Research Proposal

1 Introduction

Nowadays, a researcher's impact or productivity is mainly measured by looking at his scientific publications. Depending on the journal a certain article is published in, it is considered to have a lower or higher impact on the research field. Hence, for any researcher it is crucial to publish in high-quality journals.

Estimating a journal's quality or impact is a non-trivial task, especially for researchers being not completely familiar with their field (e.g. being at the beginning of their scientific career) or if a certain publication covers a topic which is not within one's normal field of expertise and thus relevant venues are not known too well.

To ease the process of choosing a relevant journal, different metrics were established to measure a journal's impact and to rank it among others of the same field. These metrics are known as scientometrics or bibliometrics and address the measurement of impact/performance of different journals, institutes or individual researchers. Very basic metrics covered by these terms are citation number and paper number. Within the last 15 years and due to the availability of publications through the internet, altmetrics complement this discipline.

Traditional ways to measure a journal's impact are journal rankings (e.g. SCImago[1], JCR[2]) and national journal classification lists (e.g. Denmark, Norway, Spain). Normally, journal rankings are based on mere scientometrics, mainly the Journal Impact Factor (IF) [1], whereas national classifications lists are manually compiled by certain committees and depend on peer-based judgements [2].

2 Goal of the Project/Thesis

Both approaches are subject to various criticism: The IF is discriminating against newer journals, researchers and publications and it's subject to the Matthew effect [5]. National classification lists are said to be biased (as committee members may be influenced by certain interests), thus not fully objective and are only compiled in a certain interval (e.g. one year).

2.1 Problem Statement

Therefore I would like to investigate a new approach to assess both journal and individual researcher performance based on publicly available information on the web, mainly using Google Scholar.

Instead of taking only scientometrics into account which are based on mere number of published papers and citation counts (e.g. IF, h-index [3]), a reversed technique should be examined: Starting from university rankings and based on the assumption that good researchers rather work at good institutions, it is of interest whether a reasonable journal ranking can be deduced.

[1] See http://www.scimagojr.com/journalrank.php
[2] See https://jcr.incites.thomsonreuters.com

2.2 Research Objectives

2.2.1 Verify research hypothesis

Firstly, the hypothesis that there is a correlation that researchers from good universities publish at good venues / in good journals and vice versa needs to be verified[3] (see 3.).

2.2.2 Build a reversed ranking

Secondly, a reversed ranking has to be built. The following questions express the main information needs:

(1) Which universities are ranked in which certain order?
(2) Which researchers are associated with a certain university?
(3) Which papers were published by a certain researcher?
(4) Which venues do exist?
(5) Which papers were published at a certain venue?
(6) Which researchers published at a certain venue?

An adequate database should be mainly - but not only, as it's only able to answer the information needs (2)[4] and (3)[5] – built using data from Google Scholar (*GS*) utilizing the GS scraper[6]. Different global university rankings, namely *CWUR*[7], *THE*[8] and *QS*[9], allow to answer information need (1). A list of venues (4) can partly be answered by *AMiner*[10], *CORE*[11] and *MAS*[12] (data publicly available[13]), (5) and (6) by GS and AMiner[14].

As a result, the following main steps are necessary:

1) Data consolidation: Bring together different data sources and build a data base using scraping and original data (if available)
2) Build models and API to answer information needs: University model, affiliations, venue ranking

2.2.3 Build a visual exploration tool

Thirdly, to gain insights about individual researcher performance and collaboration patterns throughout the whole research community, supporting visual exploration tools are of interest, e.g. to

- find geographic collaboration patterns (e.g. between institutions, countries)
- find out whether certain journals/venues have geographically limited impact
- visualize temporal changes concerning a venue's / researcher's impact

[3] Using CWUR[7] and QS[9] as university rankings; Can be done by manually or automatically examining Google Scholar data
[4] See, e.g., https://scholar.google.de/citations?view_op=view_org&org=16008520586621520646&hl=en&oi=io
[5] See, e.g., https://scholar.google.de/citations?user=No2ot2YAAAAJ&hl=en
[6] See https://bitbucket.org/sciplore/grespa
[7] Center for World University Rankings, see http://cwur.org/2016.php
[8] Times Higher Education, see https://www.timeshighereducation.com/world-university-rankings/2017/world-ranking#!/page/0/
[9] QS World Universities Rankings, see http://www.topuniversities.com/university-rankings/university-subject-rankings/2016/computer-science-information-systems
[10] See https://aminer.org/ranks/conf
[11] See http://portal.core.edu.au/conf-ranks/?search=&by=all&source=CORE2014&sort=atitle&page=1
[12] See, e.g., http://academic.research.microsoft.com/RankList?entitytype=3&topDomainID=2
[13] See http://datamarket.azure.com/dataset/mrc/microsoftacademic
[14] See, e.g. https://aminer.org/conference/539078f320f770854f5a8980

2.3 Expected Results

The master project should yield both a database and a website prototype to analyze journal and researcher performance. The functionality of both will be limited to English publications and the field of Computer Science. The website prototype will focus on a visual exploration tool *"GeoImpact"* allowing to see geographical impact (by citations) of venues, researchers and institutions[15].

Resulting findings from these two parts of the implementation can be further evaluated in the thesis possibly resulting in a sophisticated way to rank and explore journal, institution and individual researcher's impact and coherences among each other. A case study which compares traditional rankings (see 1) with the designed ranking could be conducted using a/multiple subset(s) of Google Scholar data (e.g. top 100 venues in Computer Science) to assess the outcome of the project. On top, a qualitative evaluation in form of an expert discussion could complement this approach.

3 Methodology

To verify both parts of the research hypothesis (2.2.1), two different approaches are necessary. Each prove requires a statistical analysis. For 1.1) the institution's top researchers are determined by mere citation count[16] which requires applying a heuristic to address academic age bias[17].

1.1) Researchers from good universities publish at good venues: Look at the top 5 papers of the top 20 researchers of the top 50 institutions and validate against venue rankings
1.2) Publications at good venues originate from researchers affiliated with good universities: Look at the top 100 papers of the top 50 venues and validate against university rankings

Implementing the reversed ranking (2.2.2) requires the following tasks:

2.1) Database design[18]
2.2) General Google Scholar scraper improvements
- Statefulness[19]
- BibTeX downloading and parsing
- Proxy support / distributed crawling

Building the visual exploration tool (2.2.3) in form of a website prototype can be divided into these tasks:

3.1) Data preprocessing, consolidation / normalization, mapping and cleansing
3.2) Web backend: web service based API
3.3) Web frontend
- Map visualization ("GeoImpact")
- Development over time
- Interactive comparison of different researchers, institutions, venues

[15] Possibly also certain institution's departments, if necessary information can be extracted from data.
[16] As Google Scholar institution profiles rank researchers by this metric only.
[17] E.g. normalizing citation count by time span between first and latest publication. This can be done looking at one researcher's "Citation indices" / "Citations per year" Google Scholar chart.
[18] In form of a relational database. Eventually examine graph representation using, e.g., Neo4J and possible benefits.
[19] Currently, the scraper does not remember which urls it crawled. Restarting results in re-crawling all urls again.

3.1 Work Plan

The resulting rough work plan looks as follows and will be completed over time.

Date	Milestone / Details
17/10/16	Discussion and refinement research proposal
14/11/16	Start project
12/12/16	Results research hypothesis validation (1.1 and 1.2)
30/01/17	Database design, improved Google scholar scraper (2.1 and 2.2)
30/02/17	Data processing implementation (3.1) and first API version (3.2)
30/03/17	Revised API and first website prototype (3.2 and 3.3)
30/04/17	Revised website prototype and project documentation (3.3)
01/05/17	Start thesis

4 References

[1] E. Garfield, "'The Agony and the Ecstasy: The History and Meaning of the Journal Impact Factor,'" *Int. Congr. Peer Rev. Biomed. Publ.*, pp. 1–22, 2005.

[2] A. Zuccala, N. Robinson-Garcia, R. Repiso, and D. Torres-Salinas, "Using network centrality measures to improve national journal," *Proc. 21st Int. Conf. Sci. Technol. Indic. València | Sept. 14-16, 2016*, 2016.

[3] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc Natl Acad Sci U S A*, vol. 102, no. 46, pp. 16569–16572, 2005.

[4] J. Wu, K. Williams, H. Chen, M. Khabsa, C. Caragea, A. Ororbia, D. Jordan, and C. L. Giles, "CiteSeerX : AI in a Digital Library Search Engine," *Proc. Twenty-Sixth Annu. Conf. Innov. Appl. Artif. Intell.*, pp. 2930–2937, 2014.

[5] Merton, Robert K. "The Matthew effect in science." *Science 159.3810 (1968): 56-63.*