Finding and Visualizing Influential Scientific Research Papers for a given research question with GELATO

Meng Li

Abstract

Scientific research papers offer a means for scientists to exchange information with their peers regarding their findings. Reading these papers can help you understand a research field, and gain insights into the key findings. However, the growing scale of scientific research papers can be overwhelming for those who are not well-versed in a particular field, making it challenging to know where to begin. In this paper, we propose a research prototype named GELATO (Graph Explorations: Languages and Tools) and develop an automated system based on it to find and visualize influential scientific research papers in a citation network for any research questions. To provide a clearer understanding of the pipeline, we will focus on sentiment analysis as our research question and highlight the top 3 influential scientific research papers.

Keywords

GELATO; citation network; influential.

1 Introduction

In the past few years, there has been a significant increase in the quantity of scientific research papers that are published in journals and conferences. To assist individuals who are not well-versed in a particular research domain, the development of an automated system capable of identifying and visualizing influential scientific research papers is of great importance.

The task primarily involves two steps: (1) extracting all scientific research papers related to a particular research question, and (2) finding and visualizing the influential papers among them. The first step is to match papers with a set of user-defined words, which are searched for within the title, abstract, and keywords fields of these papers. These user-defined words are keywords relevant to a given research question. The second step is to compute a score for these extracted papers. Sorting the scores, selecting the top K papers, and visualizing them.

Various languages and tools can be employed to process network data, each with its unique strengths and capabilities. Query languages are good at matching patterns, which can be used to extract scientific research papers based on research questions. General-purpose languages are more flexible, which can be used to define advanced scoring functions. GELATO (Graph Explorations: Languages and Tools) is a research prototype which can handle network data using a number of languages and tools, e.g., Python, R, Cypher, SPARQL, and Clingo. For example, the network data originally stored in a Neo4j database could be queried with Cypher first, then the output of it could be imported into a Blazegraph database and queried with SPARQL as well. GELATO makes it possible to reuse the output of one query language or tool as an input to another query language or tool. Two types of functions are available for GELATO. One type are generic functions that are not limited to any specific network dataset, e.g., to calculate the number of nodes and links, average degree, diameter, clustering coefficient, and average path length for a given network. The other are custom functions which are based on the network you are working with and the problems you are interested in. In this case, the scoring function should be written in a custom function.

In a nutshell, the goal of this project is to develop an automated system based on GELATO to find and visualize influential scientific research papers for a given research question in a citation network.

2 Background

The declarative query language Cypher is designed for property graphs stored in Neo4j, a widely used non-relational database. SPARQL is a query language which can retrieve and manipulate data stored in Resource Description Framework (RDF) format. These two languages are widely used for knowledge graph construction (Hogan et al., 2021), and basic measures computation (Warchał, 2012). General-purpose languages, such as Python and R, are popular for advanced metrics computation (Hagberg et al, 2008; Csardi and Nepusz, 2006).

3 Data

We use the latest citation network dataset (Tang et al., 2008), which contains 5,259,858 papers and 36,630,661 citation relationships. The dataset (version 14) is extracted from DBLP and last updated on Jan. 31, 2023. Since DBLP is a bibliographic database that specializes in topics related to computer science. The automated system could handle research questions related to the field of computer science, but might be challenging for other fields due to the lack of data.

4 Methods

In this section, we provide a detailed overview of our approach. We start with data preparation to extract a subgraph of the citation network based on the research field. We then compute basic measures based on GELATO's generic function. Next, we extract papers related to a specific research question based on GELATO's custom function with Cypher queries. Furthermore, we compute a score for each extracted paper and get the influential papers with Python. Finally, we visualize these influential papers with Python. To give an intuitive information, we will focus on sentiment analysis as our research question and highlight the top 3 influential scientific research papers.

4.1 Data preparation

Due to the size of the citation dataset (Tang et al., 2008) (20.12 GB) and the limited computing resources, we filtered the JSON file based on the *fos.name* field before loading it to the Neo4j database to select papers in the field of natural language processing. Figure 1 shows the schema of the citation network stored in the Neo4j database. There are three types of nodes, i.e., Paper, Author, and Venue. The Paper node stores paper ID, title, abstract, keywords, published year, and citation number. The Author node stores author ID, and the Venue node stores paper venue name. There are three types of relationships, i.e., CITED, WROTE, and PUBLISHED_IN. The CITED relationship represents a paper cited another paper, the WROTE relationship indicates authors of a paper, and the PUBLISHED_IN relationship shows the place where a paper was published. A subgraph (334.5 MB) with 204,156 papers, 731,097 CITED relationships, 250,837 WROTE relationships, and 88,464 PUBLISHED_IN relationships were extracted.

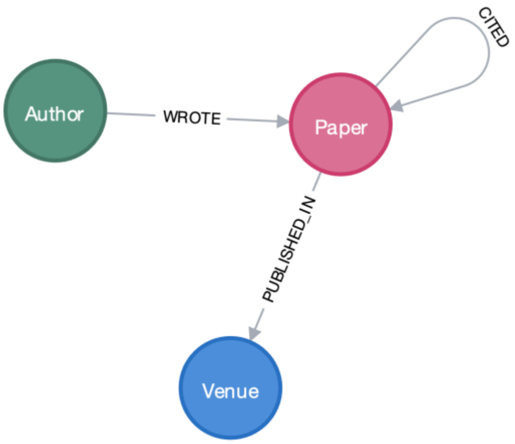


Figure 1: Schema of the citation network.

4.2 Basic measures

This section is to compute basic measures of the citation network in the field of natural language processing. Table 1 shows the basic measures, such as number of nodes and links, average degree, diameter, clustering coefficient, and average path length, computed by the GELATO's generic functions.

Metrics	# of nodes	# of edges	avg degree	diameter	clustering coefficient	avg path length
Value	296,914	1,070,398	7.21	29	0.026	0.002

Table 1: basic measures of the papers in the field of natural language processing.

4.3 Research question extraction

Since we are interested in the research question of sentiment analysis, we need to find all the papers related to it. We match papers with the term of "sentiment analysis" ("Sentiment analysis" and "Sentiment Analysis" are also considered.) in the *keywords*' list. As some papers do not have the *keywords* field, we included papers with the term of "sentiment analysis" in the *title* and *abstract* fields as well. Figure 2 shows the citation network of papers related to sentiment analysis. The network contains 1364 papers and 4644 CITED relationships. Standalone nodes (687 papers) are excluded. One thing I want to emphasize is that we can use a list of words instead of matching the term of "sentiment analysis". With GELATO, you just need to update the configuration file by adding more words to the list.

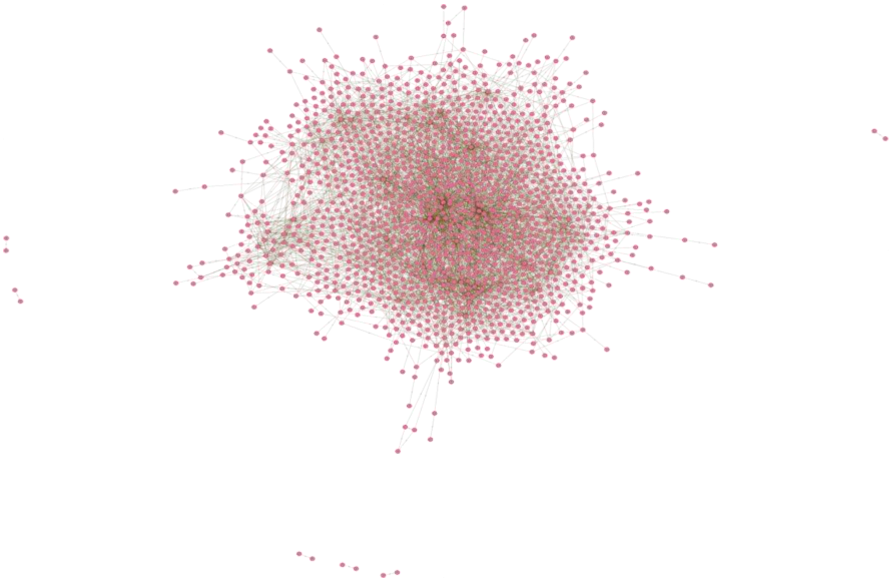


Figure 2: Citation network of papers related to sentiment analysis.

4.4 Influential papers detection

Influential papers detection is a popular topic over the past few decades (Zhang et al, 2016). A score metric should be able to show the importance of each paper. By ranking these papers based on their scores, we can get the top k important papers related to a specific research question (e.g., sentiment analysis). Basically, the higher the score, the more important the paper will be. Therefore, we consider two types of algorithms as the score metric. One is based on the citation number, the other is based on different centrality measures.

Once we selected the top k important papers related to sentiment analysis, we could generate a list of papers in chronological order based on the published years. Table 2 is the list of top 3 influential scientific research papers for the research question of sentiment analysis based on degree centrality.

Paper ID	Year	Title
53e9b8a1b7602d970447a7a2	2005	Recognizing contextual polarity in phrase-level sentiment analysis
53e9bbbcb7602d9704809491	2010	SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining
53e9b891b7602d9704463ede	2011	Lexicon-based methods for sentiment analysis

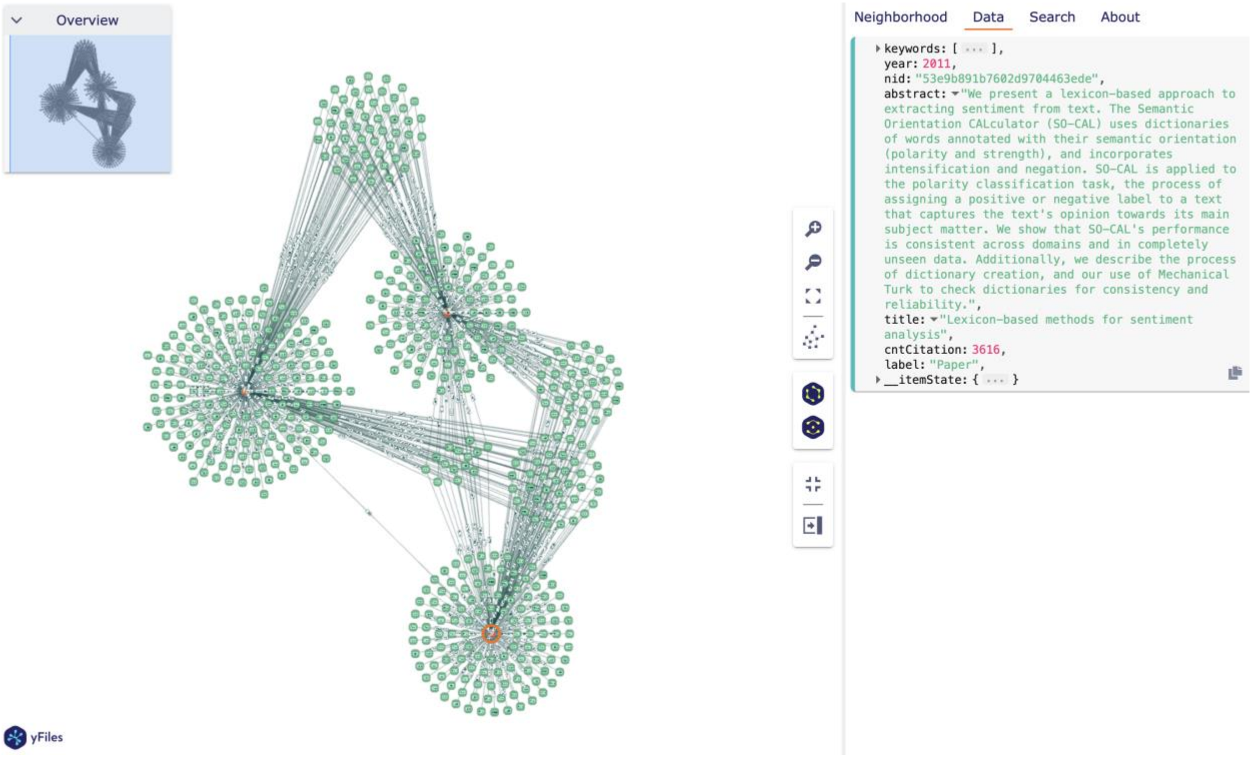
Table 2: top 3 influential papers of sentiment analysis based on degree centrality.

4.5 Influential papers visualization

Since these 3 influential papers have a huge number of citations, we only show the directed connected nodes of these papers while visualizing this network and highlighting these influential papers. As this plot is to assist individuals who are not well-versed in sentiment analysis, we only include the papers with more than 100 citations here.

5 Results

Figure 3 is the graph of top 3 influential papers about sentiment analysis with the default layout based on the yFiles graphs for Jupyter. If you click a node, the detailed information about the selected paper will be shown on the right side. In figure 3, the bottom orange node was clicked.



```
▸keywords: [ ··· ],
 year: 2011,
 nid: "53e9b891b7602d9704463ede",
 abstract: ▾"We present a lexicon-based approach to
 extracting sentiment from text. The Semantic
 Orientation CALculator (SO-CAL) uses dictionaries
 of words annotated with their semantic orientation
 (polarity and strength), and incorporates
 intensification and negation. SO-CAL is applied to
 the polarity classification task, the process of
 assigning a positive or negative label to a text
 that captures the text's opinion towards its main
 subject matter. We show that SO-CAL's performance
 is consistent across domains and in completely
 unseen data. Additionally, we describe the process
 of dictionary creation, and our use of Mechanical
 Turk to check dictionaries for consistency and
 reliability.",
 title: ▾"Lexicon-based methods for sentiment
 analysis",
 cntCitation: 3616,
 label: "Paper",
▸__itemState: { ··· }
```

Figure 3: Top 3 influential sentiment analysis papers visualization with default layout.

Since yFiles generated an interactive HTML file, we can also choose different layouts, e.g., circular layout, hierarchic layout, organic layout, orthogonal layout, radial layout, tree layout, orthogonal edge router, and organic edge router. Figure 4 is an example with the hierarchic layout.

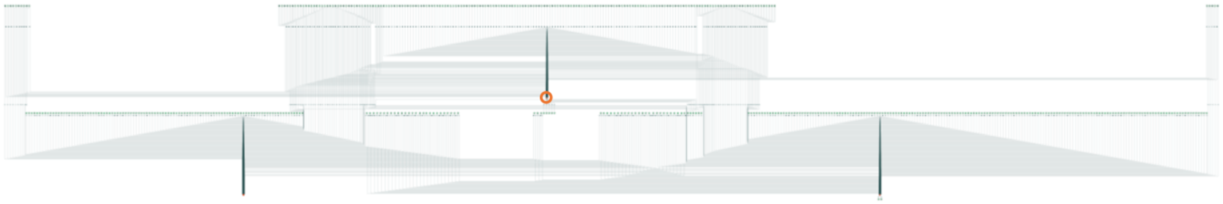


Figure 4: Top 3 influential sentiment analysis papers visualization with hierarchic layout.

6 Conclusion

The automated system based on GELATO can identify and visualize top 3 influential papers for sentiment analysis, which can assist those who are not familiar with this research question. After reading these 3 papers, they will understand this field and know how to start their research.

Top 3 papers extracted here are “Recognizing contextual polarity in phrase-level sentiment analysis”, “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining”, and “Lexicon-based methods for sentiment analysis”. All of them are early works of this field. Once we set higher number of K, more papers will be included: more recent and advanced papers.

This automated system also implies that GELATO can be a useful tool for network analysis project. With GELATO, we can use Cypher to extract all papers related to a particular research question and use Python to compute the scores of each extracted paper. The flexibility of using different languages at different steps makes GELATO a good option for developing an analysis pipeline. Since GELATO is based on the snakemake workflow management system (Mölder et al., 2021), it also supports parallel execution and ensures that each analysis step will only be executed once.

6 Limitations and future work

The scoring function of the automated system only supports citation number and centrality for now. More advanced metrics could be developed with Python or R as a custom function in GELATO in the future.

References

Csardi, Gabor, and Tamas Nepusz. "The igraph software package for complex network research." *InterJournal, complex systems* 1695, no. 5 (2006): 1-9.

Hagberg, Aric, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. No. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane et al. "Knowledge graphs." *ACM Computing Surveys (CSUR)* 54, no. 4 (2021): 1-37.

Mölder, Felix, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster et al. "Sustainable data analysis with Snakemake." *F1000Research* 10 (2021).

Tang, Jie, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. "Arnetminer: extraction and mining of academic social networks." In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 990-998. 2008.

Warchał, Łukasz. "Using Neo4j graph database in social network analysis." *Studia Informatica* 33, no. 2A (2012): 271-279.

Zhang, Sheng, Danling Zhao, Ran Cheng, Jiajun Cheng, and Hui Wang. "Finding influential papers in citation networks." In *2016 IEEE first international conference on data science in cyberspace (DSC)*, pp. 658-662. IEEE, 2016.