

Topic and Affiliation Analysis of Database Publications in Computer Science Field

Team members:

Fatma Dogan

Sona Hasani

1. Objective:

Our objective is to analyze the topics, keywords, author's affiliations and categories of the published papers in databases in computer science field in order to:

- Find the level of collaboration among different research institutes in different research areas
- Find the most popular research topics that attract more collaborators
- Find the level of collaboration between different individuals on different topics

We are going to design and implement an informative visualization of the results of our project to enhance the understanding of the level of collaboration and popularity of the different research topics among the published researchers in different areas of computer science.

2. Dataset:

We choose ACM Digital Library as our dataset since it provides the information that we need including papers titles, authors, their affiliation, keywords and categories. However, the ACM library does not provide any API for programmers to download the required data. We will implement a program to collect the required data from the ACM Digital Library.

3. Deliverables:

- A website will be designed and implemented to introduce the project and the references that are used for the project. Additionally, the website presents a demo of the project that enables users to interact with the project and submit their customized queries and see the results.
- A technical report will be written on the project describing the literature review, design and implementation of the project and evaluation of the results.
- A PowerPoint presentation will be submitted

4. Challenges:

- Our main challenge is to gather the required data from the ACM Digital Library since it does not provide any API for programmers to access and download data. We are going to implement a program to crawl the ACM website and collect the required data.

- Another challenge is the lack of unique identifiers for authors. If there are multiple authors with the same name from the same institute or if the same author is registered with different institutes, it is not possible to identify them using the data provided by the ACM digital library.
- Moreover we need to come up with an efficient and informative visualization in order to convey comprehensive information about the results of our project to the users. We will tackle this challenge by further research in the field of visualization and using the existing visualization techniques for similar data that are proven to be efficient and understandable.

5. Evaluation:

According to our studies so far, a similar project has been done on affiliation analysis by a database research team in Leipzig University. (http://dbs.uni-leipzig.de/en/publication/title/affiliation_analysis_of_database_publications). They do not use the ACM digital library but the final result of our project is comparable with theirs. We also plan to conduct further research on existing similar projects and find some benchmarks that we can compare our final results with them. Additionally we are going to ask some of the students in the class to use our final product and provide feedback on the usability and efficiency of our project.

6. Partitioning the tasks:

We partitioned the tasks but we have not divided the tasks between the team members yet. We will provide the detailed assignment of the tasks to the team members with our report.

- Literature review on existing work on affiliation analysis and visualization of similar data
- Data preparation including implementing the data collection tool, cleaning the data, and creating the dataset
- Implementing the main project on retrieving the collaboration between authors and institutes on different topics using the provided dataset
- Design and implementation of the visualization of the results
- Creating a website to introduce the project with an interactive demo of the project
- Writing a technical report on the project describing different steps of this research from literature review to evaluation of the final results
- Creating presentation slides