Yates and Dixon *Source Code for Biology and Medicine* (2015) 10:16
DOI 10.1186/s13029-015-0046-2

SOURCE CODE FOR BIOLOGY AND MEDICINE

METHODOLOGY Open Access

PageRank as a method to rank biomedical literature by importance

Elliot J. Yates* and Louise C. Dixon

Abstract

Background: Optimal ranking of literature importance is vital in overcoming article overload. Existing ranking methods are typically based on raw citation counts, giving a sum of 'inbound' links with no consideration of citation importance. PageRank, an algorithm originally developed for ranking webpages at the search engine, Google, could potentially be adapted to bibliometrics to quantify the relative importance weightings of a citation network. This article seeks to validate such an approach on the freely available, PubMed Central open access subset (PMC-OAS) of biomedical literature.

Results: On-demand cloud computing infrastructure was used to extract a citation network from over 600,000 full-text PMC-OAS articles. PageRanks and citation counts were calculated for each node in this network. PageRank is highly correlated with citation count ($R = 0.905$, $P < 0.01$) and we thus validate the former as a surrogate of literature importance. Furthermore, the algorithm can be run in trivial time on cheap, commodity cluster hardware, lowering the barrier of entry for resource-limited open access organisations.

Conclusions: PageRank can be trivially computed on commodity cluster hardware and is linearly correlated with citation count. Given its putative benefits in quantifying relative importance, we suggest it may enrich the citation network, thereby overcoming the existing inadequacy of citation counts alone. We thus suggest PageRank as a feasible supplement to, or replacement of, existing bibliometric ranking methods.

Keywords: PageRank, Bibliometrics, Citation count, Impact factor, Journal ranking

Background

MEDLINE is the premier bibliographic database of the U.S National Library of Medicine (NLM), containing over 22 million biomedicine-related entries. With approximately 750,000 new citations added in 2014, it is essential to identify literature of the highest quality for priority reading [1]. High citation rates (in addition to journal impact factor and circulation rates) are proposed to be predictive of article quality [2], thus in turn, scientific importance. Factors such as bias towards review articles and variable bibliographic lengths however suggest that such methods are not always optimal [3].

Citation counts give no weighting towards articles of greater importance. Naturally, definition of such importance is a subjective task. In a static system of inter-article referencing, we observe that a citation by an article from a low distribution journal has equivalence to a citation from a large-scale systematic review. Perhaps a weighting approach would favour articles of greater perceived 'scientific gravity', however this may neglect the emerging relevance of an article's spread through the scientific community. Therefore a method of objectively weighting literature importance would be highly beneficial.

The PageRank algorithm, originally used for link analysis by the search engine, Google [4], provides one such method of ranking by importance. The concept, originally applied to web pages, proposes that a web page itself carries a greater importance if linked to by other high importance pages. Thus for a closed system of total web pages online, a system of merit can be constructed based on assigning a relative weighting (as a proportion of the entire database) to each web page.

Much as web pages are interconnected through hyperlinks, scientific articles are themselves linked via their citations. As such, this study seeks to investigate PageRank-based bibliometrics as an alternative to citation counts alone.

* Correspondence: exy071@bham.ac.uk
College of Medical and Dental Sciences, University of Birmingham, Birmingham B15 2TT, UK

BioMed Central

© 2015 Yates and Dixon. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Methods

The PubMed Central open access subset (PMC-OAS) represents a more liberally-licenced part of the PubMed Central collection [5], freely available online. Contributing journals provide selected full text articles in eXtensible Markup Language (XML) format, specifically for data mining purposes.

PMC-OAS was here chosen, both due to ease of accessibility, though also as a training corpus allowing concept validation prior to expansion to the entirety of MEDLINE. With over 600,000 unique manuscripts included, the dataset amounts to some 40Gb uncompressed [6]. Data parsing and computation was performed in three steps (Fig. 1).

XML parsing

With data ingestion going beyond the capability of traditional desktop computing, on-demand cloud-computing infrastructure was leveraged to parallelise metadata extraction. This commodity cluster environment represents a readily-available, low-cost method of scaling up 'embarrassingly parallel' computational tasks [7].

XML parsing was performed in parallel on four compute nodes (2Gb RAM, 2 virtual CPU cores) using a hand-written Python [8] parser in under two hours (Appendix 1). PubMed identification (PMID) numbers of 'outbound' citations were extracted from each article's reference list and used as reference keys for every citation vertex in the graph of article nodes.

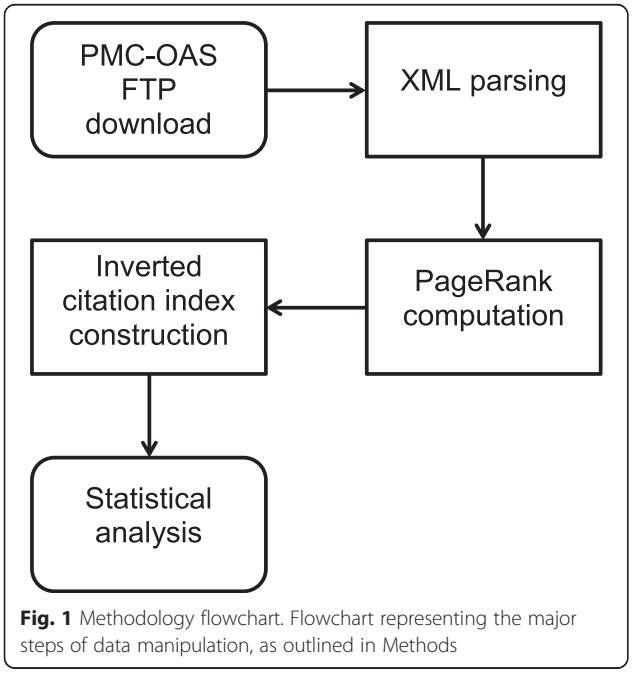

Fig. 1 Methodology flowchart. Flowchart representing the major steps of data manipulation, as outlined in Methods

PageRank computation

PageRank computation was performed on a single compute node (specifications as previous) using an open source C++ based implementation of the algorithm [9]. The algorithm can be summarised as per Fig. 2, where *pi* represents the set of all unique PMIDs in the citation network (and *PR(pi)* its individual PageRank), *d* is the dampening factor (d = 0.85 here), *N* is the total number of unique PMIDs, *M(pi)* represents the set of all inbound citations to *pi*, *PR(pj)* represents the PageRank values of all inbound citations to *pi* and *L(pj)* is the number of outbound citations of *pj*.

A dampening factor was originally introduced in PageRank to model an imaginary surfer randomly clicking on links, that will eventually stop clicking. 0.85 suggests an 85 % probability that at any step, this imaginary surfer will continue to click. Due to the recursive nature of the algorithm, a convergence value (epsilon) of 0.00001 was used to guarantee precision. The algorithm was used as per the reference implementation except where otherwise described.

Inverted citation index creation

MapReduce, a programming model for large corpus processing, also developed at Google, was used to create an 'inverted citation index'. This distributed computational approach allows near linear scalability with increasing cluster size [10], thus facilitating a route for future corpus expansion. The inverted citation index generates a list of 'inbound' citations for each article node in the graph, with a corresponding total citation count.

The high-level programming language, Pig [11] was used as a layer on top of MapReduce for near-natural language manipulation of the dataset. A Pig script was written to facilitate numeric comparison between derived citation count and calculated PageRank (Appendix 2).

Statistical analysis

Statistical analysis was performed using IBM SPSS version 21.0.0.0 [12].

Results

The PageRank algorithm processed and ranked a total of 6293819 unique PMIDs as graph nodes, with 24626354 vertices, representing corresponding outbound citations. A random, 5 % sample of the data was taken (using SPSS randomisation) for statistical analysis. This figure comfortably exceeds the sample size calculation ($n = 385$ required, Raosoft [13]), detailed in Appendix 3.

PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

Fig. 2 PageRank algorithm. PageRank algorithm representation. Set of unique PMIDs in citation network [*pi*], individual PageRank [*PR(pi)*], dampening factor [$d = 0.85$], total number of unique PMIDs [*N*], set of all inbound citations to *pi* [*M(pi)*], PageRank values of all inbound citations to *pi* [*PR(pj)*] and number of outbound citations of *pj* [*L(pj)*]

PageRank is shown to be a surrogate of literature importance

A statistically significant correlation between PageRank and citation count was observed ($P < 0.01$) with a high correlation coefficient ($R = 0.905$). Simple linear regression was performed, obtaining $R^2 = 0.819$ with the fitted regression line being statistically significant ($P < 0.01$), illustrated in Fig. 3.

As such, given the current role of citation count as a marker of literature importance, we demonstrate PageRank to be a similar such surrogate due to high degree of correlation. In light of this finding, we suggest that novel rankings would likely remain broadly similar and thus suggest that implementation of PageRank into the ranking of biomedical literature is feasible.

Top of the corpus comparison

If the putative benefits of PageRank in quantifying importance are to be observed, it must be through outliers from those otherwise highly correlated with citation count. Such outliers may have been preferentially weighted by the algorithm, based on perceived importance. Due to the training subset size, it would be infeasible to account for such examples, however a top of corpus comparison allows some speculative inspection.

The top ten ranking articles of the corpus were compared by descending PageRank (Table 1). This table size was chosen for illustrational ease as graphical whole corpus analysis, aside from regression testing, was outside the scope of this research. From

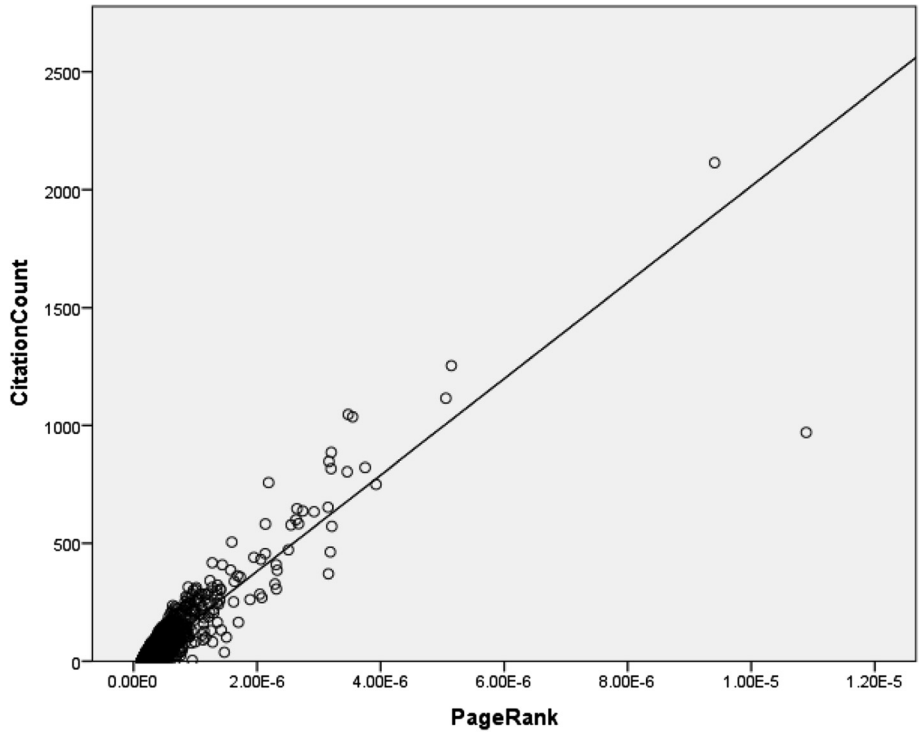

Fig. 3 PageRank versus citation count. Scatter plot of PageRank versus citation count for random, 5 % sample of data. $R = 0.905$ ($P < 0.01$), $R^2 = 0.819$ ($P < 0.01$)

Table 1 Top of the corpus comparison

PubMed ID (PMID)	Paper title	PageRank (E-5)	Citation count
9254694	Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.	3.19	6291
2231712	Basic local alignment search tool.	2.88	5385
10802651	Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.	2.37	4293
11846609	Analysis of relative gene expression data using real-time quantitative PCR and the 2(−Delta Delta C(T)) Method.	1.95	6012†
7984417	CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.	1.78	3899
942051	A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.	1.62	3850
21546353	MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.	1.58	3431
17488738	MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.	1.47	3075
5432063	Cleavage of structural proteins during the assembly of the head of bacteriophage T4.	1.13	2881
3447015	The neighbor-joining method: a new method for reconstructing phylogenetic trees.	1.12	2171

Top of the corpus comparison ($n = 10$), sorted by PageRank, descending. Paper titles were sourced from PMID via PMC-OAS look-up, though were not included in the initial XML extraction. Rankings accurate as of January 2015

inspection, citation count decrement order matches that of PageRank (as expected from the high degree of correlation), with the exception of citation 11846609 (†), a method article with a lower relative PageRank ranking to its citation count.

Whilst this represents a single example, we hypothesize that a method article is likely to be widely cited by those utilising its techniques, however this gives little information about the importance of such implementers. As such, we suggest that this correlation outlier has been proportionally 'down-ranked' by the PageRank algorithm in relation to the rest of the comparative head.

Whilst further work is required to validate such claims, we suggest this finding may build upon the notion of PageRank's potential benefits in outweighing citation count alone. If the method is truly able to better weight those articles with higher importance rather than mass citation, we propose that its implementation into the ranking of biomedical literature may be warranted.

Discussion

PageRank can be trivially calculated on commodity cluster hardware

The use of on-demand cloud computing infrastructure for data extraction and computation allows for scalability with increasing corpus size. In the event of increasing article burden, additional XML parsing nodes could be employed with linear cost and throughput. Despite the uncompressed corpus totalling approximately 40Gb, the fully citation-extracted form was <500 Mb. Therefore, we suggest that growth by an order of magnitude (in the range of entire MEDLINE database size) could still be stored on a single commodity hard drive.

Whilst the PageRank calculation was performed on a single node, expansion beyond 2Gb of RAM on a single computer is becoming cheaper and widely available [14]. The use of MapReduce for inverted citation network creation allows near-linear scalability, similar to XML parsing, and can thus be trivially re-evaluated as the corpus grows. PMC-OAS is updated daily, thus all metrics can be recalculated in a matter of minutes (minus the cost of data parsing), as required by the maintainer.

Expanding automated XML processing to MEDLINE as a whole is problematic

The PMC-OAS full-text articles are freely available in XML format, facilitating automated citation extraction. Unfortunately, the vast majority of MEDLINE articles are not open access, meaning that full-text access in not trivially available without bulk licencing programmes. Furthermore, the lack of XML-based metadata in non-open access articles limits the capability for rapid citation network generation.

Efforts have been made to parse bibliographic data from papers [15, 16], however attempts are limited by paid access to such articles in addition to the efficiency of extraction from a variety of article distribution file formats. We thus identify expansion beyond this 600,000-article training corpus as a major barrier to non-proprietary bibliometrics.

Articles appearing in PMC-OAS, referenced articles, which were not included in the corpus. This means that the latter's PMID appeared in the citation network and thus received a PageRank. However, due to the limited inclusion set of this work, the PageRank (and thus relative ordering) is by no means final and would inevitably change should expansion to the whole of MEDLINE be feasible.

Other methods of importance quantification

Thus far, importance analysis has been derived from article citation networks alone. However, importance is a non-static entity, with the impact of papers going beyond that of, who cites who. Indeed, importance of a particular work may be represented by its spread through the scientific community, rather than an 'acknowledgement-based' system of the traditional publishing model. Social media may provide a real-time window into this community dissemination.

Altmetrics, the use of the social web for insight into article impact [17], has previously shown promise in correlation with citation count and may therefore add to bibliometrics through real-time importance weighting [18]. Consideration of social impact is beyond the scope of this research, though provides an exciting avenue for further exploration, perhaps in conjunction with PageRank.

Conclusions

PageRank is a novel method for determining the importance of biomedical literature. The possibility of commodity cluster hardware use and value re-calculation following corpus expansion suggests that curation of an open access citation network is not beyond the limits of a single maintainer. Whilst further work will inevitably be required to expand the network beyond the XML data-mining corpus of the PubMed Central open access subset, the 600,000-article training corpus provides a starting platform for PageRank's addition to existing importance ranking methods.

Appendix

Appendix 1

Python source code for XML citation network extraction. Script finds .nxml files within the directory where all PMC-OAS XML full-text documents were extracted. Subsequently generates in key, value format corresponding [citer PMID, cited PMID] and stores as a comma separated value (CSV) file.

```
import xml.etree.ElementTree as ET
import glob


for nxml in glob.glob('papers/**/*.nxml'):
        citations = []


        tree = ET.parse(nxml)
        pmid = tree.iterfind('front/article-meta/article-id[1]').next().text


        for parent in tree.iterfind('back/ref-list'):
                for child in parent.iterfind('.//pub-id'):
                        if child.attrib['pub-id-type'] == 'pmid' and child.text != None:
                                citations.append(child.text)


        file = open("citer_cited.csv", "a")
        for cite in citations:
                file.write(pmid + "," + cite + "\n")
```

Appendix 2

Pig source code for citation network generation and further PageRank comparison. Accepts key, value list of citer, cited PMIDs (as per Appendix 1) and groups by cited PMID before co-grouping with corresponding PageRank value and citation count.

```
citer_cited = LOAD 'citer_cited.csv' USING PigStorage(',') AS (citer: chararray, cited: chararray);

pageranks = LOAD 'pageranks.csv' USING PigStorage(',') AS (cited: chararray, pagerank: float);


grouping = GROUP citer_cited BY cited;

citation_count = FOREACH grouping GENERATE group AS cited, COUNT(citer_cited.cited)-1 AS cite_count;


co_grouping = COGROUP citation_count BY cited, pageranks BY cited;


main_output = FOREACH co_grouping GENERATE FLATTEN(pageranks.cited), FLATTEN(pageranks.pagerank), FLATTEN(citation_count.cite_count);


ranking = ORDER main_output BY $1 DESC;

STORE ranking INTO 'cite_count_pagerank.csv' USING PigStorage(',');
```

Appendix 3

Statistical analysis performed on derived dataset, accurate as of January 2015. Tables generated using IBM SPSS version 21.0.0.0.

Kolmogorov-Smirnov test for normality

Table 2 Tests of normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
PageRank	.383	314664	.000
CitationCount	.399	314664	.000

[a]Lilliefors Significance Correction

Pearson's Correlation test

Table 3 Correlations

		PageRank	CitationCount
PageRank	Pearson Correlation	1	.905[a]
	Sig. (1-tailed)		.000
	N	314664	314664
CitationCount	Pearson Correlation	.905[a]	1
	Sig. (1-tailed)	.000	
	N	314664	314664

[a]Correlation is significant at the 0.01 level (1-tailed)

Simple Linear Regression

Table 4 Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.905[a]	.819	.819	4.844

[a]Predictors: (Constant), PageRank

Table 5 ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33389613.691	1	33389613.691	1423001.392	.000[b]
	Residual	7383297.502	314662	23.464		
	Total	40772911.193	314663			

[a]Dependent Variable: CitationCount
[b]Predictors: (Constant), PageRank

- Equation for the regression line:
  - CitationCount = −27.861 + (204365203.423 x PageRank)

Table 6 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	−27.861	.027		−1023.789	.000
	PageRank	204365203.423	171318.510	.905	1192.896	.000

[a]Dependent Variable: CitationCount

Abbreviations
PMC-OAS: PubMed Central open access subset; NLM: National Library of Medicine; XML: eXtensible Markup Language; PMID: PubMed identification; FTP: File Transfer Protocol; CSV: Comma-separated values.

Competing interests
The authors declare no competing interests in this work. The project was self-funded and was not financially supported by any external organisation.

Authors' contributions
EY designed the methology and performed PageRank calculation.
LD performed statistical analysis and guided the methodology.
LD and EY prepared the manuscript and proof read the final version.
Both authors read and approved the final manuscript.

Acknowledgements
Prof Jamie Coleman, School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham for his kind support throughout.

Received: 5 February 2015 Accepted: 9 November 2015
Published online: 09 December 2015

References

1. NLM: Fact Sheet MEDLINE ®. [http://www.nlm.nih.gov/pubs/factsheets/medline.html]. Accessed 2 Apr 2015.
2. Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. JAMA. 2002;287(21):2805–8.
3. Adam D. The counting house. Nature. 2002;415(6873):726–9.
4. Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. In: Stanford Digital Library Working Paper SIDL-WP-1999-0120. Stanford University. 1999.
5. PMC: Open Access Subset [http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/]. Acccessed 2 Apr 2015.
6. PMC: FTP Service [http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/]. Acccessed 2 Apr 2015.
7. Leykin A, Verschelde J, Zhuang Y. "Parallel Homotopy Algorithms to Solve Polynomial Systems". Proceedings of ICMS 2006. 2006.
8. Python Software Foundation: Python programming language [https://www.python.org]. Acccessed 2 Apr 2015.
9. Github: Panos Louridas (louridas) PageRank C++ implementation [https://github.com/louridas/pagerank]. Acccessed 2 Apr 2015.
10. Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA. 2004.
11. Apache Software Foundation: Pig [http://pig.apache.org]. Acccessed 2 Apr 2015.
12. IBM: SPSS [http://www-01.ibm.com/software/uk/analytics/spss]. Acccessed 2 Apr 2015.
13. Raosoft: Sample size calculator [http://www.raosoft.com/samplesize.html]. Acccessed 2 Apr 2015.
14. Brock DC, editor. Understanding Moore's law: four decades of innovation. Philadelphia: Chemical Heritage Press; 2006. ISBN 0941901416.
15. Zou J, Le D, Thoma GR. Locating and parsing bibliographic references in HTML medical articles. Int J Doc Anal Recognit. 2010;13(2):107–19.
16. Zhang X, Zou J, Le DX, Thoma GR. A structural SVM approach for reference parsing. BMC Bioinformatics. 2011;12 Suppl 3:S7. doi:10.1186/1471-2105-12-S3-S7.
17. Melero R. Altmetrics - a complement to conventional metrics. Biochem Med (Zagreb). 2015;25(2):152–60.
18. Thelwall M, Haustein S, Larivière V, Sugimoto CR. Do altmetrics work? Twitter and ten other social web services. PLoS ONE. 2013;8(5), e64841.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central