point, divide into strings of $editsessions$, edits that occur within 1 hour of the previous edit. Then $labourhours$ are determined by subtracting the looking at the total time between the first and last edit in each edit session, and then summing the labours of each edit session. [**?**, **?**].

For an exogenous measure of article quality, v_a, we use a group of 5 text analysis metrics performed on Wikipedia articles at the lastest time in the snapshot. These are ratio of mark-up to readable text, number of headings, article length, citations per article length, and outgoing intrawiki links. To reduce the dimensionality of these 5 metrics, we perform Principal Component Analysis, and accept the principal component. Variance explained by the first principal component, was as high as .7 and never below .5 http://www-users.cs.umn.edu/ morten/publications/wikisym2013-tellmemore.pdf, http://mailer.fsu.edu/ bstvilia/papers/quantWiki.pdf [**?**].

4.1 Interpretation of w* algorithm in the context of open collaboration

To understand the results, we must have a firm grasp on what α and β mean. They are more easily understood by roughly rewriting $\mathbf{w}^*$ as:

$$w_e^* \sim k_e^{1-\beta} \langle k_a^{-\alpha} \rangle_e$$

$$w_a^* \sim k_a^{1-\alpha} \langle k_e^{-\beta} \rangle_a$$

where $\langle k_a^{-\alpha} \rangle_e$ is the arithmetic average of k $k_a^{-\alpha}$.

Since we see beta and alpha are close to being additive inverses, we can just study what it means for them to increase and decrease.

For Editors. As beta approaches one from infinity then editors are less judge by their the amount of contributions and more about the quality of the articles they contribute to. As beta becomes more negative below one, then the amount of articles is more important in predicting success. So 'gnomier' editors are more successful in this case. So we can see beta as a style marker for a category.

The lower beta, the more a diversified editor will be successful in a Category, and the higher beta, the more a targeted quality-writing author will be successful.

The output of $\mathbf{w}^*$ are a pair of rankings w_e^* and w_a^* for editors and articles respectively.

Figure 4: Matrix M ordered by decreasing order of edits on both contributors and articles dimensions.

$$\begin{cases} w_e^* = A(\sum_{a=1}^{N_a} M_{e,a} k_a^{-\alpha}) k_e^{-\beta} \\ w_a^* = B(\sum_{e=1}^{N_e} M_{e,a} k_e^{-\beta}) k_a^{-\alpha} \end{cases} \quad (5)$$

Next we collect exogenous metrics as comparison for both w_e^* and w_a^*, which we call v_e and v_a.

It is important to note that we are not competing with these metrics, but take them as state-of-the-art, Grand Truth, to which we calibrate against. We adopt a "less is more" approach. We using the exogenous variable to proxy

Yes, indeed for authors, what we are measuring is a weaker proxy for the gnomieness.

but for artilces the exogenous metric is quite, good, i.e. similar to what we are measuring

4.2 Testing against Wikipedia Metrics

We turn to the evaluation of the method against metrics

-
-

Having our endogenous and exegenous variables now, we perform a recursive grid search over the two dimensions of α and β to find a maximum correlation between our rankings from $\mathbf{w}^*$ and our exogenous variables. Our grid search operates on the interval $[-5, 5]$ with a resolution of 0.2 in on each axis. Importantly we search for negative values of α and β, which is not done in the Economics Domain. t

4.3 Finding trends

5. While the implementation presented here is strictly similar to **??**, the interpretation is slightly different in the context of group collaboration. Indeed, while countries competes for selling products, the hypothesis here is that Wikipedia contributors cooperate, at least in a very informal way, for improving the quality of articles.

5. RESULTS

6. DISCUSSION

6. *Refer to problems here, if any.*

7. CONCLUSIONS

Having our endogenous and exogenous variables now, we perform a recursive grid search over the two dimensions of α and β to find a maximum correlation between our rankings from $\mathbf{w}^*$ and our exogenous variables. Our grid search operates on the interval $[-5, 5]$ with a resolution of 0.2 in on each axis.

We also use a maximizing algorithm to find the values of α and β which maximize the spearman ρ rank correlation between our endogenous and exogenous rankings. This is performed for both of internal users ranks versus labour hours

Figure 5: Landscape

and internal article ranks and aggregated actionable article metrics.

Importantly we search for negative values of α and β, which is not done in the Economics Domain.

7.1 Finding Trends over Snapshots

The calibration technique for a category is performed at each of the ten points in the snapshot history. This allows us to track trends of ρ, α, and β over time.

5. While the implementation presented here is strictly similar to, the interpretation is slightly different in the context of group collaboration. Indeed, while countries competes for selling products, the hypothesis here is that Wikipedia contributors cooperate, at least in a very informal way, for improving the quality of articles.

8. RESULTS

8.1 high Correlations with Exogenous Variable

THIS IS EVEN BETTER THAN HH and by the way we have an explanation. read on to find out.

2.Once we have fitted alpha and beta we achieve high correlations with exogenous variables. We our correllation rho on users ranges from x to y. And on articles its better going from w, to v. These are quite high, and means that our new editor and article metrics are related to the state-of-the-art metrics that exist at the moment on Wikipedia. This gives w^* a basis as an alternative metric.

Also from snapshotting we see that rho increases over time, sometimes as much at 70% from 2006 to 2014. This means that w^* benefits from incorporating more contribution history. Is this true for both articles and users?

What do we say about the fact that user correlation is worse? -That infact editor fitness is not related to hours investment that much? -That in the "alternative economy" we are less concerned anyway about measuring users because it all disappears in the collaborative approach to making artilces.

Although rho is stable and high, α and β vary a lot between categories, and overtime within a category. α is a measure of how important it is for quality that many users edit, with lower alpha, we have a more collaborative category, where edits are more equal and egalitarean. With alpha high, the category is more being rewarding users that operate more individualistically. Beta is inversely related to alpha so the same can be said but the directions of the arguments reversed. This means we can talk of the characteristic of a category, compared to one another and compared over time.

8.2 Negative values of alpha and beta

This is unexpected, and different from HH, but show anti-competitiveness. It means the contributions of less fit authors are important.

Figure 6: rtime

Figure 7: timeseries or maybe table

8.3 Controversial Categories

We inspeceted controversial categories. And found alphas and betas that show...

8.4 Maybe "finding trends" subsection goes here.

We inspect the ρ maximum achievable Spearman rank correlation between w_e^* versus v_e and w_a^* versus v_a for any values of α and β over our snapshots. Behaviour of ρ over time differ when considering editors, or articles. For editors, in all categories we see a trend of ρ increasing over time. While the values of ρ for articles are higher much higher at earlier points in the snapshots, they do not clearly increase over time, and even in some categories we see a small loss of predictive power. Since our exogenous metrics for editors and articles are different, the difference in the trend of correlation could be affected either by $w*_e$ evolving over time differently than $w*_a$, or the v_e evolves over time differently than v_a, we did not manage to disentangle this problem.

8.5 correlation of alpha and beta as a single paramter metric for categories

We find that the maximizing values of α and β are strongly anti-correlated. In particular for articles, the pearson correlation coefficient never drops is never below 0.8 for any category. For editors, the correlation is below 0.8 in three categories, and tend towards zero. The reduction in α-β anti-correlation, is itself correlated to the size of the category as determined by the number of edits, 0.65 for editors, and 0.85 for articles.

This would mean this in that for sufficiently small categories, approximately those with less 100,000 edits, or approaximately 1,500 articles in our data, we can make the subsitituion $\alpha = -\beta$

This would mean that:

For editors Since alpha beta are correlated high beta -> high gamma low beta -> low gamma high gamma -> best editors are less gnomie, more specialized low gamma -> best editors are more gnomie, less specialized

For articles, as beta decreases from one and alpha increases then, the fitness of users become more important than the number of users editing. This is like, when collaboration fails because people are taking ownership. Lower alpha, and negative alpha mean more edits are more important to the success of an article.

high beta/low alpha -> high gamma low beta/high alpha -> low gamma high gamma -> best articles characterized by high quantity of contributor low gamma -> best articles charachterized by high quality of contributors

Most categories fluctuate in this measure over time. Sometime there are spikes in gamma to the positive end. This would indicate??

If we look at articles in 2013 films, we do see a clear trend. Gamma starts high and ends low. That means that high quality articles shift from being edited from many people to being edited by few higher quality contributors.

It seems natural to say that α, the importance of editor article quality, and β the importance of editor quality would be related. For articles this would mean that they lie on a spectrum from edited by fewer specialized editors to many diverse editors. For editors, this would mean that editors lie on spectrum of editing many ubiquitous articles to fewer obscure articles. Let us note that this does not necessarily have to be the case. It could be that the best articles are edit both by many people and also by many specialized editors, rather than "either-or". And likewise it could be the case that the best editors edit both many ubiquitous, and many obscure articles."

And in fact, gamma becomes less of a good assumption as the category size grows.

Except that α, β correlation is itsself correlated to the category size.

8.6 Ranking Evolution

We have three options for interpreting the evolution of ranking over time. First, what we observe is genuine to the category, that our results are specific to the categories we plot. Second, that this is a phenomena related to Wikipedia, or more generally online collaboration platforms. And third is that this an artefact of the properties of ranking.

This paper [**?**] shows that it is not inherent to ranking, so at least it is peculiar to Wikipedia.

We have to design a small test, for instance based on hamming distance or similar, to account for the properties of ranking.

Super Users persist over time, see the band of

Improved correlations when remove bots. Higher correllations with binary matrices. Implies that edit counts are a counter productive and

3. Highest correlations are when we have negative alpha and beta, which is about being anticompetetive.

9. DISCUSSION

A measure of controversy. Lately there have been conflicts in what categories represent. [**?**]

And that the super-users contribute more to sexism.[**?**]

Now we have measures of importance of superuser-contributors.

10. LIMITATIONS

We do not directly measure the *capabilities* of editors. For instance, what does one editor do best to improve an ar-

ticle, among the five metrics (ratio of mark-up to readable text, number of headings, article length, citations per article length, and outgoing intrawiki links) we have used to assess the quality of an article ?

We have not looked at the evolution of the quality of correlation as a function of iteration steps.

“less is more” : Why is it that when we incorporate more information in the matrix M, the results are not better ?

Quality of the exogenous metrics, especially for editors. It would be definitely make to look at real capabilities.

We only look at the ranking not at the real quality/expertise values ? Can we learn more the real values about the gap between articles/editors ?

Conclusions go here