A Comparative Study of Academic and Wikipedia Ranking

Xin Shuai
School of Informatics and Computing
Indiana University
Bloomington
IN, USA
xshuai@indiana.edu

Zhuoren Jiang
College of Transportation Management
Dalian Maritime University
Dalian, China
jzr1986@hotmail.com

Xiaozhong Liu
School of Library and Information Science
Indiana University
Bloomington
IN, USA
liu237@indiana.edu

Johan Bollen
School of Informatics and Computing
Indiana University
Bloomington
IN, USA
jbollen@indiana.edu

ABSTRACT

In addition to its broad popularity Wikipedia is also widely used for scholarly purposes. Many Wikipedia pages pertain to academic papers, scholars and topics providing a rich ecology for scholarly uses. Scholarly references and mentions on Wikipedia may thus shape the "societal impact" of a certain scholarly communication item, but it is not clear whether they shape actual "academic impact". In this paper we compare the impact of papers, scholars, and topics according to two different measures, namely scholarly citations and Wikipedia mentions. Our results show that academic and Wikipedia impact are positively correlated. Papers, authors, and topics that are mentioned on Wikipedia have higher academic impact than those are not mentioned. Our findings validate the hypothesis that Wikipedia can help assess the impact of scholarly publications and underpin relevance indicators for scholarly retrieval or recommendation systems.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

Keywords

Wikipedia; Scholar Impact; Citation Analysis

1. INTRODUCTION

Science Citation Index established by Garfield [2] makes citation statistics a gold standard for the assessment of scholarly impact. Citation data is held to be a valid and reliable indicator of scholarly impact because it represents an explicit and objective acknowledgement of influence by expert authors. Yet, the Web 2.0 is revolutionizing scholarly practices. A growing number of scholars discuss and share the research literature on Twitter and Facebook [6], organize it in social reference managers like Mendeley, and review it in blogs [5]. The increasing role of social media in scholarship requires new ways to assess impact beyond traditional approaches on the basis of citation data. Wikipedia, as a collaboratively edited, multilingual, and free Internet encyclopedia, has become an important source for the creation, distribution, and acquisition of scientific knowledge. Kittur et al. shows that over 25% of pre-2008 articles in Wikipedia are related to natural or social sciences [3].

Wikipedia editors frequently reference scholarly entities, such as papers, scholars, and topics. We refer to such mentions as "Wikipedia citation", implying that their value or influence has been explicitly recognized by the Wikipedia community. Unlike "academic citations" that represent the explicit recognition of expert scholars, the authority of a "Wikipedia citation" is uncertain and will need to be examined further.

Several studies have compared "academic citations" with "Wikipedia citations". Nielsen [8] showed that citations in Wikipedia correlate well with statistics from the Journal Citation Reports. Evans and Krauthammer [1] investigated this relationship at the journal article level and found that PubMed journal articles that are mentioned in Wikipedia have significantly higher academic citation counts than an equivalent random article subset. Although these findings show that Wikipedia citations are an indicator of academic impact, their results are limited to journals or articles published in the same journal.

Here we extend this line of work to larger-scale data and across a broader research area for a more diverse set of scholarly entities. This paper makes an effort to quantitatively compare the rankings of articles, authors, and topics selected from ACM Digital Library publication data on the basis of their academic citations and Wikipedia citations. Our major findings include:

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.

- Academic and Wikipedia rankings exhibit a positive correlation across all types of scholarly entities.
- Papers, authors, and topics that are mentioned on Wikipedia have a higher average academic impact than those that are not mentioned.
- Wikipedia mentions are biased towards high-impact scholars who publish many papers.
- Wikipedia mentions are biased towards trending topics that occur in many articles.

2. RELATED WORK

Studying the influence of social web on the scholarly community opens a new direction for scientometrics. Priem et al. propose scientometrics 2.0 [4] and Altmetrics [5] to measure scholarly impact from social media data. Xin et al. [6] compare different types of online responses to newly submitted preprint publications, namely article downloads, Twitter mentions and citations. Nielsen [8], Evans and Krauthammer [1] correlate journal citations to Wikipedia and academia. Our study extends this work by comparing the rankings of three different scholarly entities, i.e. papers, authors, and keywords, from a large-scale dataset in the field of Computing between scholarly sources and Wikipedia.

3. PROBLEM DEFINITION

The framework of our study is shown in Figure 1. P is a set of scientific papers, A is a set of authors, K is a set of keywords (topics) and W is a set of Wikipedia articles. $X = P \cup A \cup K$ is a set of heterogeneous scholarly entities.

[**Academic Citation**] AC includes paper citation $AC_p = \{(p_i, p_j)|p_i \ cites \ p_j, p_i, p_j \in P\}$, author citation $AC_a = \{(a_i, a_j)|a_i \ cites \ a_j, a_i, a_j \in A\}$ and keyword citation $AC_k = \{(k_i, k_j)|k_i \ cites \ k_j, k_i, k_j \in K\}$

[**Wikipedia Citation**] $WC = \{(w_i, x_i)|w_i \ mentions \ x_i, w_i \in W, x_i \in X\}$ represents the acknowledgement of the scholarly entity x_i from Wikipedia article w_i

[**Paper Citation Network**] $G_p = \{P, A, K, AC_p\}$ is a directed and unweighted heterogeneous network.

[**Author Citation Network**] $G_a = \{A, AC_p, F_a\}$ is a directed and weighted network derived from G_p, where F_a is a weight function that maps each edge (a_i, a_j) to an positive integer $f(a_i, a_j) \in \mathbb{N}^+$ that corresponds to the number of citations passing from a_i to a_j.

[**Keyword Citation Network**] $G_k = \{K, AC_k, F_k\}$ is a directed and weighted network derived from G_p, where F_k is a weight function that maps each edge (k_i, k_j) to a positive integer $f(k_i, k_j) \in \mathbb{N}^+$ that corresponds to the number of citations passing from k_i to k_j

[**Wikipedia Interlinking Network**] $G_w = \{W, E\}$ is a directed and unweighted network composed of Wikipedia articles and their internal links, where $(w_i, w_j) \in E$ indicates that w_i contains a hypertext linking to w_j.

[**Ranking Function**] $R(X, \Theta)$ maps $X = \{x_i\}$ into a sorted permutation X^* where $\Theta(x_i) \geq \Theta(x_j)$.

[**Academic Ranking**] $AR(X, \Theta)$ ranks X based on statistics calculated from ACM database, where $\Theta(x_i)$ can be Θ_{af} (the frequency of x_i), Θ_{ac} (the citation of x_i in G_x), or Θ_{ap} (the Pagerank of x_i in G_x).

[**Wikipedia Ranking**] $WR(X, \Theta)$ ranks X based on statistics calculated from Wikipedia, where $\Theta(x_i) = \sum_{w_j} S(w_j)$ satisfying that $(w_j, x_i) \in WC$. Especially, $S(w_j)$ is a score attached to each Wikipedia article, which can be constantly equal to one, or the total count of edits of w_j, or the Pagerank score of w_j in G_w. Therefore, we have three types of Θ for WR: Θ_{w1}, Θ_{we} and Θ_{wp}.

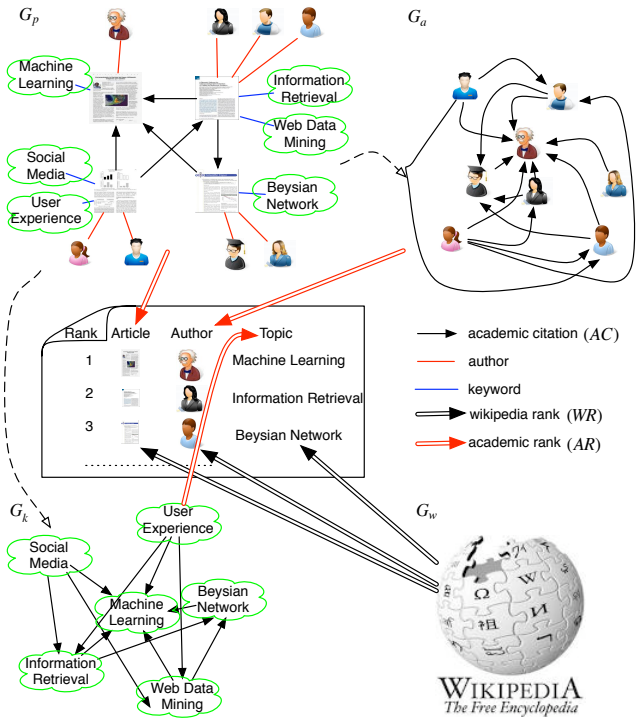

Figure 1: Academic ranking and Wikipedia ranking

The main goal of this study is to examine the relationship between $AR(X, \Theta)$ and $WR(X, \Theta)$. Specifically, we investigate three questions:

- What's the correlation between $AR(X)$ and $WR(X)$?
- How do different types of Θ affect above correlation?
- Are those scholarly entities mentioned in Wikipedia ranked higher in $AR(X)$ than those not mentioned?

4. METHODS

Two types of data sources, i.e. the ACM database and Wikipedia data, are involved in our study. Our analysis consists of three steps. First, we select a set of papers, authors, and keywords from the ACM database. Second, we build a Wikipedia search index and query it for these set papers, authors, and keywords. Third, we rank all papers, authors, and keywords based on $AR(X, \Theta)$ and $WR(X, \Theta)$, and compare the resulting rankings.

4.1 ACM data selection

The ACM database stores over 200K papers published in computer science journals and conferences, from 1980 to 2010. Most papers contain three types of metadata: a title, author, and keywords field. Each type of metadata needs careful filtering before we search for them in Wikipedia, since short queries (title, author names, keywords) may lead to false positives. We therefore decide to remove titles, authors, and keywords from our ACM database that match the properties in Table 1.

metadata	removed type
title	less than three non-stop words e.g. "On the Level", "Behavioral synthesis"
author	less than six tokens or one-letter first name e.g. "C. Richard" or "Li Li"
keyword	single word or occurs only once in database e.g. "program", i.e. "system initiated dialogue"

Table 1: Criteria to remove titles, authors, and keywords from our ACM database.

Once P, A, and K are selected, we construct G_p, G_a and G_k, based on which $AR(X, \Theta_{af})$, $AR(X, \Theta_{ac})$ and $AR(X, \Theta_{ap})$ are obtained.

4.2 Wikipedia search

The export data of the English version of Wikipedia is publicly available in XML format[1]. We use a Perl script *WikiPrep*[2] to parse this data and remove all discussion and talk pages, leaving only article pages. We further remove category, file, disambiguation, and redirect pages to make sure that each article page is about an explicit concept.

Since the to-be-ranked sets of papers, authors, and keywords come from the ACM database, the Wikipedia articles that are used to build the search index should match the topics of the ACM repository. Our preliminary tests showed that including Wikipedia articles that do not match ACM fields in the process of searching will enormously reduce the retrieval precision. Therefore, we adopt a two-step Wikipedia index building process on the assumption that the set of keywords K selected from ACM database are representative topics in computer science. First, all Wikipedia article pages are indexed using Lucene[3]. Second, we run a full-text search in the complete Wikipedia index for every keyword from K. In the retrieval results, only those Wikipedia articles containing at least two keywords are selected to construct G_w and build a computing-exclusive index named Wiki_{cs}. For each node in G_w, we compute the PageRank score and query its historical edit frequency using WikiMedia API[4]. Finally, we search P, A and K in Wiki_{cs} and generate $WR(X, \Theta_{w1})$, $WR(X, \Theta_{we})$ and $WR(X, \Theta_{wp})$.

4.3 Evaluation measures

When X is searched in Wiki_{cs}, only a subset of X is found in Wikipedia and denoted as X_{wiki}. The complement of this subset X is denoted X_{nowiki}. We propose two measures to perform two types of ranking comparisons: (1) the ranking of X_{wiki} in ACM and Wikipedia , (2) the ranking of X_{wiki} and X_{nowiki} in ACM.

First, we use Spearman's rank correlation coefficient [7] to measure the quantitative relation between $AR(X, \Theta)$ and $WR(X, \Theta)$:

$$\rho = \frac{\sum_i (I_{x_i} - \bar{I_{x_i}})(I_{y_i} - \bar{I_{y_i}})}{\sqrt{\sum_i (I_{x_i} - \bar{I_{x_i}})^2 (I_{y_i} - \bar{I_{y_i}})^2}}, x_i, y_i \in X_{\text{wiki}} \quad (1)$$

where I_{x_i} and I_{y_i} are the index of x_i and y_i in X^* ranked by $AR(X, \Theta)$ and $WR(X, \Theta)$, respectively, and

$$\bar{I_{x_i}} = \frac{\sum_{x_i \in X_{\text{wiki}}} I_{x_i}}{|X_{\text{wiki}}|}, \bar{I_{y_i}} = \frac{\sum_{y_i \in X_{\text{wiki}}} I_{y_i}}{|X_{\text{wiki}}|} \quad (2)$$

Next, we define a Normalized Average Ranking (NAR) to measure the average ranking positions of a subset of entities in the whole ranking list.

$$\text{NAR}(X_{\text{sub}}) = \frac{\bar{I_{x_i}}}{\max(I_{x_j})}, x_i \in X_{\text{sub}}, x_j \in X, X_{\text{sub}} \subseteq X \quad (3)$$

where, X_{sub} can be either X_{wiki} or X_{nowiki}, and $\max(I_{x_j})$ is the maximum ranking index in X

5. RESULTS AND DISCUSSION

We obtain 122,350 papers, 163,172 authors[5] and 35,518 keywords from the ACM database. In addition, we select 357,345 Wikipedia articles from the Sep 2, 2012 Wikipedia dump, and build the search index Wiki_{cs}. After querying all selected papers, authors, and keywords in Wiki_{cs}[6], we find 3836 papers, 17564 authors and 20891 keywords that are mentioned on Wikipedia. We can see that only a very small portion of papers (0.03) and authors (0.1) but more than half of keywords occur in Wikipedia.

The statistics of G_p, G_a, G_k and G_w are shown in Table 2.

measure	G_p	G_a	G_k	G_w
num of nodes	122,350	163,172	35,518	357,345
num of edges	395,152	2,877,723	1,746,734	17,947,480
density	2.64-e^5	0.0001	0.0014	0.0001
clustering	0.1326	0.2754	0.4784	0.2903

Table 2: Descriptive statistics of paper, author, keywords, and Wikipedia citation networks.

The Spearman rank correlations between $AR(X)$ and $WR(X)$ using different Θ, i.e. Θ_{af}, Θ_{ap} and Θ_{ac}, for $AR(X, \Theta)$ and Θ_{w1}, Θ_{wp} and Θ_{we} for $WR(X, \Theta)$) are shown in Table 3.

The overall scholarly ranking and Wikipedia ranking are positively correlated. The p-values for all correlations are less than 0.001, hence we can reject the null-hypothesis that the two variables are not correlated. However, none of correlation coefficients are greater than 0.5, indicating significant differences between the scholarly community and the community of Wikipedia editors in recognizing the impact or importance of scholarly entities.

We note several interesting results. First, the inclusion of Wikipedia article weight does not improve the ranking correlation. Wikipedia ranking using Θ_{w1} consistently shows

[1]http://dumps.wikimedia.org/enwiki/
[2]http://www.cs.technion.ac.il/ gabr/resources/code/wikiprep/
[3]http://lucene.apache.org/
[4]http://www.mediawiki.org/wiki/API:Query

[5]We do not consider name ambiguity here and leave it for future work
[6]Titles and keywords use exact match, and authors use fuzzy match

	$WR(X, \Theta_{w1})$	$WR(X, \Theta_{wp})$	$WR(X, \Theta_{we})$
Paper			
$AR(P, \Theta_{ap})$	**0.1716**	0.1342	0.0913
$AR(P, \Theta_{ac})$	0.1652	0.1316	0.0863
Author			
$AR(A, \Theta_{af})$	0.2159	0.1319	0.0635
$AR(A, \Theta_{ap})$	**0.2521**	0.1753	0.0882
$AR(A, \Theta_{ac})$	0.2422	0.1684	0.0811
Keyword			
$AR(K, \Theta_{af})$	**0.4638**	0.3818	0.3478
$AR(K, \Theta_{ap})$	0.3235	0.2646	0.2340
$AR(K, \Theta_{ac})$	0.3311	0.2714	0.2437

Table 3: The Spearman rank-order correlation between scholarly ranking and Wikipedia ranking using different ranking criteria. The academic ranking is based on frequency of occurrence (Θ_{af}), citation PageRank (Θ_{ap}), and citation frequency (cited) (Θ_{ac}). The Wikipedia ranking takes into account Wikipedia page weight, including one (Θ_{w1}), editing frequency (Θ_{we}) and inter-linking PageRank (Θ_{wp}).

better correlation with the academic ranking than Θ_{wp} and Θ_{we}. It seems that "Authority" of Wikpedia page cannot provide information to better weight the importance of scholarly outcomes[7]. Second, Wikipedia pages tend to cite scholars of high scholarly reputation more than those who are "merely" productive. Θ_{af} tends to rank productive authors higher while Θ_{ap} and Θ_{ac} tend to rank the most cited authors higher. Ranking with Θ_{ap} shows higher correlation with Wikipedia ranking than Θ_{af} and Θ_{ac}, implying that reputable authors who published influential works are more favored by Wikipedia community than those that are new and productive. Third, Wikipedia tends to mention trending research topics more than classical or traditional topics. The situation of keyword ranking is different than author ranking, i.e. frequently occurring keywords are more likely to be mentioned in Wikipedia than most cited keywords ("classical" topics). It implies that Wikipedia editors prefer to discuss trending scientific topics when they create new, or edit existing Wikipedia pages.

	$AR(*, \Theta_{af})$	$AR(*, \Theta_{ap})$	$AR(*, \Theta_{ac})$
Paper			
P_{wiki}	N/A	0.3912	0.3316
P_{nowiki}	N/A	0.7576	0.7063
Author			
A_{wiki}	0.5307	0.4808	0.4362
A_{nowiki}	0.8185	0.7731	0.7290
Keyword			
K_{wiki}	0.3274	0.5369	0.5034
K_{nowiki}	0.5068	0.7125	0.6731

Table 4: The average academic ranking comparison between Wikipedia mentioned (X_{wiki}) and non-mentioned (X_{nowiki}) scholarly entities

Table 4 shows the differences in academic ranking between scholarly entities that are mentioned and those that are not mentioned in Wikipedia. It is apparent that papers, authors, and keywords that are mentioned on Wikipedia are ranked higher in the scholarly community than those are not mentioned. We use t-test to examine the differences and all corresponding p-values (consistently smaller than 0.001) confirm the statistical significance. It indicates that Wikipedia is a good social filtering system that recommends high impact papers, authors, and topics to the public.

[7]It is also possible that we did not find the correct way to measure Wikipedia article importance.

6. CONCLUSIONS

This paper compares the scholarly ranking and Wikipedia ranking of a set of selected papers, authors, and keywords (topics), from the field of Computer science. We run a full-text search of those papers, authors, and keywords in Wikipedia, and separate the sets into two subsets: X_{wiki} denotes the set of papers, authors, and keywords that are mentioned in Wikipedia, while X_{nowiki} represents the set that are not mentioned. First, we compute the Spearman rank-order correlation coefficient between scholarly and Wikipedia rankings of X_{wiki} using different ranking methods. We find that the two rankings are statistically significantly correlated and that the Wikipedia community favors reputable authors and trending topics. Second, we compare the average ranking of X_{wiki} and X_{nowiki} and found the former had much higher values than the latter. This implies that Wikipedia does serve as a collaborative social filtering system which is able to favor "classical" papers, authors, and topics, and recommend them to the general public.

7. ACKNOWLEDGMENTS

One of the authors Xin Shuai thanks the National Science Foundation for its support of his work under grant SBE #0914939

8. REFERENCES

[1] P. Evans and M. Krauthammer. Exploring the use of social media to measure journal article impact. *AMIA Symposium*, 2011(January):374–81, 2011.

[2] E. Garfield and R. K. Merton. Citation indexing-its theory and application in science, technology, and humanities. *garfield.library.upenn.edu*, Jan. 1979.

[3] A. Kittur, E. H. Chi, and B. Suh. What's in wikipedia?: mapping topics and conflict using socially annotated category structure. CHI '09, pages 1509–1512, New York, NY, USA, 2009. ACM.

[4] J. Priem and B. H. Hemminger. Scientometrics 2.0: New metrics of scholarly impact on the social web. *First Monday; Volume 15, Number 7 - 5 July 2010*, 2010.

[5] J. Priem, H. A. Piwowar, and B. M. Hemminger. Altmetrics in the wild: Using social media to explore scholarly impact. *ArXiv e-prints*, Mar. 2012.

[6] X. Shuai, A. Pepe, and J. Bollen. How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations. *PloS ONE*, 11(7), 2012.

[7] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3-4):441–471, 1987.

[8] F. Ąrup Nielsen. Scientific citations in wikipedia. *First Monday*, 12(8), 2007.