SciTrend: Visualizing scientific trends in publications

Mimi Yang
Department of Electrical Engineering
Stanford, USA
mxy@stanford.edu

Benjamin Wang
Department of Mechanical Engineering
Stanford, USA
bwang17@stanford.edu

ABSTRACT

We have developed a system and process of visualizing scientific publication trends use the PubMed API to access recent publications concerning a seed keyword. A histogram of the number of papers published per year related to the seed keyword demonstrates the growth and maturation of that material or technique in the bio-medical field. Further understanding of research trends is obtained by parsing the publication titles and enumerating the associated words by time, showing the temporal popularity of terms associated with the seed keyword. Advanced analysis can be conducted by examining relative popularity of common title terms per year using treemaps. The outlined method guides a researcher in understanding and discovering the history and development of past research, popular current research directions, and future applications in a chosen research field.

INTRODUCTION

Bibliometrics is a method of quantitative analysis of scientific publications. Many current techniques in bibliometrics are focused on citation analysis, looking at how scholars and organizations cite one another in publications. From this data, we can understand the networks and collaborations of scholars and the development and focuses of scientific inquiry over time.

The focus of the majority of bibliometric tools is to underline the relational aspect of publications. In any popular area of science, there are groundbreaking publications that instigated the origin of the field of study. These pioneer publications are almost always cited in any proceeding publication in that area. As equally important, the most prolific writers of the area are commonly cited.

The progress in scientific research is heavily dependent on the previous works of other scientists and engineers. Specifically individual fields within science and engineering emerge, grow and decline over time depending on a variety of factors. Publications of research in scientific journals is a strong indicator of the research interest for a specific topic. In most cases there is always a "seed" paper that starts a field or topic of investigation, leading to many other research projects that stem and reference the original first paper. Visualizations of paper genealogy have been completed in the past [1], with global visualizations of entire databases also implemented [2]. However, none of these programs take into account how a specific fundamental research topics (ex. carbon nanotubes, graphene) first started and evolved to modern day applications. As such, the focus of this paper is to develop a process of visualizing how these important fundamental discoveries have impacted the landscape of modern science and engineering.

For this project, we use the PubMed API to access recent publications concerning a keyword. The publications can be sorted by time, and the publication count over time shows us the temporal popularity of research related to that keyword. We also track the popularity of words used in the titles of the publications found from the original keyword. This search gives us the an understanding of topics and terminology typically associated with the seed keyword, allowing us to see what are "hot topics" in this area of research. For a researcher, having the capability to track the topics being researched in a given scientific area is important in shaping that person's future directions of scientific inquiry. Most importantly, this tool gives a summation of the newest developments in research in order to stay on the cutting edge of innovation. Many bibliometrics are concerned purely with the strength of connection between publications, professors, and keywords without an emphasis of what is happening now. Those tools are often skewed by the original research that instigated the whole field of study. The following visualization maintains the ability to determine the time of the onset of specific research, but offers much more as a method to discover current research direction and future applications.

PREVIOUS WORK

Tools such as PaperLens [3] and CiteVis [4] analyze influential authors and references by analyzing the citation mapping. To visualize these relationships, the connections are often demonstrated using network graphs [3–6]. While these visualizations contain a lot of valuable information, the graphical format is not conducive of locating a specific paper within the network to determine that paper's specific contributions and connections. In general, the visualizations can be a mess of wires

and nodes that do not facilitate understanding of the relationships. Chinchilla- Rodriguez [7] slightly remedies the mess by grouping publications according to country and organizes a spherical graph according to overall domestic and international citations. However, this graphical simplification also loses the ability to find connections between individual papers.

Tools such as Citeology [1] and CiteRivers [8] organize the papers temporally. This enables the inclusion of individual points for each paper along the time axis, allowing for further analysis of specific papers. Laying out the publications along the x- axis permits addition of a relational notation along the y-axis. For example, Citeology includes wires that span the y-axis to connect the references and citations between papers. For the intended purpose, Citeology may be one of the most concise yet informational visualization format.

To enhance the amount of information available, graphical interactions can be included [8, 9]. Specifically, once the relationships between publications are mapped, we may be interested in reading and knowing specific details about a particular influential paper. Giving access to the full citation information of a chosen paper enhances the usability and usefulness of a graphical visualization.

The previously cited works are primarily focused on analyzing the relationships between papers and authors based on citation analysis. While that objective can greatly facilitate a researcher in delving into the literature of their project, the overarching knowledge and view obtained from this foray may be too narrow. To obtain a broader understanding about the topics being studied in a scientific area, we propose organizing publications based on publication title words. By searching a specific keyword and tallying the other associated title words found in publications, we can determine popular subareas of research over time. This offers a overview of the research activity as a whole rather than concentrate on finer detail interactions. By plotting the popularity of associated title words over time, we can include interactions to also incorporate the paper-to-paper relationships that most bibliometrics are concerned with.

METHODS

The National Center for Biotechnology Information (NCBI) provides access to a plethora of biomedical and genomic information. The center also provides E-utilities to query for specific scientific research, including the Esearch tool to download universal identifiers (UIDs) and summaries of relevant research. The data used in this SciTrend visualization is sourced from Esearch queries of the PubMed database. We refine the list of publications included in the visualization by querying according to publication keywords (known as other text [OT]

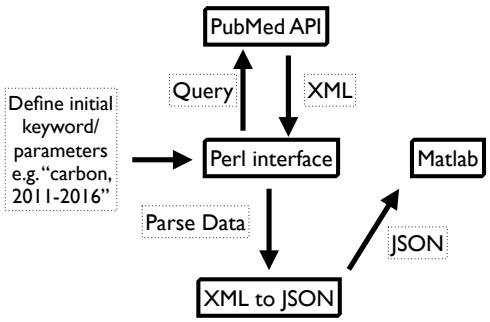

Figure 1. Process flow for data extraction and import. The keyword and time parameters are used to query the PubMed database. The resulting XML file of all of the search results matching the initial keyword query are parsed into an JSON file and imported into Matlab.

in PubMed). As the authors are responsible for selecting keywords that they feel are representative of the core concepts and innovations of their paper, we argue using keywords as a refinement metric produces a list of relevant and important work.

Esearch returns a XML file of UIDs and a list of descriptors for each publication. For this specific visualization, we are mainly concerned with the publication date and title. Creating a visualization from the raw XML introduces a long wait time due to the large volume of data that is processed, so we parse the XML file into a JSON file for manageability. In introducing this processing step, we can isolate the publication data that is necessary for the visualization and reformat the data and syntax to be compatible with later Matlab processing.

The processing of large data sets generally require cleaning, removal of stop words, and stemming. This is generally the most time consuming component of data mining, and also the most resource intensive for large datasets. We implement a very basic stemming algorithm to combine plural words, however more elaborate stemming programs exist that would provide more accurate results. We then clean up the data and remove

For our test case, we use our Perl PubMed API interface to query the main keyword "carbon" with a time period from 2011 to 2016.

The final resulting sub-keywords and year data is then plotted and mapped using Plotly. We visualize our keywords using TreeMaps, which provide a visual representation of the distribution of keywords as a function of area. The data is displayed as nested rectangles for the hierarchical data.

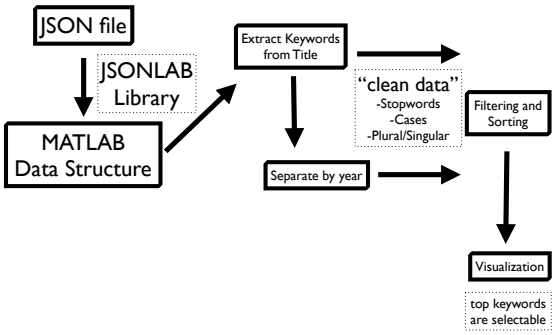

Figure 2. Process flow for generating a visualization from the JSON data. The data JSON file is imported into Matlab using the JSONLAB Library. From the generated structures, we extract and parse the titles of the publications into singular words. Prepositions are removed from the title word list, and the counts for plural and singular versions of words are merged together. The title words that meet the user's filter are then displayed in a visualization.

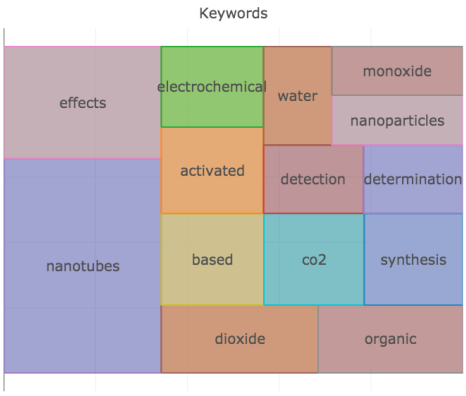

Figure 3. TreeMap of sub-keywords with the main-keywords bein g "carbon" for years 2011 - 2016.

RESULTS

For our test case, we query the parent keyword "carbon" for a time period between 2011 - 2016. Each of our TreeMaps had a total of 15 bins. It is easily concluded that the top two sub-keywords for the entire time period are "nanotubes" and "effects". Moreover, the secondary sub-keywords such as "activated" and "dioxide" are easily attributed to carbon and can be used to iterate the search for a deeper look at how the keywords are connected. In Figure 4, we see that by plotting all of the individual TreeMaps by year, we can see what trends become more prevalent and what trends are not as important over time. To plot each of the TreeMaps by year, we filter the dataset by year and then extract all keywords by title. The side by side comparison allows for an overall of the scientific trends by year, while providing relative importance of each keyword in terms of the size of the rectangle. A possible interaction to the TreeMaps would be weighing certain keywords differently than other keywords based on a few potential algorithms. It would also be interesting for the user to tag certain sub-keywords that are related to the research topic investigated, and running individual queries using the sub-keywords as the main keywords. This is the benefit of using this system of searching through trends, with simple iterations.

DISCUSSION

Another important visualization is comparing TreeMaps by year, which gives the user a sense of how various trends change over time. A downfall of using TreeMaps is that it is relatively difficult to compare specific keywords across time. A histogram is a better visualization for comparing two or more sub-keywords across time. This is useful because it allows for the user to see when certain trends start grow in comparison to other sub-keywords. In Figure 5, we have the histogram interface that has a main histogram(A) and sub-histogram(B). Comparing individual keywords between TreeMaps is rather difficult due to difficulties comparing area between maps. We have implemented a histogram/bar graph representation that shows various keywords as a function of year. As seen in Figure 5, subfigure A shows the total number of the main keyword over time. Subfigure B shows the sub-keywords over time and allow for simple and easy comparison between sub-keywords.

Figure 3 shows the TreeMap for the keyword "carbon" from 2011-2016. It is interesting to note that with a bin size of 15, we get a usable distribution of reference keywords. It is possible to change the number of rectangles in the TreeMap, which would give the user the ability to adjust the granularity of the results. The year by year TreeMap also allows for general trend identification, with focus on seeing if there are any overwhelmingly popular keywords.

At this level the keywords are not weighted and are simply shown as a number count. It would be interesting to weight the keywords based on search popularity, journal impact factor, or another metric that would allow for user control of weights. Another interesting feature would be showing the connection between certain keywords to other keywords, allowing for linking and relationships.

CONCLUSION AND FUTURE WORK

The above mentioned method and visualizations guide a researcher in understanding and discovering the history and development of past research, popular current research directions, and future applications in a the various research fields. This process gives us the an understanding of topics and ter-

Figure 4. The use of TreeMaps as a function of different times, showing the changes in sub-keywords across years.

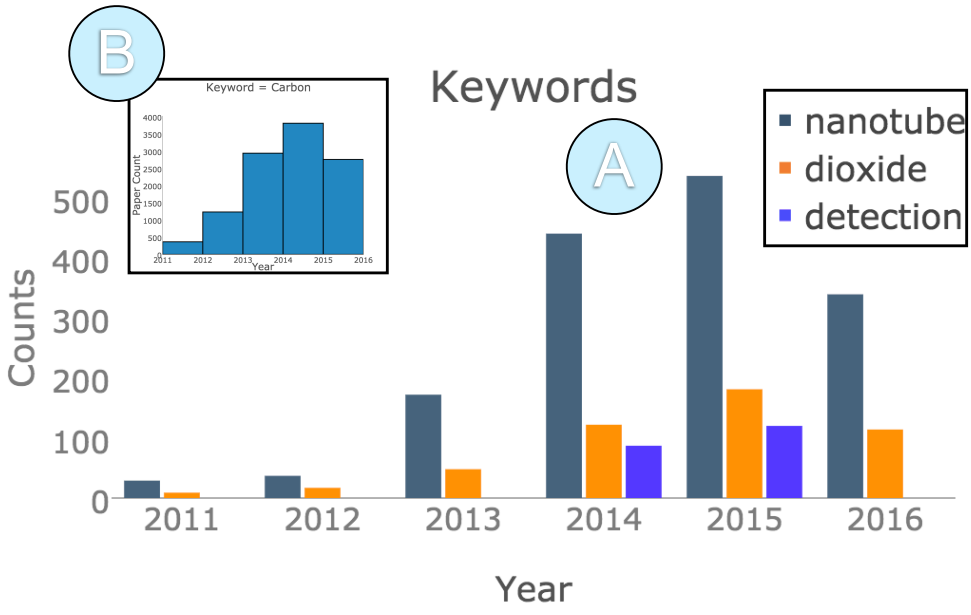

Figure 5. The keyword histogram contains (A) multiple sub-keywords compared over each year and (B) the main keyword compared over each year.

minology typically associated with a seed keyword, allowing us to see what are "hot topics" in this area of research. The majority of bibliometrics are primarily concerned with delineating the connection between publications, professors, and keywords while this tool is unique in that its primary purpose is to give a broader sense of the newest developments in research. This process flow can be applied to a variety of scientific areas to give seasoned researchers ideas for future projects or to give new researchers impressions of the history of their research area.

The full process flow and visualizations were operated on publication title words sourced from PubMed data. This data source was selected due to the existing API infrastructure for querying and download paper descriptions. Future work may include implementing a similar algorithm for other publisher and journal access API to increase the utility of this method. In addition, the visualizations presented are mainly focused on words found in publication titles. While we have argued that the words included in the publication titles are representative of the core ideas and concepts presented in the papers, it may be insightful to examine the other keywords used to tag the queried publications. At this time, the Esearch utility does not expose all keywords associated with a paper, but comparing visualizations of the keywords versus visualizations of the publication title words may provide significant perception of past and current research trends.

REFERENCES

1. J. Matejka, T. Grossman, and G. Fitzmaurice. "Citeology: visualizing paper genealogy". Proceedings of ACM CHI '12 Extended Abstracts, pages 181-190. ACM, May 2012.
2. D. George and R. Knegjens, Paperscape, http://paperscape.org/, (May 25, 2016)
3. B. Lee, M. Czerwinski, G. Robertson and B. B. Bederson, "Understanding eight years of infovis conferences using PaperLens", Proc. of the IEEE Symposium on Information Visualization,page 216.3, Washington, DC, USA, 2004. IEEE Computer Society.
4. J. Stasko, J. Choo, Y. Han, M. Hu, H. Pileggi, R. Sadanaand and C. D. Stolper, "CiteVis: Exploring conference paper citation data visually", Posters of IEEE InfoVis, 2013.
5. H. Small, "Visualizing science by citation mapping", Journal of the American Society for Information Science, 50(9), 799-813.
6. P. Glenisson, W. Glänzel, F. Janssens , B. De Moor, "Combining full text and bibliometric information in mapping scientific disciplines", Information Processing and Management: an International Journal, v.41 n.6, p.1548-1572, December 2005.
7. Z. Chinchilla-Rodríguez, M. Benavent-Pérez, F. de Moya-Anegón, S. Miguel, International collaboration in Medical Research in Latin America and the Caribbean (2003-2007), Journal of the American Society for Information Science and Technology, v.63 n.11, p.2223-2238, November 2012.
8. F. Heimerl, Q. Han, S. Koch and T. Ertl, "CiteRivers: Visual analytics of citation patterns", IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 01, 2016.
9. P.C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner and J. Thomas, IN-SPIRE Infovis 2004 Contest Entry, Proc. IEEE Symp. Information Visualization, Oct. 2004.