Science maps for information retrieval

Bascur Cifuentes, J.P.

Summary in English

Science maps are a widely used tool in scientometric analysis. One of their main advantages is that they reveal the structure of data, which for an analyst can both reveal unknown academic topics and address their own blind spots. These strengths make them well suited for information retrieval tasks, and researchers frequently employ science maps for this purpose. However, most existing research on science maps focuses on the accuracy of topic detection, while much less attention has been paid to their capabilities for information retrieval. This dissertation addresses this gap by exploring one overarching question: What is the effectiveness of science maps for information retrieval, and how can we enhance it?

The dissertation consists of an introduction, four Chapters (Chapters 2–5) that answer subquestions of the overarching question, and a conclusion. Each of these four subquestion Chapters is based on a peer-reviewed publication.

Chapter 1: Introduction

This Chapter defines what science maps are and to which information retrieval tasks they relate. It situates the research within the broader academic literature, introduces the research questions addressed in later Chapters, and presents the additional contributions of the dissertation beyond the research questions.

Chapter 2: An interactive visual tool for scientific literature search: Proposal and algorithmic specification

This Chapter addresses the question "How can science maps be designed to support information retrieval?" by proposing a tool that integrates science maps with an interactive retrieval process. The tool enables users to iteratively "scatter" a set of documents into clusters of related and then "gather" selected clusters into a refined subset. In addition, we developed an algorithm to position clusters in a visualization by minimizing the empty space while preserving the meaning of the distances between the clusters.

Chapter 3: Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews

This Chapter addresses the question "How effective are science maps for making systematic reviews?". In the evaluation we modeled information retrieval as an iterative process of selecting clusters and generating subclusters, using different user models with varying preferences for recall and precision. The results showed that science maps outperformed the boolean queries for about half of the reviews. This indicates that science maps are best used as a complement to, rather than a replacement for, other retrieval tools.

Chapter 4: Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks

This Chapter addresses the question "Do science maps represent some topics better than others?" using Medical Subject Headings as the ground truth for topics. We measured clustering effectiveness for each topic individually and then aggregated by topic category. The best-represented categories were Organisms and Diseases, while Geographical entities were the least well represented. A topic was considered well clustered if its documents were concentrated in a few clusters and those clusters contained few unrelated documents. The evaluation was done on both citation-based and text-similarity-based maps across three different granularities. The analysis also showed that each

category had similar clustering effectiveness on both types of maps.

Chapter 5: Use of diverse data sources to control which topics emerge in a science map

This Chapter addresses the question "How can the representation of specific topics be improved in a science map?" by comparing the performance of topics across science maps created with different data sources. We compared eight sources: text similarity, citations, co-authorship, patents, policy documents, and several forms of social media. The evaluation method is similar to Chapter 4, but modified to facilitate the comparison of a higher number of maps. A key modification was to compare the quality of maps for a given topic only at granularities where the topic was represented by the same number of clusters in both maps, which made the comparison more straightforward and flexible. Results showed that different sources could shift which topic categories performed best, but overall clustering quality tended to decrease when moving beyond text and citation networks. However, this performance loss could be mitigated by merging data sources.

Chapter 6: Conclusion

The Chapter summarizes the findings of the research questions and answers the overarching research question. It emphasizes that science maps can effectively support information retrieval, particularly when combined with other tools, and that science maps vary in their suitability across academic topics. The Chapter also outlines directions for future work, including how to obtain better performance from diverse data sources, the use of large language models, and the importance of prototyping and software sustainability.