Investigation of the PubMed scientific literature using a graph and parallel computing based approach

Stijn Arends - 377499[1†*]

[†]Equal contributors

*Corresponding author

[1] Institute for Life Science & Technology, Hanze University of Applied Sciences, Groningen, The Netherlands

Key words: literature, graph theory, parallel computing, big data, spark

Introduction

Since the first publication in 1665, over time the number of publications in scientific journals has only increased. How papers are structured and published has also changed over time. For example, during the late 1970 and the 1980s the ethics of scientific publication evolved[1]. The need for research funds and awards has led to competition amongst research scientists, which in turn has led to a change in how the research was published. For instance, the h-index (i.e. measure of the number of publications published, as well as how often they are cited[2]) has become an important measure amongst scientists. One way a scientist might increase their h-index is by often referring to their own work. Another way is for a group of authors that work regularly together to refer to each other a lot. In this project, the structure of publishing in the scientific literature was investigated. This was done by processing, and graph based analyses for the entire PubMed literature database whilst making use of parallel and distributed computing.

Methods

PubMed Data

The data consisted of the entire PubMed literature database which was distributed over 1063 XML files where each file contained 30000 articles, so in total there were 31,89 million articles. Each file had the same structure, however, for some files information was missing. For example, for some files, the PMID was not available or stored somewhere else from where it was normally supposed to be. Moreover, the way references to other papers were noted was also not consistent.

Parsing the data

In order to parse the data in parallel a star topology network was set up, i.e. one computer was used to act as a server that was responsible for distributing the data (PubMed article) and at least one other computer (preferably more) acted as a client, which was responsible for actually processing the articles.

Only the essential information was extracted from the articles, such as PMID, name of author, title, name of co-authors, journal, language used, publish date, and references. As mentioned above, the way that references were written down was not consistent and a distinction could be made between two different types: using the citation plus the PMID, and only using the citation which consisted of the author names, title, publish date, and journal. The PMID of the reference was desired as this was needed to create the citation graph (i.e. nodes are PMIDs and references are edges to other PMIDs). Therefore, regular expressions were used to attempt to extract the author names plus the title of the reference so that in a later stage, once all the articles have been parsed, they could be mapped to the data frame to extract the PMID of the references. Therefore, three additional columns were added to capture the name of authors and title for each reference, where applicable, as well as the type of reference (i.e. reference using PMID or authors + title). Finally, the data that was processed was written out to multiple JSON files.

Constructing the citation graph

The processed articles were stored inside a graph for further analysis. First, an adjacency list was created that contained all the vertices (articles) with the set of neighbours (references). Next, a vertex list was created for all the vertices with any outgoing neighbours (i.e. articles without references). Finally, the vertex attributes were extracted and saved. This was again done making use of the star topology network to reduce the processing time.

The graph that contained the entire dataset was too large to load, therefore, a random subset of 5000 articles was taken from the data, resulting in a graph containing 166563 articles, which was used for further analysis. From this subset another subset was taken for visualization purposes, this subset contained 1165 articles.

Results

The first question that was meant to be answered was: 'how large a group of co-authors does the average publication have?'. By counting the number of co-authors for each article and then taking the mean it was determined that the average number of co-authors was 3,12. Second, it was found that 43.42% of the time an author publishes with a similar group of co-authors. This was calculated by first collecting all the co-authors an author has had and then to see how often there was an intersection between two lists of co-authors. Third, the question 'do authors mainly reference papers with other authors with whom they've co-authored papers (including themselves)?' was explored. By putting all the co-authors an author has had into one list and then calculating how often the author or co-author of a reference was inside of this list we found that only 14.7% of the time an author referred to a paper that included an author with who they have previously worked with (Fig. 1). Next, we wondered whether the distribution in time for citations of papers, in general, was different from highly cited papers. Papers were determined to be highly cited when they got more citations than 99.5% of all the papers, which in this case was more than 5. The time span between the most recent and oldest citation was taken for all articles and based on that it was found that on average papers, in general, had a time span of 0.14 years and the highly cited papers had an average time span of 7.6 years (Fig. 2A & Fig. 2B).

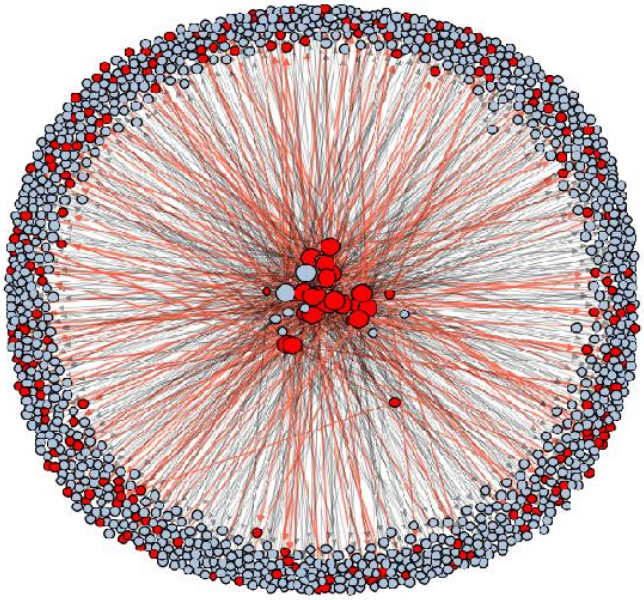

Fig. 1 A graph depicting articles referring to other articles where the main author has worked with either the main author or the co-authors of the referenced article (depicted in red.

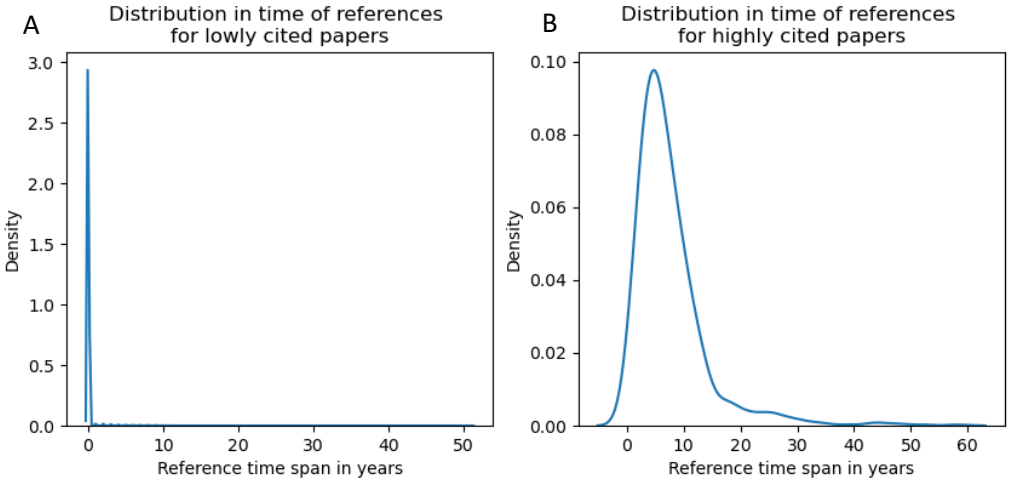

Fig. 2 The distribution in time for citations of papers in general (A) and for highly (i.e. papers with more than 5 citations) cited papers (B).

Additionally, it was also investigated whether or not papers that share key words cite each other more often (i.e. is there a correlation between citations and the number of key words papers share). For each paper, the key words were compared to the key words of each paper that has cited it and the key words of each paper it has referenced. By taking the number of times key words matched and dividing it by the total amount of citations and reference a fraction was calculated that indicated how often papers that refer to each other contain similar key words. From this we found that 24.6% of the time papers cite each other they contain at least one similar key word (Fig. 3). Moreover, we also looked if this correlation is different for the most-cited papers, which turns out not to be the case because compared to the 24.6% across all papers the most-cited papers only had similar key words in 1.7%. Likewise, the correlation between citations and language used was looked at. Not too surprisingly, in 97.8% of the citations the two papers were written in the same language (Fig. 4). Finally, the most cited paper, author, and highest index were calculated. The most cited paper titled: 'Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.', from the author Livak, K.J. was the most cited paper with 48 citations. However, the author with the most citations was

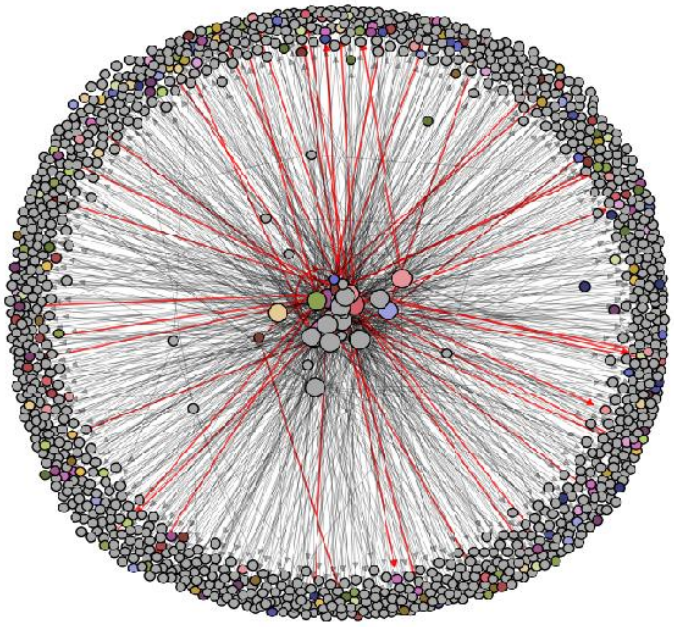

Fig. 3 Citation graph showing papers that share similar key words. Gray are articles that do not contain key words, other colours represent a key word.

Wang, Y. with a total of 155 citations. The author with the highest h-index was Siegel, R. with an h-index of 3.

Discussion

In this project, the PubMed literature database was investigated which was based on a graph and parallel computing approach.

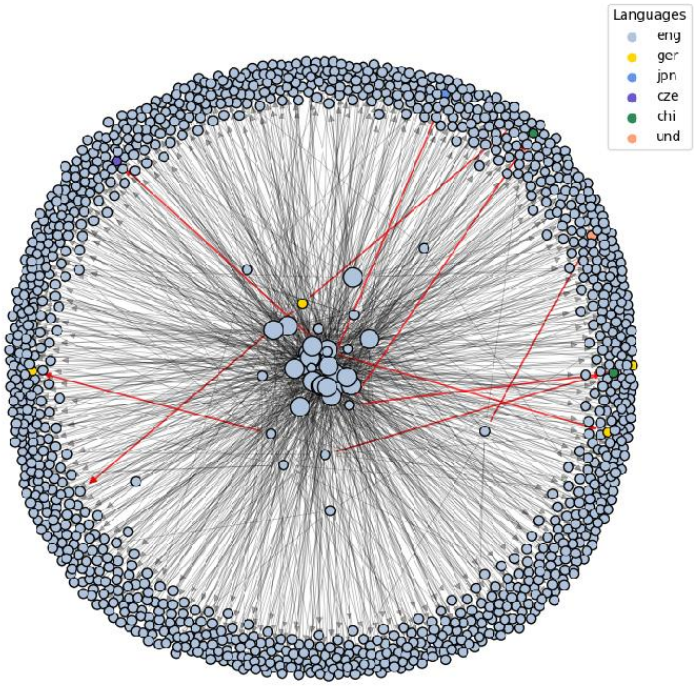

Fig. 4 Citation graph showing papers that cite other papers that were written in another language.

We found that the average article has 3.12 co-authors and that authors on average publish 43.42% of the time with the same or a similar group of co-authors. More interestingly, we found that authors on average cite other articles that they wrote themselves or contain authors with who they have worked with 14.7% of the time. Also, the distribution in time for citations of papers, in general, is much different compared to papers that are highly cited, 0.14, and 7.63 years respectively. Furthermore, we also found that there is not really a correlation between citations and the number of key words that papers share (0.246). Yet, there is a big difference between the correlation between citations and the number of key words papers share for papers in general and highly cited papers, as the correlation is only 0.017 for highly cited papers. Additionally, it was found that papers mostly reference other papers that were written in the same language (97.8%). Finally, the most cited paper was written by Livak, K.J., the most cited author was Wang, Y. and the author with the highest index was Siegel, R.

After the data was processed the articles were separated into three groups: no references, PMID references, and author + title as references. To make full use of the data the references containing the author names and the titles were to be mapped to the complete data frame to extract the PMID of the article. However, due to the sheer size of the data we were unable to achieve this. This means that some information got lost, even so, there were only ~150 thousand articles that had these kinds of references out of the total ~6.5 million articles that had references which means that the loss is neglectable. Moreover, we were unable to use the entire graph to analyse the data and were forced to use a smaller subset. At the moment 166563 articles were used for the analysis and it could be argued that using

~500 thousand would also have worked. Therefore, some of the results that was acquired are not reliable, such as the correlation between citations and shared key words for the most-cited papers. This is because the number of 'most-cited' papers is very low and by chance for all these papers the citations did not share a lot of key words. This subset of most-cited papers is not representative and the correlation could very well change if another random subset were to be taken. Another example is for the time span for papers in general this would undoubtedly change if the entire dataset were to be used.

For future research, the entire data set, or at least a much larger subset, should be used in order to perform an unbiased analysis. Furthermore, it should be considered to use software that is suited to work with large graphs in parallel, such as the python package GraphFrames which is based on Apache Spark's GraphX library.

Conclusion

In this project, a subset of the PubMed Literature database was investigated based on a graph theory and parallel computing approach. The results give some general statistics about the PubMed literature database as well as an indication of how articles are structured in general. Yet, to perform a better and unbiased investigation the whole PubMed literature database should be used instead of a mere subset.

Author contributions

Stijn Arends: Methodology, Formal analysis, Writing - original draft, Review & editing.

Data Availability

Software and scripts are available at:

https://github.com/stijn-arends/programming3/tree/main/Assignment6

PubMed data is available for download at:

https://pubmed.ncbi.nlm.nih.gov/download/

References

1. Scientific Publishing | Encyclopedia.com. https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/scientific-publishing.
2. Hirsch, J. E. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A* **102**, 16569–16572 (2005).