Pub Finder: Assisting the discovery of qualitative research

Thanasis Vergoulis
Athena RC
Greece
vergoulis@imis.athena-innovation.gr

Ilias Kanellos
NTUA & Athena RC
Greece
ilias.kanellos@imis.athena-innovation.gr

Serafeim Chatzopoulos
Univ. of Peloponnese & Athena RC
Greece
schatz@imis.athena-innovation.gr

Christos Tryfonopoulos
Univ. of Peloponnese
Greece
trifon@uop.gr

Theodore Dalamagas
Athena RC
Greece
dalamag@imis.athena-innovation.gr

Yannis Vassiliou
NTUA
Greece
yv@dblab.ece.ntua.gr

ABSTRACT

In recent years, scientists are under pressure to publish more and more papers in order to survive in a very competitive environment. This trend has resulted in an explosion of the the number of published research papers in all scientific fields. Additionally, it has been shown that a large portion of these articles contains low quality research and errors. As a result, identifying the most important articles which are relevant to a subject of interest is a non-trivial, tedious task. In this work, we present Pub finder, a tool that assists the discovery of qualitative publications. This tool supports ranking and comparing scientific papers based on various impact aspects. Furthermore, it provides useful additional features like intuitive infographics and article bookmarking. Pub finder is freely available (in beta version) at http://andrea.imis.athena-innovation.gr/pubfinder/web

ACM Reference Format:
Thanasis Vergoulis, Ilias Kanellos, Serafeim Chatzopoulos, Christos Tryfonopoulos, Theodore Dalamagas, and Yannis Vassiliou. 2018. Pub Finder: Assisting the discovery of qualitative research. In *Proceedings of Hellenic Data Management Symposium (HDMS'18).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Scientific research consists of the systematic investigation of theories and hypotheses about the mechanisms of nature and the fundamentals of science and technology. The output of this process, along with details on the methodology used, is reported by scientists in research papers. These papers are reviewed by independent researchers (called reviewers) and, if their content is evaluated to be novel and interesting, they are published in scientific journals and conference proceedings. The publication of research results is an essential part of the scientific method and the published articles provide a valuable reference for subsequent research.

Published research papers contain useful information for a large variety of professionals, besides scientists, including: research and innovation policy makers, governmental administrators or fund managers to evaluate the academic performance of research institutions, recruiters in research institutions trying to enlist highly qualified candidates. As a matter of fact, academic appraisal stands to benefit strongly from effectively evaluating the impact/quality of published research papers based on content and metadata, as also evidenced by the proliferation of academic search engines.

In recent years, research is dominated by the "publish or perish" trend: scientists are under pressure to publish more and more papers in order to survive in a very competitive environment [3]. This has resulted in an explosion of the number of published research papers in all scientific fields [8]. Meanwhile, the aforementioned pressure is being related to a drop in the quality of the produced research output. It has been shown that a very large portion of the published research output is of low quality or even erroneous [5] and, to make matters worse, many false scientific findings gain coverage by the social and mass media resulting in the misinformation of the public. The aforementioned issues raise barriers to the effort of the professionals in academia and research to retrieve or analyze data from research articles.

To mitigate the aforementioned issues, many research retrieval and analytics infrastructures, such as Google Scholar, AMiner [10], Semantic Scholar, and InCites, have been developed. Apart from facilitating thematic search of research papers, such infrastructures use paper impact measures to provide insights about the influence of each article to other articles of the same field. Although a paper's influence is a useful indicator of its quality, there are also other aspects of its quality. Another important aspect of paper impact is its popularity (i.e., if the paper has hype, currently). Since popularity and influence are not necessarily correlated to each other, focusing only on one of them (e.g., on paper influence) could result in misleading conclusions. Note that each impact aspect could be useful under different types of search. For example, a student trying to write a field survey would be interested in retrieving the most influential papers of this field. In contrast, a new researcher trying to identify the latest trends in a particular research discipline may be interested in identifying popular papers.

In this work, we present *Pub finder*, a tool that assists the discovery of both popular and influential papers in the scientific literature. To guarantee satisfactory results for both types of search, the popularity and influence measures were carefully selected (see also Section 2.1). Furthermore, useful additional features, like intuitive

HDMS'18, July 2018, Larnaka, Cyprus

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Figure 1: Screenshot of the article comparison page. In this example, two articles are being compared. The first (dark gray polygon) is the most influential, while the second (light gray polygon) is better in terms of the remaining metrics.

comparative infographics on articles, and other services, such as article bookmarking, were implemented, making Pub finder a powerful tool to assist scientists in their literature review.

2 SYSTEM DESCRIPTION

2.1 Discovering popular or influential papers

A powerful search engine, based on user-provided keywords, lies in the heart of Pub Finder. The great power of this engine is that it supports ranking the retrieved papers based on their popularity or influence, depending on the user's desire. Hence, Pub Finder can be useful both for discovering papers in the cutting edge of their field and for identifying fundamental papers having a large impact in a particular discipline.

To estimate each paper's influence, Pub Finder executes the PageRank algorithm [1, 7] on the citation network stored in Pub Finder's storage components (see also Section 2.4). This citation network consists of roughly 3 million papers from DBLP. For each paper, PageRank provides a score which indicates its influence. It should be noted that PageRank scores are preferable to citation counts for measuring an article's influence, since they do not only capture the number of articles citing it, but also their importance [2].

The estimation of each papers's popularity is based on executing the FutureRank algorithm [9] on the same citation network. This algorithm is based on PageRank and HITS [6]. It applies mutual reinforcement from papers to authors and vice versa, while additionally using time-based weights to promote recently published papers [9]. These weights alleviate the bias of classic methods, such as citation counts and PageRank, against recently published papers and therefore render the method suitable for estimating paper popularity.

Finally, it should be noted that the results of keyword search are not ordered based solely on each article's popularity or influence. The relevance of each article to the user keywords is also considered. Moreover, users can determine whether keyword relevance should affect the ordering of search results, or not (the "low keyword relevance" option will result in ordering the articles based only on their popularity/influence scores).

2.2 Comparing papers and paper infographics

Pub Finder users can select a group of papers for comparison based on particular infometrics. An intuitive radar chart is used to this end (see Fig. 1). The provided chart illustrates the popularity and influence scores of each paper along with the number of times that the paper has been viewed or bookmarked (see also Section 2.3). Additionally, it provides a readability score for each article. This score is based on the Flesch Reading Ease metric [4] computed on the articles' abstract. All measures are normalized based on their greater values in Pub Finder's database.

Furthermore, Pub Finder provides, for each paper, a page of details (see Fig. 2). This page contains useful article metadata (e.g., title, authors, year of publication, references and citations) along with two intuitive infographics:

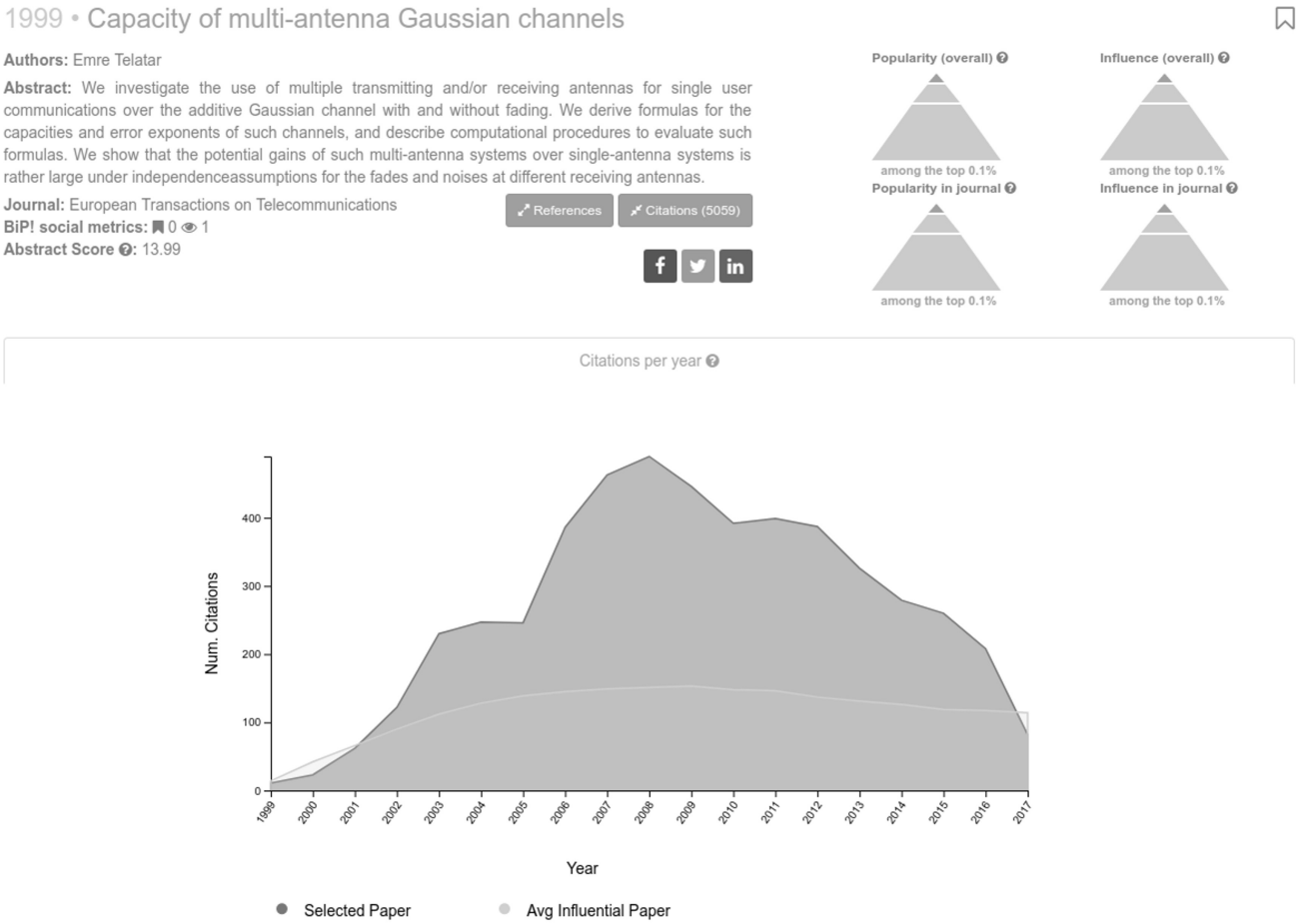

Figure 2: Screenshot of the details page of an article. The article metadata and popularity/influence pyramids appear on the top, followed by the graph displaying the paper's citation history, compared to that of the average influential paper.

- *Popularity/influence pyramids.* This infographic provides an intuition on the paper's popularity and influence score, in comparison to the corresponding scores of (a) the rest of the papers in Pub Finder's database (the first two pyramids) and (b) the other papers published in the same journal (the last two pyramids). Each pyramid is highlighted based on the percentage of articles which have a lower score.
- *Citation History Graph.* This infographic shows the number of citations the selected article received per year. Additionally, to provide further insight into the paper's citation trajectory, the interface also displays the citations per year received by the average influential paper.

2.3 Paper bookmarks

Pub Finder provides a mechanism that enables bookmarking interesting papers. A logged-in user can create a bookmark by clicking on the corresponding bookmark icon that appears in many locations of the interface (e.g., at the top-right corner of the page shown in Fig. 2). The user can browse her created bookmarks just by clicking at the corresponding menu item in Pub Finder's main menu.

2.4 Implementation details

Figure 3 summarizes the architecture of the Pub Finder system. In particular, Pub Finder consists of a number of software and data storage components. All data are organized in the following storage components:

- *Citation graph*: A text file containing the citation relationship among papers, which is stored in the Hadoop Distributed File System (HDFS) mounted on a Hadoop cluster.
- *Paper metadata database*: A relational database containing paper metadata (e.g., author lists, venues, years of publications).
- *Paper inverted index*: An inverted file built on paper titles, abstracts, and author names (based on Sphinx).

These storage components are built by, or interact with the following software components:

- *Paper parser*: This component, written in Python, is responsible for fetching, extracting and loading all paper data and metadata into the Pub Finder's storage components. The

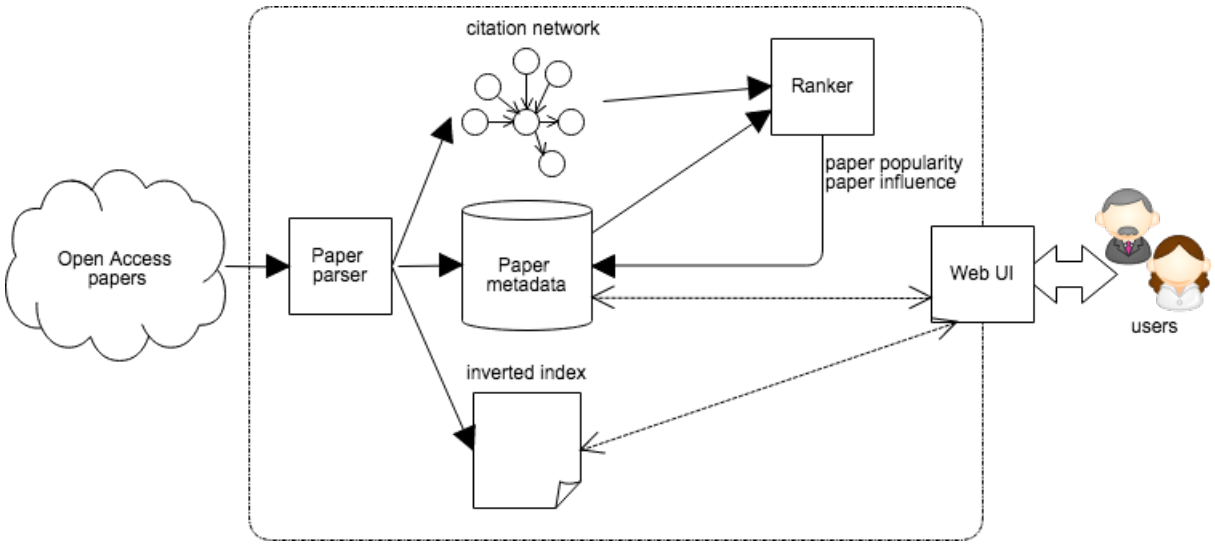

Figure 3: Pub Finder Architecture.

input of this component is an AMiner[1] dataset which contains more than 3 million papers from DBLP and 25 million citation relationships between them. The output consists of the Citation graph, the Paper metadata database, and the Paper inverted index.

- *Ranker*: This component exploits the citation network, as well as the author-paper network, to compute popularity/influence scores for each paper. It is also implemented in Python, written as iterative MapReduce (Hadoop) scripts and being executed on a okeanos[2] cluster.
- *Web UI*: This software component is responsible for the system-to-user interaction. It is mainly implemented using the Yii2 PHP framework[3]. All visualizations (e.g., radar charts, popularity/influence pyramids, citation histories), were implemented using JavaScript and, in particular, the *D3* library[4].

3 DEMONSTRATION SCENARIO

We demonstrate the functionality of Pub Finder by executing queries provided by us and by members of the audience. Furthermore, we explain the concepts of paper popularity and influence and we describe the benefits of Pub Finder in comparison to other paper search engines and research analytics platforms. Short descriptions for two demonstration scenarios follow.

Scenario 1: selecting ordering criteria. A user selects a computer science topic, such as "string matching", which has a long history, i.e., fundamental papers on the topic have been written decades ago, while novel output on the field is ongoing. The user enters "string matching" in the search bar of Pub Finder's interface, selecting to order papers based on popularity and setting keyword relevance to "high". The top retrieved papers published from 1987 to 2014. The user then switches the ranking criteria from popularity to influence. The results now change, returning papers published from 1974 to 1997. Following this, the user relaxes the criteria of keyword relevance, setting it from "High" to "low". The user then repeats the search for popular and influential papers, retrieving papers published from 1974 to 1997 for influence and papers published from 2010 to 2015 for popularity.

Scenario 2: comparing two papers. A user, again, selects a computer science topic in the search bar of Pub Finder's interface. Then, she selects two papers from the result list, a relatively old and a relatively recent. She then clicks on the "Compare" button (at the top right of the screen) and checks the strengths and weaknesses of each paper based on the presented radar chart.

[1]https://aminer.org/citation
[2]https://okeanos.grnet.gr/home/
[3] https://www.yiiframework.com
[4]https://d3js.org

REFERENCES

[1] S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems* 30, 1 (1998), 107–117.
[2] P. Chen et al. 2007. Finding scientific gems with Google's PageRank algorithm. *J. Informetr.* 1, 1 (2007), 8–15.
[3] Sarewitz D. 2016. The pressure to publish pushes down quality. *Nature* 533, 7602 (2016), 147.
[4] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.
[5] J. P. Ioannidis. 2005. Why most published research findings are false. *PLoS Med.* 2, 8 (Aug 2005), e124. https://doi.org/10.1371/journal.pmed.0020124
[6] Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46, 5 (1999), 604–632.
[7] L. Page et al. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. Stanford InfoLab.
[8] C. Pirmez et al. 2016. Scientific journal publishing is too complex to be measured by a single metric: time to review the role of the impact factor! *Mem. Inst. Oswaldo Cruz* 111, 9 (Sep 2016), 543–544. https://doi.org/10.1590/0074-02760160005
[9] Hassan Sayyadi and Lise Getoor. 2009. FutureRank: Ranking Scientific Articles by Predicting their Future PageRank.. In *SDM*. SIAM, 533–544.
[10] J. Tang et al. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 990–998.