Understanding the Relationship between Scholars' Breadth of Research and Scientific Impact

Shiyan Yan, University of Michigan, Ann Arbor
Carl Lagoze, University of Michigan, Ann Arbor

Abstract
Many existing metrics to evaluate scholars only concern their scientific impact and omit the importance of breadth of research. In this poster, we define a new metric for breadth of research based on the generalized Stirling metric which considers many aspects of breadth of research and satisfies several axioms for breadth of metrics. Also experiments on the ACM dataset show weak correlation between breadth of research measured by our new metric and scientific impact. And the variation of our metric over time illustrates a possible publication pattern for scholars
Keywords: breadth of Research; scientific impact; evaluation system
Citation: Yan, S., Lagoze, C. (2015). Understanding the Relationship between Scholars' Breadth of Research and Scientific Impact. In *iConference 2015 Proceedings*.
Acknowledgements: This research is funded by NSF 1258891 EAGER: Collaborative Research: Scientific Collaboration in Time
Research Data:
Contact: shiyansi@umich.edu, clagoze@umich.edu

1 Introduction

In Scientometrics, metrics like H-index and impact factor are straightforward. But they are relied on too heavily and do not adequately measure all aspects of scholarly impact[1]. Scientific influence of researchers is not limited to a single research community and can even cross disciplinary boundaries. A metric or a set of metrics is needed that accounts for breadth of scholars' research, so that breadth of research can be evaluated.

Another important problem of breadth of research is what is the effect of wide breadth of research on scholars' scientific impact. An empirical study of the relationship between breadth of research and scientific impact is needed. In this poster, we design a new metric based on the existing generalized Stirling metric, compares it to existing metrics and test its relationship to scientific impact.

2 Data and Methods

2.1 Data

The dataset used to extract breadth of research is from the ACM digital library. We select authors who publish at least five papers and crawl their citation numbers and H-indexes if their names are unambiguous on Google Scholar. Overall we crawled H-indexes and citation numbers for 8911 user profiles from Google Scholar in August 2014.

2.2 Methods

Two research problems in this poster are: First, how to measure the breadth of research for scholars; Second what's the relationship between breadth of research and scientific impact.

2.2.1 Breadth of Research Measurement

The first research problem has been studied by many scholars. There are many measurements of diversity or interdisciplinary, like entropy [2], Simpson's index [3] and generalized Stirling [4]:
Denote p_i as the distribution of authors' papers over $topic_i$, d_{ij} as the distance between $topic_i$ and $topic_j$.

$$Entropy = \sum_{i=1}^{n} -p_i \times \log(p_i)$$

$$Simpson = 1 - \sum_{i=1}^{n} {p_i}^2$$

$$Generalized\ Stirling = \sum_{i,j} d_{ij}^{\alpha}\ (p_i + p_j)^{\beta}$$

Among them, generalized Stirling considers not only the distribution of topics but also the similarity between topics. The farther the distance between topics where an author publishes papers is, the more diverse the author's research will be. Our new measurement is a modified version of generalized Stirling metric, defined as:
Denote d_{ij}, p_i as defined above, coh_i as the coherence of $topic_i$, which represents how close papers in the topic are:

$$Breadth\ of\ Research = \sum_{i,j} d_{ij}^{\alpha}\ (p_i + p_j)^{\beta}(Coh_i \times Coh_j)^{\gamma}$$

We modify the product of p_i and p_j in generalized Stirling to summation of p_i and p_j because the summation will give minor topics more chances to be counted into the measurement of breadth of research.

We add the coherence term into the metric because different topics have different "density" within themselves. Some topics like digital library are less cohesive topics because there are many diverse subtopics in these topics. But for topics like operation systems, researchers concentrate on several narrow subtopics. A researcher focusing on digital library should have larger breadth of research than operating systems researchers if other variables are controlled.

2.2.2 Dictionary Extraction

For calculation of p_i and d_{ij} we have to define topics and assign scholars into topics. We leverage a clustering algorithm to extract topics from papers. Before that we generate a dictionary of computer science used in the clustering.

Dictionary extraction follows these steps:
1. Extract bigrams and trigrams that occur frequently in papers
2. Extract grams from papers that conform to the pattern "grams (abbreviation)", e.g. machine learning (ML)
3. Combine the results of step 1 and step 2 (3816 terms)
4. Build a network of terms in Wikipedia through hyperlinks between different entries
5. Search terms related to grams in the dictionary of step 3 in the network and combine them with results of step 3 (6100 terms)

2.2.3 Topic Extraction and Assignment

After getting the dictionary, we count coocurrence times for every pairs of terms and calculate the similarity between different terms by:

$$Sim_{ij} = \log \frac{Cooccur_{ij} + 1}{Max(Cooccur_{ij}) + 2}$$

And we count coocurrences of terms in abstracts of papers more than those in full text because generally they have more topic signals. Using similarity matrix of terms, we run an unsupervised learning algorithm called Affinity Propagation[5] that cluster similar terms into same clusters and choose an exemplar for every cluster. Here are some clustering results:

Exemplar: digital library
Terms: citation analysis, citation index, community building, digital earth, digital library, digital library software, digital preservation, digital reference, discourse analysis, dublin core ...

Exemplar: machine learning
Terms: active learning, adaptive control, bayes classifier, belief propagation, clinical trial, computational learning theory, concept learning, conditional random field …

With the clusters of grams in computer science, we assign authors into different topics according to their papers. Every author will be represented by a word distribution over topics, which are used to calculate scores of metrics.

3 Results

3.1 Simulation Experiment

There is no common standard to decide which metric of breadth of research is better. We propose some axioms for these metrics. If they follow these axioms, they are considered to perform well:

Axiom1 Publish in New Topics: If an author publishes a paper in a new topic that he has never published in, his breadth of research should increase.

Axiom2 Publish in Old Topics: if an author publishes a paper in a topic where he has published many papers before, his breadth of research should decrease.

Axiom3 Publish in New Topics Twice: If an author publishes papers in two new topics in a sequence, the increase of breadth of research in the second time will be smaller than the increase of that in the first time.

Axiom4 Publish in Close Topics: If an author publishes a paper in a new topic close to the author's research interest, the improvement of his breadth of research should be less than that of publishing a new paper in a randomly chosen topic.

We implement four simulation experiments to test how these metrics follow axioms. The results are shown in table1.

	Entropy	Simpson's	GL Stirling α=7 β=1	New Metric α=7 β=1 γ=1
Axiom1	0.99	0.99	0.82	0.69
Axiom2	0.89	0.97	0.67	0.73
Axiom3	0.87	0.94	0.25	0.33
Axiom4	0	0	0.75	0.76

Table 1 Probability that metrics satisfying of the axioms

The results show that entropy and Simpson's perform well in the first three axioms because they don't consider distances between topics and introduce less noise. Because every new topic will be regarded equally for these metrics, they cannot follow Axiom4. Generalized Stirling and our metric perform acceptable in Axiom1 and Axiom2, but worse than entropy and Simpson's. They perform badly in Axiom3 because relatively bad performance on publishing a paper in new topic (Axiom2) will aggregate when testing the performance of publishing two papers in two new topics. But they perform well in Axiom4 because of the consideration of distances. Also we find our metric performs better than generalized Stirling in Axiom2 and Axiom3 ($p<0.001$ in proportion test), which means coherences of topics and greater weights on minor topics are beneficial when we consider variation of metrics when publishing in new topics.

3.2 Relationship between Breadth of Research and Scientific Impact

We also test the Pearson correlation between metrics of breadth of research and H-indexes of scholars. Some metrics have weak positive relationship with H-index. Others have weak negative relationship (Table 2). Because publication numbers may influence the correlation between breadth of research and scientific impact i.e. the increase of numbers of publications may bring increase of breadth of research and increase of H-index simultaneously to make them positively correlated to each other, we test the partial correlation between metrics of breadth of research to H-index controlling publication numbers (Table 2). They are weaker than Pearson correlations.

	Heading	Heading
Entropy v.s. H-index	-0.1722	-0.0769
Simpson's v.s. H-index	0.2102	0.0922
GL Stirling v.s. H-index	0.0415	0.0348
New Metric v.s. H-index	0.3828	0.1613

Table 2 Correlation between breadth of research and H-index

3.3 The Variation of Breadth of Research

We also draw a graph (Figure 1) shown average variation of metrics over publication years for scholars. Simpson's, generalized Stirling and our new metric will increase when publication numbers increase and keep stable after a long period of publications, which explains a possible publication pattern of scholars:

scholars' breadth of research increase with the increase of publications in the early stage of their career. But because of accumulation of publications, their accumulative breadth of research will not change dramatically in the late years.

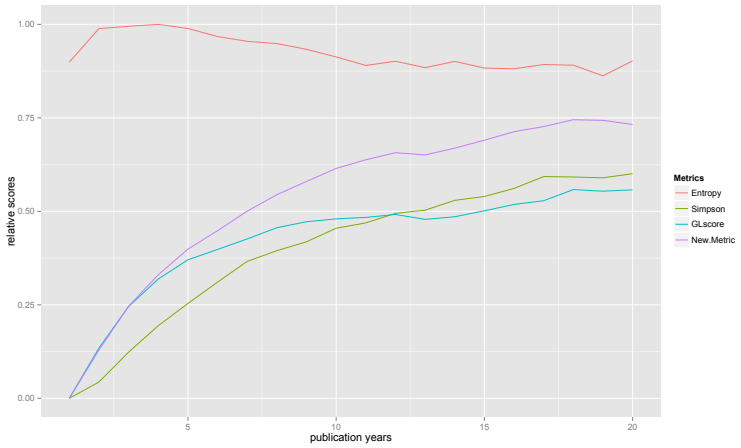

Figure 1 Variation of metrics over publication years

4 Conclusion

We design a new metric based on generalized Stirling to evaluate breadth of research for scholars in computer science. The new metric performs well in simulations of publishing papers in new topics compared to existing generalized Stirling, but not good enough in the simulation of publishing papers in familiar topics. The variation of this new metric over publication years shows a possible publication pattern of scholars. Also we find the correlation between breadth of research and scientific metrics are weak, especially when we control publication numbers.

References

1. Weingart P. Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*. 2005;62(1):117–131.

2. Weaver W, Weaver W. Recent Contributions to The Mathematical Theory of Communication 1 Introductory Note on the General Setting of the Analytical Communication Studies. 1949.

3. Simpson, E. H. (1949). Measurement of diversity. *Nature*.

4. Stirling A. A general framework for analysing diversity in science, technology and society. *J R Soc Interface*. 2007;4(15):707–19. doi:10.1098/rsif.2007.0213.

5. Frey BJ, Dueck D. Clustering by passing messages between data points. *Science*. 2007;315(5814):972–6. doi:10.1126/science.1136800.

Table of Figures

Table of Tables