10.20378/irb-54949

CiteVis: Visual Analysis of Overlapping Citation Intents as Dynamic Sets

Shivam Agarwal*
University of Duisburg-Essen, Germany
Uttiya Ghosh†
International Institute of Information Technology Bangalore, India
Fabian Beck‡
University of Bamberg, Germany
Jaya Sreevalsan-Nair§
International Institute of Information Technology Bangalore, India

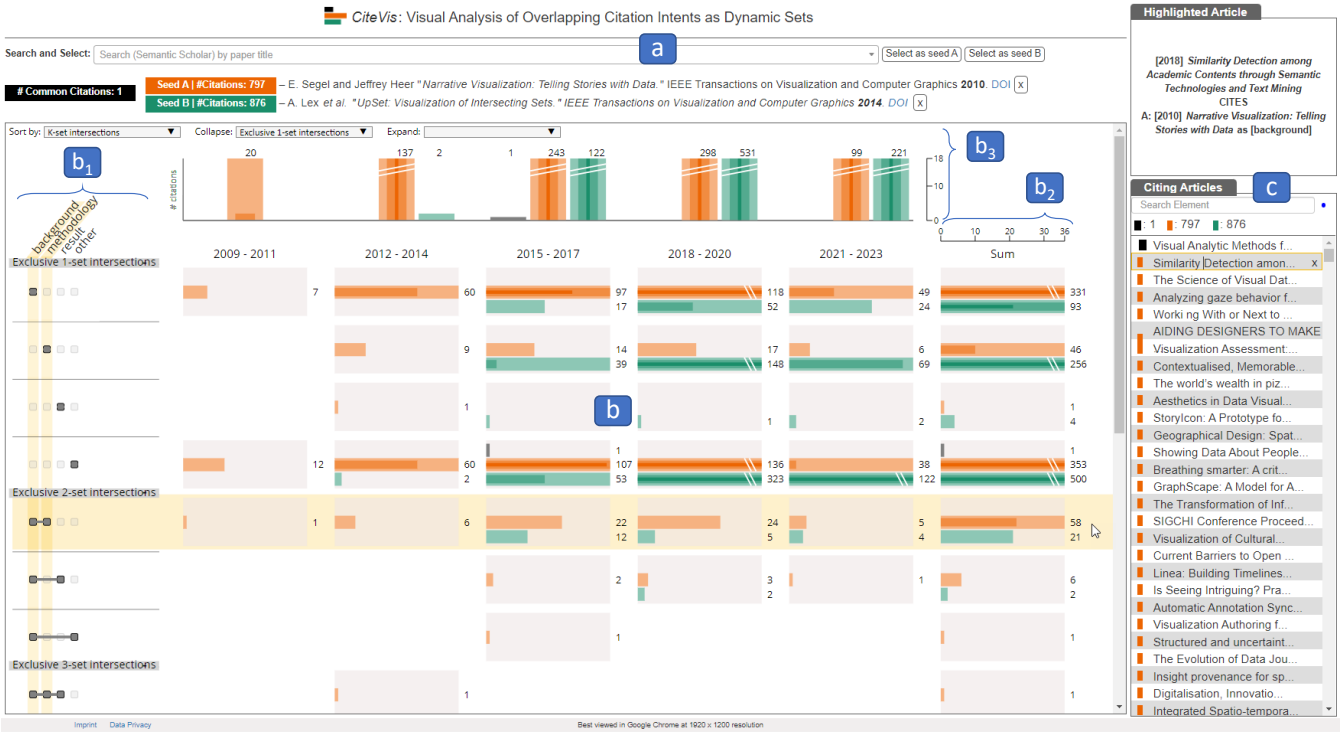

Figure 1: *CiteVis* interface: (a) search and select articles as seeds, (b) visually compare the citations, (c) explore the list of citing articles.

Abstract

A scientific article can be cited with different intents over several years. The citation intents can be inferred by classifying the citation text into different categories. With multiple citations to the same article, the citation intent categories overlap, making their analysis more challenging. We model the categories as dynamic sets and propose an approach to visualize temporal citation trends of an article across overlapping citation intents. The approach supports comparison between the citation trends of two seed articles of interest. The implemented prototype supports searching and selecting seed articles from a Semantic Scholar dataset.

1 Introduction

Citations provide a glimpse of how authors of a scientific article '*stand on the shoulder of giants*'. Existing research is used and cited in multiple ways and serves different purposes. For instance, a work is cited because it provides background of a concept, uses a research methodology, or reports certain results. The intent of a citation can be classified into different categories. The citation context—the text in the citing article that describes the cited article—provides an indication of this intent and serves as a basis for the classification [2]. Analyzing the citation intents of a *seed article* (i.e., the article in focus of the analysis) highlights its reception by the community and, analyzed temporally, might reveal shifts in reception (e.g., from serving as a technical basis to becoming a background topic). Similarly, comparing the citation intents of two seed articles over time helps understand their reception in relation to each other or to general trends in the community (e.g., comparisons get more prominent). Since a seed article can be cited with different intents by a same scientific article (at the same or different locations in the text), a citing article might be assigned multiple intents. As a result, the intent categories overlap. Hence, the data becomes challenging to analyze as it has overlapping categories and a temporal dimension.

To address the challenges, we look towards dynamic set visualizations that represent data with similar characteristics [1, 4]. For instance, *Set Streams* [1] shows the changing membership of elements through streams that branch and merge to indicate overlaps among sets. Using a similar dynamic set data model, we consider 'citing' as the criterion. However, since each article is published only once, each element (a citing article) appears in only one timestep (the publication year of the citing article) and, unlike necessary for applying other dynamic set visualization approaches, cannot be traced through time. We model each category of citation intents (e.g.,

*e-mail: shivam.agarwal@paluno.uni-due.de
†e-mail: uttiyaghosh@gmail.com
‡e-mail: fabian.beck@uni-bamberg.de
§e-mail: jnair@iiitb.ac.in

background) as a set. Hence, a citing article, becomes an element of the corresponding sets at a particular timestep, based on the citation context and the publication year of the citing article, respectively.

2 The CiteVis Approach

We present *CiteVis*[1], a visualization approach to analyze citation trends of a scientific article or compare the trends of two articles. We use the *SciCite* dataset of scientific articles from Semantic Scholar [3]. It provides citations of an article and classifies the citation intents into three categories. First, as *background* information: "The citation states, mentions, or points to the background information giving more context about a problem, concept, approach, topic, or importance of the problem in the field". Second, as *methodology*: "Making use of a method, tool, approach or dataset". And third as *result* comparison: "Comparison of the paper's results/findings with the results/findings of other work". For unclassified citations, we define an additional category: *other*. The citing articles are binned based on their publication year; the default duration of each bin is three years, as we do not expect shifts in citations intents being notable below this limit.

At the top of the interface (Figure 1a), a 'Search and Select' bar lists a few sample articles and also allows searching for an article by the title. While working also with just one seed article, to compare the trends in citation intents, our approach enables selection of two seed articles simultaneously. Colors orange and green identify the citations of seed articles *A* and *B*, respectively. Since there can be common citations, i.e., an article can cite both seed articles with the same citation intent, we represent the common citations in black.

Taking inspiration from *UpSet* [5] and *Set Streams* [1], we show each citation intent (set) as a column on the left (Figure 1b$_1$). Doing so enables us to position exclusive overlaps (common elements present only in the overlapping sets) among citation intents in individual rows, identified by black filled squares in respective columns ( ). For instance, a highlighted row (in yellow) in Figure 1 shows an exclusive overlap of the *background* and *methodology* categories, containing articles citing the seed papers with *background* and *methodology* only. We encode time on the horizontal axis, as shown in Figure 1b. In the columns representing the temporal bins, citation counts of the selected seed article are shown as bars. The last column sums the citation counts per row (Figure 1b$_2$). Vertical bars above the column labels (Figure 1b$_3$) aggregate the citation counts across the rows. The rows can be sorted from the dropdown list (Figure 1b$_3$), e.g., by prioritizing a specific category, or by cardinality in specific or all timesteps.

There can be a large variation of citation counts in different exclusive overlaps. We not only need to show the low citation counts (as they might be particularly important), but also have to facilitate comparison by using a consistent scale across the interface. To address the challenges, we visualize the citation counts using horizon bars [5] on the same linear scale (1 citation = 5 pixels), as they save space by wrapping around to show higher values than the current scale. After each wrap, bars begin from the base with smaller width and a darker shade of the respective color. The wrapping is restricted up to three times, after which, two white lines indicate that the citation count value breaks the scale ( ).

On the right, a list of citing articles is integrated in the interface (Figure 1c). The list can be searched for specific citing article(s) by title. A citing article from the list can be selected, revealing the details of the citation, including citation intent(s) in the box above.

3 Results

We select an article by Segel and Heer [6] as seed *A*. The article reviews the design space of narrative visualizations and discusses distinct genres for data-driven storytelling. From Figure 1, focusing on the orange bars in the last column, we observe that most citations of the article are in the category *background* (331) or *other* (353). As the article covers a broad field and is one of the seminal works defining the intersection of visualization and data-driven storytelling, this weight on the *background* intent is not surprising. Additionally, the article was cited 58 times with both intents *background* and *methodology*, by the same citing articles (highlighted row in Figure 1). This may be explained by a hypothesis that the citing articles used the seed article's methodology of using case studies as a basis to propose genres of narrative visualization. Likewise, the article was cited 46 times with *methodology* intent alone, followed by few citations in the other combinations of categories. The temporal trend indicates that number of classified citations are stabilizing after 2015–2017, but unclassified citations in *other* category are increasing.

Analyzing the citation intents of another article, we select the article proposing *UpSet* [5], a matrix-based static set visualization technique, as seed *B*. As shown with green bars in Figure 1, the increasing citation trend in *methodology* category (256) suggests that the article was used for its primary contribution of a visualization technique for set data. The article also witnessed citations in the *other* category (500), *background* category (93), followed by the combination of both *background* and *methodology* (21). The article also witnessed 4 citations in the *result* category, suggesting that few other set visualization techniques might have been developed and compared with *UpSet*.

There was only one article that cited both seed articles, suggesting clearly distinct themes. Comparing the citations of both articles, we also see different patterns, indicating different types of contributions. Seed *A* is cited more as *background*, while the majority of seed *B*'s citations are in *methodology* category with an increasing trend (ignoring the *other* category). While both articles have been cited with more than one intent (mostly as *background* and *methodology*) in the same citing articles, seed A's citation were higher in the combination (highlighted row in Figure 1).

4 Conclusion

The preliminary results show the potential of the proposed approach for visualizing citation intents. This can help reason about (i) the type of contributions an article makes (especially when being compared to another article), (ii) the character of joint research at the intersection of the selected articles, and (iii) shifts in the perception of an article over time. While we demonstrated the approach for a specific group of citation intents, it could easily show other categories as well (e.g., based on sentiment or research area).

5 Acknowledgments

This research was partly funded by MERCUR (project: "Vergleichende Analyse dynamischer Netzwerkstrukturen im Zusammenspiel statistischer und visueller Methoden").

[1] `https://s-agarwl.github.io/citevis`

References

[1] S. Agarwal and F. Beck. Set Streams: Visual exploration of dynamic overlapping sets. *CGF*, 39(3):383–391, 2020. doi: 10.1111/cgf.13988

[2] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *EMNLP-IJCNLP*, pp. 3615–3620, 2019. doi: 10.18653/v1/D19-1371

[3] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of NAACL*, pp. 3586–3596, 2019. doi: 10.18653/v1/N19-1361

[4] M. T. Fischer, D. Arya, D. Streeb, D. Seebacher, D. A. Keim, and M. Worring. Visual analytics for temporal hypergraph model exploration. *IEEE TVCG*, 27(2):550–560, 2021. doi: 10.1109/TVCG.2020.3030408

[5] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of intersecting sets. *IEEE TVCG*, 20(12):1983–1992, 2014. doi: 10.1109/TVCG.2014.2346248

[6] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE TVCG*, 16(6):1139–1148, 2010. doi: 10.1109/TVCG.2010.179