AbsClust Whitepaper
or Navigating in the Sea of Knowledge

A. Trelin, J. Gau, T. Henning

December 5, 2023

1 Introduction

In the realm of scientific research, finding valuable articles has become a tough task, demanding a lot of time and mental effort. Unfortunately, this often leads to wasted scientific potential, with many articles going unread. The reasons behind this challenge are interconnected and include the explosive growth in the number of scientific articles (Fig. 1, A). The volume of scientific literature is surpassing researchers' ability to keep up with new developments, fueled by the "publish or perish" principle [1]. Total number of articles in technical sciences doubles every ≈13 years [2]. This flood of publications not only requires a significant time commitment but also raises concerns about the sustainability of scientific work. As the pool of scientific knowledge continues to expand, navigating this vast sea of information becomes increasingly difficult.

Adding to the difficulty is the fact that general search algorithms are not well-suited for scientific searches. Unlike casual web searches, scientific inquiries require a nuanced understanding of specific terms, methods, and context. Standard search engines, designed for broader use, struggle to grasp the subtle nuances of scientific language. This makes it challenging for researchers to find articles that precisely meet their needs. Moreover, many of the state-of-the art scientific databases only provide simple full-text search, i.e. the result of the search is all the articles, containing given word/phrase. Taking into account scale of the science, researcher gets tens of thousands articles, potentially containing the one he/she is looking for, but the results are mostly unstructured. In other words, the task of finding the article of interest among thousands of the search query results must be solved manually, and is similar to finding a needle in haystack.

According to our survey among 23 researchers, on average scientists spend 2.21 ± 0.64 hours/week searching for articles (Fig. 1, B) To make matters worse, this problem is getting more pronounced over time. The rate at which new articles are produced shows no signs of slowing down, making it even harder for researchers to stay current with the latest advancements. According to our survey, 72% noticed worsening during their career.

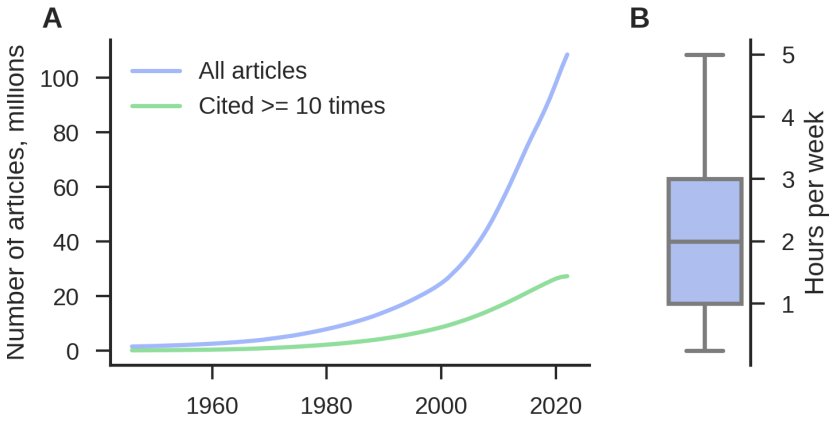

Figure 1: A: total number of articles published by year (according to Microsoft Academic database [3]). B: amount of time, spent by researchers searching for articles (own study).

2 Proposed method

The core concept of Absclust's innovation is the visualization of search results instead of ranking, which allows to provide easily understandable concepts. In general, humans can better perceive and process visual information than any other form of information representation [4]. This fact suggests that finding a correct method for visualizing thousands of results for a specific search query would be much more useful than presenting users with raw titles/abstracts. The idea behind AbsClust is to visualize the data by creating so-called Subject Maps (see Fig. 2), i.e., diagrams where each article is represented as an individual point, and semantically similar articles are positioned close to each other. In visual form, a researcher can examine thousands of articles simultaneously, allowing them to quickly identify relevant articles.

The idea behind semantic map can be easily explained in terms of "semantic distance". E.g. by comparing three articles:

1. Rugina, Radu, and Martin Rinard. "Automatic parallelization of divide and conquer algorithms." ACM SIGPLAN Notices 34.8 (1999): 72-83.

2. Piqueira, J. R. C., Cabrera, M. A., & Batistela, C. M. (2021). Malware propagation in clustered computer networks. Physica A: Statistical Mechanics and its Applications, 573, 125958.

3. Cichocka, A., Marchlewska, M., & Biddlestone, M. (2022). Why do narcissists find conspiracy theories so appealing?. Current Opinion in Psychology, 47, 101386.

one can intuitively say that articles 1 and 2 are much more similar to each other then to article 3. Put differently, the semantic distance between articles

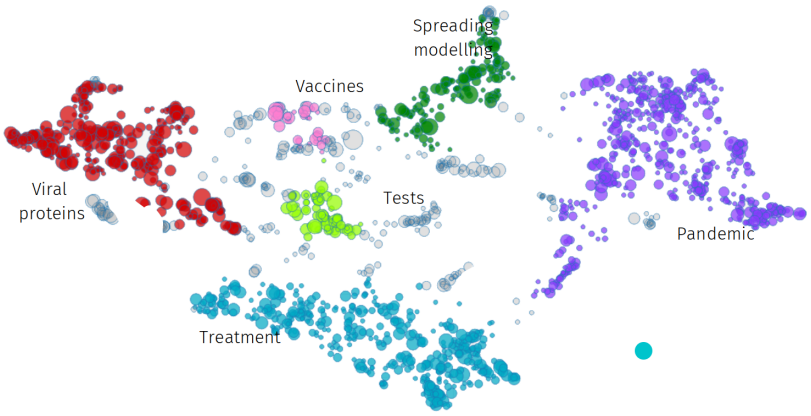

Figure 2: An example of the Subject Map for search query "coronavirus". The subject is chosen due ease of understanding for broad audience.

1 and 2 is smaller than that between 1 and 3 or 2 and 3. Although there is no strict definition for semantic distance, it can be effectively computed with natural language processing algorithms. If distances between all articles are known, Subject Map can be obtained by finding positions of the articles, such that Euclidean distance between them approximately equals to the semantic distance. This can be achieved e.g. with multidimensional scaling algorithm [5].

3 Case study: visual search vs textual search

Experiment description

To evaluate the proposed approach, a comparison experiment was designed and conducted. Experimental group (17 participants in total varying from undergraduate students to professors, 14 with technical sciences or engineering background, 2 with economical background, 1 with pedagogical background) get the scientific search task, consisting of finding research articles for the subject "influence of social media on mental health". The subject was selected by criteria of being generally understandable by a broad range of scientists, but unconnected to the main scientific interest of all the participants. The visual search (VS) system was briefly introduced to the experimental group (15 minutes presentation of main features). Participants were instructed to find the most relevant articles for a given search task, at the same time covering different aspects of it (to exclude finding multiple highly relevant but very similar articles) and write them down. The same search task was performed twice, once using classical textual search (TS, one of the well-known scientific search solutions) and once using VS approaches. Same amount of time (7 minutes) was allocated for both tasks.

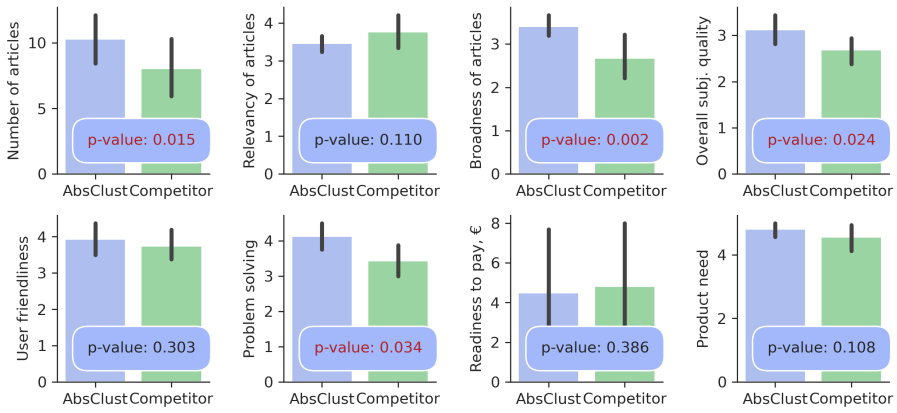

Figure 3: Visualization of metrics, collected during experiment. Metrics, which demonstrated statistically significant difference between AbsClust and competitor are highlighted in red.

Table 1: Summary of metrics, obtained from the experiment

Metric name	Description	Scale	Kind
Number of articles	How many articles have been found during fixed time (7 minutes)	$0-\infty$	Experimental
Relevancy of articles	Participant scored relevancy of articles found by another participant	$1-5$	Experimental
Broadness of articles	Participant scored broadness of articles found by another participant	$1-5$	Experimental
Overall subj. quality	Participant reported overall quality of articles found by oneself	$1-5$	Survey
User friendliness	Participants reported ease of using the given tool	$1-5$	Survey
Problem solving	Participants reported ease of task solving with given tool	$1-5$	Survey
Readiness to pay, €	Participants reported readiness to pay to use the given tool	$0-\infty$	Survey
Product need	Participants reported will to use the given tool if provided for free	$1-5$	Survey

After completing tasks, every participant passed the lists of articles found to the next participant, which scored both lists by assigning relevance score to every of the found articles as well as overall quality and diversity scores. After the experiment, participants were asked to fill the survey collecting individual experiences from the participants regarding tool usability, effectiveness, and satisfaction, providing specific numbers for a holistic evaluation of the tools. Criteria are tabulated in the Tab. 1. After the experiment, the data was processed by calculating p-values for paired t-test using one-sided alternative hypothesis. When a value was missing, the whole pair was excluded from calculating statistics. Final results are visualized in the Fig. 3.

Results discussion

4 out of 8 metrics were found to be statistically significant (with standard 95% level of confidence):

Number of articles Over 7 minutes of experimental time the users found on average 35% more articles using VS compared to TS. This can be attributed to the usage of semantic distance on the map: finding one relevant article means finding simultaneously many other in the vicinity of the first one.

Broadness of articles With this metric we measured how well different aspects of the problem are tackled. Here, VS achieved 30% higher scores when compared to TS. This is connected to the fact that visual map helps to see high-level structure inside the search results. For example, the easiest strategy to find highly relevant but diverse articles would to simple pick one example (perhaps the most cited one) in every cluster.

Overall subjective quality After the experiment, participants reported in the survey that they subjectively score articles, found with VS higher then with competitor, which can be interpreted as a sign of higher satisfaction of VS users compared to textual search, however exact reasons of this difference is unclear due to its subjective nature.

Problem solving According to the survey, participants reported that accomplishing of the task with VS was easier then with competitor. Similar to the previous metric, this might be a sign of higher satisfaction, but exact reasons are unclear due to highly subjective nature of the metric.

Other metrics have demonstrated no statistically significant difference. It is interesting to note that experiment demonstrated no significant difference in articles relevance between AbsClust and competitor, which focuses on relevancy of the search results.

4 Conclusion

We performed initial exploration of the novel VS approach in comparison to traditional TS approach applied to scientific literature search, that yielded evidences of the former advantages. The comparison experiment, involving 17 researchers and a comprehensive set of metrics, found statistically significant improvements across key parameters, including the average time required to locate an article, the breadth of the scope of identified articles, and subjective user satisfaction. These results are an early proof of the potential of the VS system, applied to weakly structured documents (such as scientific articles). These findings call for further investigation and broader comparisons. While our results are promising, a more extensive examination is essential to assess the VS system's effectiveness across diverse user groups and contexts.

References

[1] Fire, Michael, and Carlos Guestrin. "Over-optimization of academic publishing metrics: observing Goodhart's Law in action." GigaScience 8.6 (2019): giz053.

[2] Bornmann, Lutz, Robin Haunschild, and Rüdiger Mutz. "Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases." Humanities and Social Sciences Communications 8.1 (2021): 1-15.

[3] Sinha, Arnab, et al. "An overview of microsoft academic service (mas) and applications." Proceedings of the 24th international conference on world wide web. 2015.

[4] Potter, Mary C., et al. "Detecting meaning in RSVP at 13 ms per picture." Attention, Perception, & Psychophysics 76 (2014): 270-279.

[5] Mead, Al. "Review of the development of multidimensional scaling methods." Journal of the Royal Statistical Society: Series D (The Statistician) 41.1 (1992): 27-39.