Topic and Affiliation Analysis of Publications in Computer Science Field

Midterm Report

**Team members:**

Fatma Dogan

Sona Hasani

# Objective:

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

* Find the level of collaboration among different research institutes in different research areas
* Find the changes in collaboration between different research institutes over time and for different research areas
* Find the most popular research topics that attract more collaborators
* Find the trending research topics over years and among different research institutes
* Find the most similar research institutes in terms of publication in certain research topics
* Find the pattern of collaboration between different research institutes based on their geographical location

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 different research topics among different research institutes in different areas of computer science.

# Change of initial plan

Initially we planned to use individual researchers’ information too but after discussing about different aspects of the project we decided to limit our analysis only to the research institutions (authors’ affiliations) and not the individual researchers. Moreover, at first we wanted to analyze the title of the papers as a part of topic analysis, however later we decided to limit the topic analysis only to the provided keywords and tags for each paper since there are multiple tags and keywords are provided for each paper and they include a comprehensive collection of the different topics that the paper belongs to.

# Dataset:

We chose ACM Digital Library as our data source since it provides the information that we need including papers titles, authors’ name and affiliation, keywords, tags, publisher, publication year, and categories. By far we have collected data from Transactions on Database Systems (TODS) **(add number of records)** and VLDB**(add number of records)**. ACM Digital Library does not provide any API for programmers to download the required data, therefore, we developed a program in python to crawl it and collect the data we require.

# Components of the Project:

* A data collector program written in python to crawl the ACM Digital library on a list of predefined conferences and journals and collect data about authors, affiliation, publication year, keywords, and tags.
* A data analysis tool will be implemented to use the collected data and retrieved the information for customized input queries
* A visualization tool will be implemented to visualize the result of the data analysis tool
* 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 visualization tool that enables users to interact with the project and submit their customized queries and see the results.

# Tasks to be accomplished in future:

* Creating the project website with interactive interface that enables the user to submit customized query to the system and receive statistics as well as the visualization of the results of their query online
* Design and implement the database and import our collected data into that to utilize the data analysis tasks
* Developing data analysis tool to answer the predefined questions that users ask about different aspects of research collaboration between research institutes
* Implementing the visualization component and embedding it into the project website in order to enable user to have a better understanding of the answer to their query

# Challenges:

* Our main challenge was to collect the required data from the ACM Digital Library since it does not provide any API for programmers to access and download the data. We implemented a program to crawl the ACM Digital Library and collect the required data. Challenges we faced after collecting the data are :
* Difference in encoding for non-English characters in authors names
* Missing values for some of the fields in some papers: affiliation, volume, keywords…
* Mismatch between the number of authors in a paper and number of affiliations (two authors from the same research institute)
* ACM Digital Library website blocks the data collecting program after a certain amount of data is collected. We need to spread the data collection task over time. Therefore, collecting all of the data takes much longer time that what we initially planned.
* To find and use an optimal visualization tool and embed it into our website in order to make it an interactive and effective visualization tool
* To find the geographical location of the research institutes
* Finding an accurate method to evaluate the accuracy of the outcome of our project is another major challenge for us.

# Evaluation:

According toour 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. Also the ACM Digital Library provides some statistics on the number of the papers that research institutes published together. We would like to find a way to use those statistics to evaluate our results too. We 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.