Mining Google Scholar Citations: An Exploratory Study

Ze Huang and Bo Yuan

Intelligent Computing Lab, Division of Informatics,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, P.R. China
workthy@hotmail.com, yuanb@sz.tsinghua.edu.cn

Abstract. The official launch of Google Scholar Citations in 2011 opens a new horizon for analyzing the citations of individual researchers with unprecedented convenience and accuracy. This paper presents one of the first exploratory studies based on the data provided by Google Scholar Citations. More specifically, we conduct a series of investigations on: i) the overall citation patterns across different disciplines; ii) the correlation among various index metrics; iii) the personal citation patterns of researchers; iv) the transformation of research topics over time. Our results suggest that Google Scholar Citations is a powerful data source for citation analysis and provides a solid basis for performing more sophisticated data mining research in the future.

Keywords: Google Scholar Citations, Citation Analysis, Tag Cloud, Clustering

1 Introduction

Citation analysis refers to the investigation of the frequency and patterns of citation records (i.e., references to published or unpublished sources) in scholarly literature. It has been widely used as a method of bibliometrics to evaluate the quality of journals and to establish the links among works and authors. For example, it is possible to identify groups of people that collaborate frequently or how a piece of work is related to existing studies [1]. It is also important for researchers to choose the right journals as well as track the development of specific research topics [2]. In many academic institutes, citation record is also being used as one of the major selection criteria in the process of recruiting and promotion.

There have been some critics [3，4] on the potential misinterpretation of citation in evaluating journals and researchers: i) the number of citations can be manipulated to deliberately increase the impact of a journal; ii) the influence of self-citation and negative citation needs to be taken into account; iii) it may be difficult to find a good tradeoff between the number of citations and the number of publications; iv) different research disciplines may have significantly different typical citation numbers. Nevertheless, citation record provides a practical and quantitative performance measure, which has been accepted widely in academia.

Online bibliographic databases such as Web of Science (published by Thomson ISI), SciVerse Scopus (published by Elsevier) and Google Scholar (a freely accessible web search engine released in 2004 by Google) have brought tremendous benefits to researchers across different disciplines [5]. In nowadays, Web of Science covers over 12,000 journals and 150,000 conference proceedings but its most famous citation index Science Citation Index Expanded only covers around 8,000 journals. In the meantime, SciVerse Scopus contains nearly 19,500 titles from 5,000 publishers worldwide with 46 million records most of which are journal articles.

Different from subscription-based commercial databases, Google Scholar retrieves bibliographic data of academic literature by automatically crawling over the Web and uses a ranking algorithm, which relies heavily on citation counts, to display the search results. Although its exact coverage is not known to the public (some publishers may not allow Google Scholar to crawl their databases), some studies show that Google Scholar usually returns the highest number of citations compared to other similar services [6, 7]. Note that conference papers are extensively indexed in Google Scholar and their citation counts are calculated in combination with journal articles. This is particularly important for disciplines such as computer science where many high quality research outcomes are published in conferences.

Although most bibliographic databases provide comprehensive search functions, there is an inherent issue making the accurate evaluation of individual researchers a very challenging task: the disambiguation of authors. In many occasions, different researchers share exactly the same name (e.g., even different names may appear to be identical in terms of spelling when translated into English) and it is necessary to group the publications corresponding to the same author before doing any further analysis. Despite of some recent progress in this area, it is still not a fully reliable procedure [8]. After all, a researcher may have different affiliations and collaborate with different people and publish papers in seemingly irrelevant disciplines.

The official launch of a new service in the name of Google Scholar Citations[1] (GSC) in late 2011 provides a different solution to the above issue. Built on the top of Google Scholar, it allows researchers to create personal accounts and add, manually or automatically, papers published by them. By doing so, each registered researcher has a mini-homepage (similar to a blog) with a list of papers and citation counts. In other words, GSC gives researchers the freedom to maintain the list of publications to ensure the highest accuracy and integrity of data. For example, the bibliographic information of each paper is user-editable, which means that errors accidentally introduced during the crawling of web pages can be corrected in a straightforward manner. Also, similar to many social networking services, researchers can follow new articles and citations of any registered researchers.

In this paper, we present one of the first exploratory studies on the analysis of the structured data in GSC. We will investigate: i) the overall citation patterns across different disciplines; ii) the correlation among various index metrics; iii) the personal citation patterns of researchers; iv) the transformation of research topics over time. Section 2 describes the data collection procedure including the content structure of GSC. Section 3 presents the major results of citation analysis while Section 4 focuses on the analysis of topic trends. This paper is concluded in Section 5 with some discussion and directions for future work.

[1] http://scholar.google.com/citations

2 Data in GSC

In GSC, researchers can manually label their research disciplines. In many cases, each registered researcher has three to four discipline labels. To explain how we collected the data, all web pages in GSC are divided into two types in this paper: *Discipline Level* (DL) pages and *Author Level* (AL) pages.

2.1 Web Page Description

The major information in GSC is shown in Table 1. DL pages display authors in a certain discipline (e.g., data mining). For example, with "label: data_mining" as the keyword, GSC returns a name list of 10 authors who have identified themselves as in the *data mining* discipline, sorted based on the total citation number in descending order. By clicking the "Next" link, the next 10 authors (if any) will be displayed.

Each author in the list has an URL linking to his/her personal page (AL page), which shows detailed publication information. The AL page can be divided into 3 sections from top to bottom: i) author profile; ii) citation indexes table; iii) a list of papers that the author has published, sorted by each paper's citation number in descending order. The paper list shows maximally 20 or 100 papers and additional papers (if any) may be accessed by clicking the "Next" link. This action was simulated in our program to gather the information of all papers corresponding to an author.

Table 1. The content description of two types of web pages.

Page Type	*Content*	*Details*
DL Page	Author List	URL links to each author's personal page
AL Page	Author Profile	Affiliation, Disciplines, Home Page
	Citation Index Table	Citations, h-index, i10-index (All & Recent)
	Paper List	Title, Author, Year, Citation Number

Note that, the citation indexes table has two columns. The first column consists of statistical values based on an author's entire publication records, and the other one is based on recent papers published within 5 years (e.g., since 2007 as of 2012).

2.2 Data Collection

Since GSC does not provide APIs to the public, we collected the data by analyzing the web page source code with a crawler program. The crawler extracted required information through pattern match using regular expression.

For DL pages, we focused on 6 related disciplines: Data Mining (DM), Artificial Intelligence (AI), Bioinformatics (Bio), Information Retrieval (IR), Machine Learning (ML) and Pattern Recognition (PR). So far, many disciplines in GSC did not have sufficient number of registered authors (e.g., 200 authors in Sociology). Note that the same author may appear in multiple disciplines and authors whose papers had zero citation were not counted. For AL pages, the 6 indexes in the citation index table and publication information (title, year of publication and citation count) were retrieved. Papers without any citation were excluded.

Totally, we collected up to 1000 authors in each discipline and no more than 100 papers for each author. It took less than 30 minutes for the crawler to collect the data, and the dataset used in the following experiments was collected on March 11, 2012.

3 Citation Analysis

3.1 Index Metrics

The number of total citations is a most commonly used index metric to quantify the impact of an individual's research output. However, it is far from sufficient to compare and evaluate research work comprehensively. Recently, many new metrics were designed and enhanced, such as the h-index [9] and the g-index [10]. In GSC, only three metrics are adopted: the total number of citations (TC), the h-index and the i10-index. The h-index attempts to address both the productivity and the impact factors and is defined as the maximum number h so that there are h papers each with citation number $\geqslant h$. The i10-index was introduced in July 2011, which indicates the number of academic papers of an author that have received at least 10 citations.

Fig. 1 shows the TC values of the top 30 scholars in each of the 6 disciplines. It is clear that researchers especially the most eminent ones in AI tend to have much larger TC values compared to disciplines such as IR and PR. This fact suggests that TC is not a reliable metric for evaluating the impact of individuals in different disciplines.

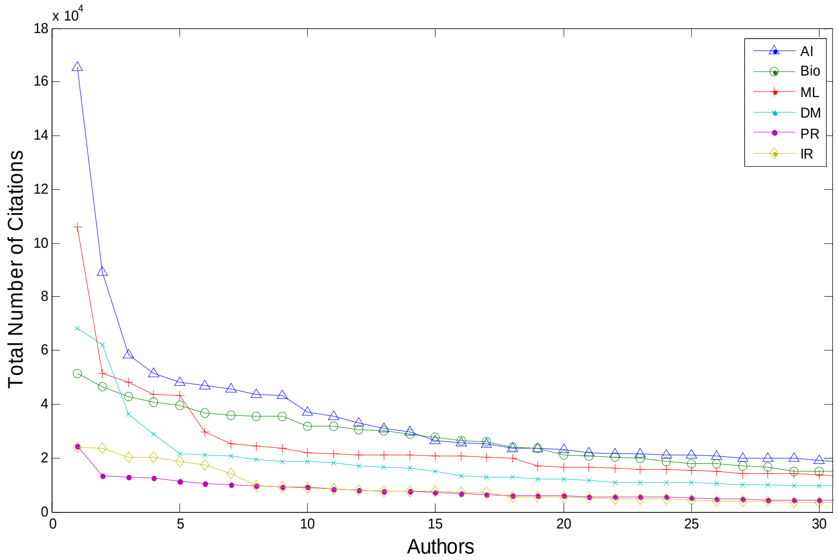

Fig. 1. A comparison of the total number of citations of researchers in different disciplines

As mentioned above, currently three index metrics are calculated in GSC, which are divided into two groups based on all publications and recent publications respectively. The Pearson Correlation Coefficient was used to measure the dependences among these metrics. Fig. 2 shows a scatter plot based on the values of the h-index and the i10-index, over the entire dataset (i.e., all disciplines and all publications). Intuitively, there was strong linear correlation between the two index metrics.

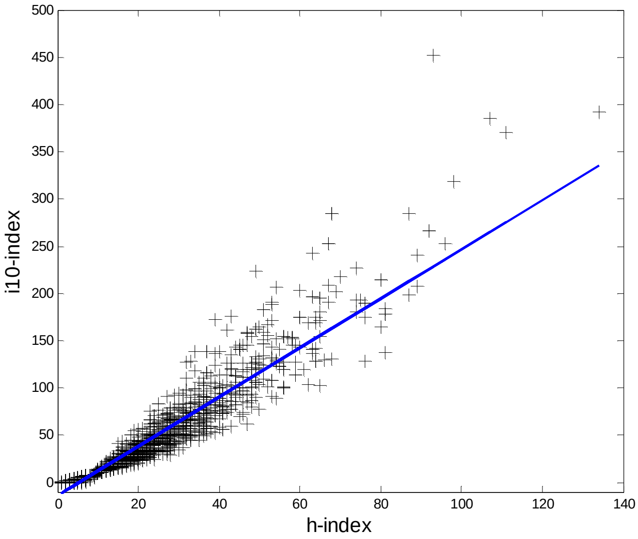

Fig. 2. The relationship between the h-index and the i10-index

In fact, the correlation coefficient (CC) between the h-index and the i10-index was around 0.95 (very strong correlation). In the meantime, the CC value between TC and the h-index was around 0.77, which was similar to the CC value between TC and the i10-index (around 0.75). Table 2 shows the CC values among the three index metrics in each discipline (the number of authors is shown in parentheses). In all disciplines, the CC values were more than 0.7 and some disciplines (e.g., IR) had stronger CC values between TC and the h-index/i10-index than others (e.g., AI & ML).

Table 2. The coefficients among index metrics in different disciplines.

Discipline	*All Publications*			*Recent Publications (Since 2007)*		
	TC vs. h	TC vs. i10	h vs. i10	TC vs. h	TC vs. i10	h vs. i10
IR (511)	0.8495	0.9004	0.9389	0.8583	0.9185	0.9289
AI (1000)	0.7829	0.7738	0.9710	0.8276	0.8366	0.9636
Bio (999)	0.8019	0.7282	0.9374	0.7752	0.7205	0.9180
DM (879)	0.7862	0.7669	0.9327	0.8052	0.8225	0.9306
ML (1000)	0.7690	0.7240	0.9512	0.7604	0.7498	0.9511
PR (539)	0.8434	0.8583	0.9306	0.8404	0.8613	0.9334

3.2 Personal Citation Pattern

One of the major benefits of GSC is that it provides the most accurate citation profiles for individual researchers. Among many potential research questions that can be addressed, we focused on the personal citation patterns at this stage. Regardless of the TC value of a researcher, it is interesting to investigate how the citations are distributed among his/her publications. For example, some authors may have a small number of highly cited documents (e.g., review papers and books are often cited heavily) while other authors may have citations distributed relatively evenly.

To testify this hypothesis, cluster analysis was conducted as follows. Firstly, only researchers with more than 200 citations and at least 10 papers were selected (3539 authors in total). Secondly, the papers of each researcher were sorted based on the citation numbers in descending order. Thirdly, a 10D vector was created with the first element corresponding to the proportion of the citations of the top 10% papers among all citations of the specific researcher. Similarly, the second element corresponded to the citation proportion of the second 10% papers and so on. Note that the actual number of papers was rounded to the nearest integer towards minus infinity for the first 9 subsets and all rest papers were assigned to the 10th subset (the number of papers in it may be higher than average).

As a result, each researcher was represented by a data point in the 10D space, specified by the distribution of citations. Fig. 3 shows the results of clustering using the K-Means method (K=2) and the Euclidean distance as the similarity measure.

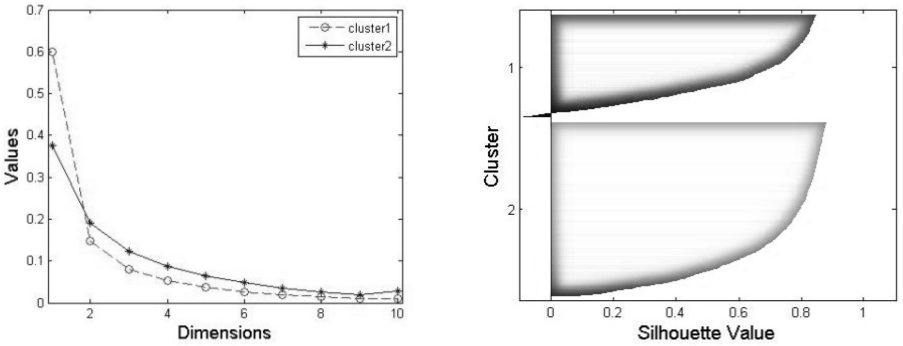

Fig. 3. Personal citation pattern: cluster centroids (left); clustering evaluation (right)

Fig. 3 (left) shows the two cluster centroids after clustering. It is evident that authors in *cluster 1* tended to have the majority of citations concentrated on the few very best papers. In fact, the top 10% papers contributed around 60% of the total citations. By contrast, papers of authors in *cluster 2* received citations in a more uniform manner. Fig. 3 (right) demonstrates the quality of clustering using the Silhouette method. There were more authors in *cluster 2* (2225 authors or 62.87%) than in *cluster 1* (1314 or 37.13%) in *cluster 2*. Moreover, only few data points had negative Silhouette values (the average Silhouette value was 0.6717) and it is clear that the two clusters were reasonably well structured.

4 Topic Trend

The titles of papers may provide some vital clues on how topics or keywords evolved in a discipline. One solution is to show the title texts in the form of tag cloud using IBM Word Cloud Generator [2] where the font size of a word in the cloud is proportional to its frequency in the text. We divided the papers into 2 subsets: before 2007 and since 2007 to observe the change over time. Fig. 4 shows an example of the keywords in the field of IR. For better visual effect, some common title words such as *based*, *approach*, *systems*, *analysis*, *using* were ignored as these words appeared frequently across all disciplines. Additionally, we filtered out *information*, *retrieval* and *search*, due to their large number of occurrences in both clouds.

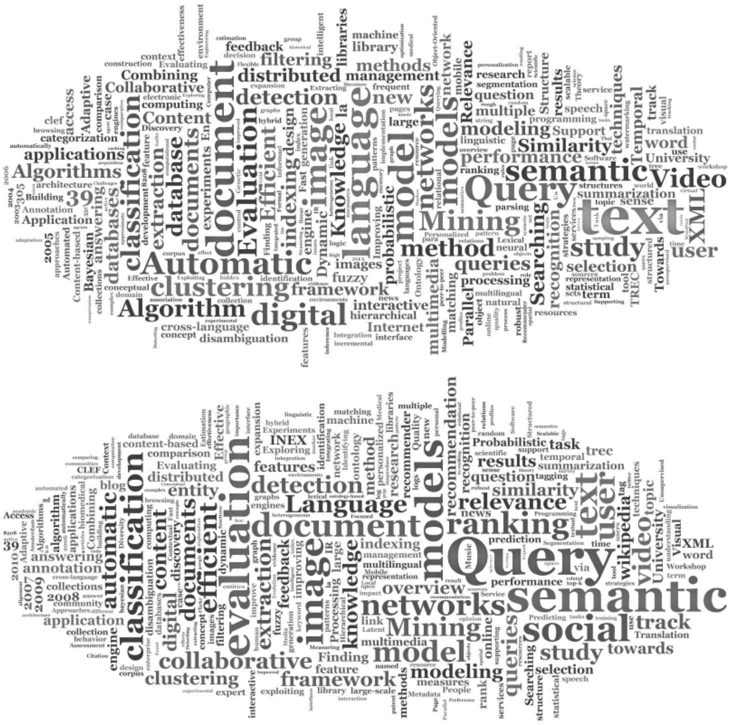

Fig. 4. Tag clouds of titles in IR: top (before 2007); bottom (since 2007)

By comparing the two clouds corresponding to two consecutive time periods, it is possible to find some interesting clues. For example, the font size of *text* dropped down while the font size of *semantic* scaled up, which may suggest that semantic retrieval has been growing rapidly as the mainstream research topic in IR. Meanwhile, the word *social* appeared only in the bottom cloud, indicating that a new research direction related to social networks has emerged in recent years.

[2] http://www.wordle.net/

5 Conclusion

Citation analysis has attracted significant attentions from researchers in all aspects. With the availability of online searchable citation databases such as Web of Science and Google Scholar, it is now possible to conduct decent analysis using data mining and knowledge discovery techniques. In this paper, we presented one of the first studies on Google Scholar Citations, which provides well organized citation records in terms of discipline and authorship. We found that different disciplines had different numbers of typical citations and there were strong correlations among the h-index, the i10-index and the total number of citations. For individual researchers, we identified two distinct groups of authors in terms of the distribution of citations. Finally, we demonstrated the effectiveness of tag cloud in discovering the topic trend in certain research field. In the future, with the increasing number of registered researchers, we will be able to collect more comprehensive data sets and conduct more thorough analysis. It would also be interesting to combine the domain knowledge with the results of data analysis to provide more insights into each discipline.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 60905030).

References

1. White, H.D., McCain, K.W.: Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972–1995. Journal of the American Society for Information Science and Technology, 49(4), pp. 327–355 (1998)
2. Chen, C.: CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. Journal of the American Society for Information Science and Technology, 57(3), pp. 359–377 (2006)
3. Gisvold, S.E.: Citation Analysis and Journal Impact Factors – Is the Tail Wagging the Dog? Acta Anaesthesiol Scand, 43(10), pp. 971–973 (1999)
4. MacRoberts, M.H., MacRoberts, B.R.: Problems of Citation Analysis: A Critical Review. Journal of the American Society for Information Science and Technology, 40(5), pp. 342–349 (1989)
5. Bakkalbasi, N., Bauer, K., Glover, J., Wang, L.: Three Options for Citation Tracking: Google Scholar, Scopus and Web of Science. Biomedical Digital Libraries, 3(7) (2006)
6. Harzing, A., Wal, R.: Google Scholar as a New Source for Citation Analysis. Ethics in Science and Environmental Politics, 8(1), pp. 61–73 (2008)
7. Meho, L.I., Yang, K.: Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science versus Scopus and Google Scholar. Journal of The American Society for Information Science and Technology, 58(13), pp. 2105–2125 (2007)
8. Torvik, V., Smalheiser, N.: Author Name Disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data, 3(3), Article 11 (2009)
9. Hirsch, J.: An Index to Quantify an Individual's Scientific Research Output. Proceedings of the National Academy of Sciences, 102(46), pp. 16569–16572 (2005)
10. Egghe, L.: Theory and Practise of the g-index. Scientometrics, 69(1), pp. 131–152 (2006)