Viziometrics: Identifying Central Figures in Scientific Papers

Olga Kazakova*
University of Washington
Seattle, USA

Poshen Lee**
University of Washington
Seattle, USA

Bum Mook Oh***
University of Washington
Seattle, USA

Jevin West****
University of Washington
Seattle, USA

Bill Howe*****
University of Washington
Seattle, USA

Abstract

Scientific communication depends on visual representations of data, results, analysis, and models. Given the important role of these information objects, we have built a figure-centric search engine called VizioMetrics.org. We have used millions of figures from PubMed Central to better understand effective visual communication and scholarly impact. In this poster we present preliminary results for automatically identifying key figures in scholarly papers. We conducted a large-scale survey asking authors to identify their central figures — the single visualization that encapsulates key aspects of a paper. If participants were able to identify such figures, they were asked to indicate what the selected figures represent. Our results show that for over 90% of evaluated papers the authors were able to identify a single central figure. In most cases such figures represent results and the most common figure class is the composite, followed by the diagram. We then use this training set to test early-stage algorithms for identifying these graphical abstracts.

Keywords: VizioMetrics; central figure; figure classification; figure content; image search; scientific literature, visual content.

Index Terms: Figure Retrieval, Information Retrieval, Data Visualization, Bibliometrics, Viziometrics, Machine Vision.

1 Introduction

VizioMetrics.org is an image search engine for the scholarly literature [2]. It includes open source algorithms and software for automatically classifying figure types [2]. The search and classification are performed on 8 million images from PubMed Central. Currently we are working on a feature that automatically identifies a "central figure" in a scientific publication. We have defined the central figure as a single visualization that encapsulates key aspects of a paper, a graphical summary that captures the content of the article for readers at a single glance. This feature can enable us to rank the results of the queries according to their importance within the paper in which they appeared. The existing algorithm for central figure identification relies on an NLP approach, which evaluates the similarity between a given figure caption and the abstract of a paper. However, before we could rely on its the results, we needed to evaluate its performance.

One of the most straightforward ways to assess the performance of an NLP algorithm is to estimate its accuracy. This requires access to labelled data, which can be treated as a ground truth. Until this research such datasets were not readily available. Therefore we had to obtain the dataset ourselves. Since PubMed contains email addresses of the authors, we decided that one of the most reliable ways to obtain labelled data was to ask the authors themselves to identify "central figures" in their publications. We sent 488,590 survey invitations to the emails addresses that we found in PubMed.

Apart from evaluating the performance of the algorithm, this survey helped us answer the following questions: 1) how often the authors themselves can identify a central figure in their own publication; 2) what figures communicate to the reader; and 3) what classes (e.g. table, diagram, etc.) central figures belong to. We found a strong indication that in most cases scientific papers do have a central figure. Our findings not only help us improve the VizioMetrics platform, but also contribute to our understanding of the role of visual content and help us define the most effective ways in which it can be used. Our dataset provides an insight into the content and class of central figures. It also contains information about the papers in which those central figures appeared as well as impact factors of journals where the papers were published.

2 Results

As of June 15th 2017 we collected data on 8165 distinct papers from 6115 distinct authors. Over 50% of evaluated papers were published in years 2011 – 2014. Only 874 papers were indicated not to have a figure that satisfies our definition of central figure. The collected data shows that for 89.3% of evaluated papers, authors were able to identify a single central figure. Therefore, it is reasonable to conclude that central figure is a concept that objectively exists in modern scientific literature.

Figure 1 shows the distribution of central figures across different classes of visual content identified by VizioMetrics [2]: multi-part, diagram, plot, table, and equation. Multi-part figure class dominates the distribution with 37.8% of central figures identified belonging to it. However diagram and plot are also quite popular, as they encompass 29.4% and 18.8% of central figures respectively. The least popular class of figures is equation (less than 1% of central figures), which is expected given the field of the papers. However, it is important to mention that our dataset of figures only includes equations and tables present in papers in image format. Thus, in our survey LaTeX formatted equations and tables for example would not be listed among the images in a particular paper.

Figure 2 presents the distribution of central figures across the types of content that we identified as suitable for biomedical literature. As you can see in 67% of the cases central figures are shown to represent results. Methods and model follow right after with 13.6% and 12.4% of central figures representing them respectively. Discussion is responsible for only 5% of central figures. In 2% of the papers the authors indicated the content as "other."

Our current algorithm was able to identify the central figure correctly in 37.4% of the papers, for which authors were able to confirm the existence of a central figure. While this figure might seem small, according to our calculations on average evaluated

* kazako@uw.edu
** sephonlee@gmail.com
***bmo5@uw.edu
**** jevinw@uw.edu
*****billhowe@uw.edu

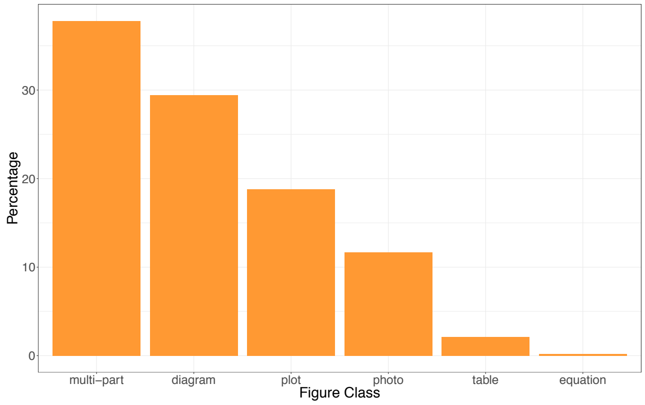

Figure 1: Distribution of central figures across various classes identified by VizioMetrics.

papers have about 5 figures. Therefore, the average accuracy of a random central figure selection would be 20%. Thus, our algorithm is performing better than a random selection. However, it is reasonable to conclude that adjustments need to be made to the algorithm. The fact that over 60% of authors selected figures with captions that are not most similar to the abstract indicates that we may need to employ computer vision techniques

3 Discussion

The main findings of our research are: 1) According to the authors, the majority of evaluated papers have a single central figure; 2) An absolute majority of central figures represent results; 3) Composite figures are a leading class of central figures; 4) According to our findings pure NLP analysis of similarity between figure captions and paper abstracts is not sufficient for developing a high performing central figure identification algorithm. We also introduced the term "central figure" to describe a single most important figure in a scientific article which captures the content of the article for readers at a single glance.

Our work supports some of the findings of previous research conducted in the area of information visualization in scientific literature. We found that despite the fact that plots (graphs) and tables can both be used for presenting data, plots are a much more popular class when it comes to presenting key information. This is inline with findings described by Cleveland, who showed the overall increase in fractional graph area (FGA) in sciences going from social to mathematical, to natural science [4]. It also agrees with findings by Smith et al, who point out that harder sciences seem to be more graph oriented versus table oriented [1]. The fact that equations are rarely found among central figures is consistent with findings by Fawcett and Higginson [3].

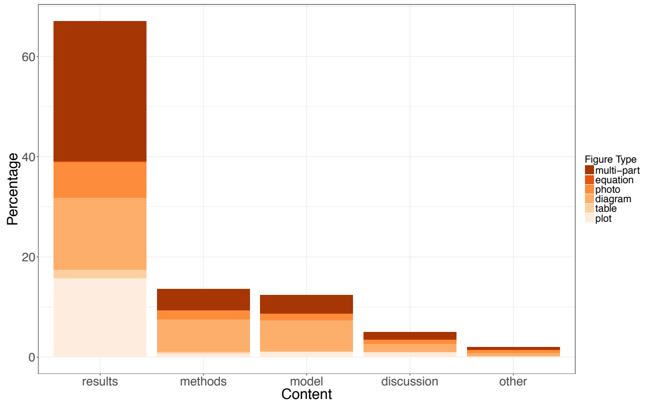

Figure 2: Distribution of central figures across types of content.

Overall, our work fills the gap in the study of the use of visual content in scientific literature when it comes to the evaluation of relative importance of figures within a paper. It also gives insights into what the most important figures in papers represent and what classes of figures are commonly used to communicate the main idea of scientific work. Our data set can be used as a ground truth for developing central figure search algorithms, which can be integrated into image search engines like VizioMetrics.org.

4 Conclusion

In the scientific literature, particularly in biomedical field, images are a valuable source of information, capable of summarizing and communicating ideas faster and more in depth than text. We are seeing more journals and conferences suggesting and sometimes even requiring authors to provide "graphical abstracts" with their submission, which are often defined as "concise, illustrative reflection of the content of your article." Therefore it seems reasonable to pose a question about the existence of a single most important figure, a graphical summary, or what we call a "central figure" in scientific articles. Our research has led to the collection of a unique dataset, which shows that in the majority of cases papers do have a single central figure. It provides insight into the content of these figures and the relative importance of figure types within scientific papers. It also provides researchers with a training set for designing new computer vision algorithms. We plan to make the data set freely available to the broader research community.

Acknowledgements

The authors wish to thank Sean T. Yang from the University of Washington for his work in helping us extract fulltext data from the PubMed Central database.

References

[1] Lawrence D. Smith, Lisa A. Best, D. Alan Stubbs, Andrea Bastiani Archibald, and Roxann Roberson-Nay, "Constructing knowledge: The role of graphs and tables in hard and soft psychology", The American Psychologist, 57(10), p. 749, 2002.

[2] Po-shen Lee, Jevin D. West, and Bill Howe, "Viziometrics: Analyzing visual information in the scientific literature." IEEE Transactions on Big Data , 2017.

[3] Tim W. Fawcett and Andrew D. Higginson, "Heavy use of equations impedes communication among biologists", Proceedings of the National Academy of Sciences USA 109, pp. 11735–11739, 2012.

[4] William S. Cleveland, "Graphs in Scientific Publications", The American Statistician, 38(4), pp. 261-269, 1984.