Topic and Affiliation Analysis of Publications in Computer Science Field

Project Report

Team members:

Fatma Dogan

Sona Hasani

1. Introduction:

In this work we study the collaboration between different organizations (Universities, Research Labs, Private companies, …) on scholarly published papers in Computer Science and Engineering field. We find the amount of collaboration between different organizations based on the number of the papers that they have published together. We also compare the similarity of every two organizations using the keywords of the papers that they have published. For this study we use the information about published data in the conferences under ACM digital library. We provide an interactive visualization for our results.

2. Dataset

For this project we used the 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. 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. We used BeautifulSoup library for this purpose.
Following is some of the challenges that we faced in the process of collecting the data:

- 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.

2.1 Data Collection

In total we collected the information of 74316 papers which is 1/6 of the total papers in ACM digital library today. The conferences which we collected their data and the number of papers collected from each are as following:

VLDB : 2352

KDD : 2457

SIGMOD/POD : 507

CIKM : 4692

SIGMOD : 3116

PODC : 1638

ANLC : 291

WWW : 3392

ISSS : 360

SIGIR : 3450

GIS : 1204

SIGGRAPH : 11135

SIGDOC: 1046

POPL:1706

ICCAD:3024

LNCS:19993

HPDC: 1202

ICSE : 7245

ISCA : 1780

ICMI : 1122

ICS : 1307

CCS : 2194

The following is a sample of one single record from our dataset.

journal_name: Null
publisher_name: ACM
Authors: Fram, David M.; Almenoff, June S.; DuMouchel, William
Title: Empirical Bayesian data mining for discovering patterns in post-marketing drug safety
Date: 8/24/2003
Volume:Null
Issue:Null
Issn:Null
first_page: 359
last_page: 368
Keywords: association rules; data mining; empirical Bayes methods; pharmacovigilance; post-marketing surveillance
author_info: ['David M. Fram', '81546930756', 'June S. Almenoff', '81100442362', 'William DuMouchel', '81100642918']
affiliation_info: ['GlaxoSmithKline, Research Triangle Park, NC', '60020649', 'AT&T; Shannon Laboratory, Florham Park, NJ', '60008383']
Tags:['association rules', 'data mining', 'data mining', 'empirical bayes methods', 'pharmacovigilance', 'post-marketing surveillance', 'scientific databases']

2.2 Data Preparation:

After some basic data cleaning we performed the following operations on the data to make it more consistent and usable.

- Among the 74316 papers we collected only 63069 had affiliation information therefore we did not use 11247 records of our data.
- We combined keywords and tags associated with each paper and removed the duplicates. In the remaining of the paper when we refer to keywords, we mean the combined list of keywords and tags.
- All of the affiliations of different authors are concatenated in one single string named “affiliation info”. We spitted the affiliations, and removed the duplicates. For this word we only focus on the main organization level. The different departments of the same organization share the same unique ID associated with that organization.

3. Data Analysis

We studied our data from different aspects. But in this report we focused on two main approaches and we leave the rest for future work.

3.1 Collaboration

We defined the collaboration number of each organization as the number of the papers that it has published with co-authors from other organizations.

The following table shows the top 50 most collaborative organizations.

	Organization Name	Published papers	Collaborated papers
1	Microsoft Research	1169	835
2	Carnegie Mellon University	1407	665
3	IBM Thomas J. Watson Research Center	949	615
4	U. C. Berkeley	855	441
5	Stanford University	900	433
6	University of Illinois at Urbana-Champaign	918	432
7	International Business Machines Corporation	674	406
8	Microsoft Research Asia	457	396
9	Microsoft Corporation	598	389
10	M.I.T.	725	378
11	Yahoo! Research	486	326
12	Tsinghua University	532	314
13	Georgia Institute of Technology	555	302
14	National University of Singapore	529	300
15	Intel Research Berkeley	379	282
16	University of Toronto	572	280
17	The University of Tokyo	582	264
18	University of Texas at Austin	554	261
19	Google Inc.	363	261
20	IBM Almaden Research Center	430	258
21	Purdue University	555	254
22	University of Maryland	522	250
23	ETH Zurich	480	249
24	University of California San Diego	461	248
25	Cornell University	544	242
26	IBM Research Bengaluru India	363	241
27	Hong Kong University of Science and Technology	333	231

28	Technion Haifa Israel	423	230
29	Microsoft Research Ltd. United Kingdom	285	225
30	University of Southern California	522	216
31	UCLA	422	211
32	INRIA	316	210
33	Bell Laboratories	387	208
34	University of Waterloo	467	206
35	Rutgers University	325	203
36	The Pennsylvania State University	402	201
37	Chinese Academy of Sciences	373	199
38	University of Wisconsin-Madison	438	198
39	Tel Aviv University	291	197
40	EPFL	333	195
41	Hewlett-Packard Labs	300	193
42	Columbia University	336	185
43	Princeton University	330	185
44	Peking University	298	180
45	Technische UniversitÃ¤t MÃ¼nchen Germany	390	178
46	University of California Irvine	396	177
47	Arizona State University	307	170
48	University of British Columbia	383	167
49	University of Massachusetts	478	166
50	The Chinese University of Hong Kong	314	165

We also studied the collaboration between the most collaborative organizations. The following table shows a sample of the collaboration between some of the most collaborative organizations.

Organization 1	Organization 2	collaboration
Microsoft Research	Microsoft Corporation	96
Microsoft Corporation	Microsoft Research	96
Tsinghua University	Microsoft Research Asia	73
Microsoft Research Asia	Tsinghua University	73
IBM Thomas J. Watson Research Center	International Business Machines Corporation	71
International Business Machines Corporation	IBM Thomas J. Watson Research Center	71
Stanford University	Microsoft Research	49
Microsoft Research	Stanford University	49
Hong Kong University of Science and Technology	Microsoft Research Asia	45
Microsoft Research Asia	Hong Kong University of Science and Technology	45
Peking University	Microsoft Research Asia	42
Microsoft Research Asia	Peking University	42
University of Illinois at Urbana-Champaign	Microsoft Research	40
Microsoft Research	University of Illinois at Urbana-Champaign	40
Microsoft Research	Carnegie Mellon University	40
Carnegie Mellon University	Microsoft Research	40
University of Illinois at Urbana-Champaign	IBM Thomas J. Watson Research Center	32
IBM Almaden Research Center	International Business Machines Corporation	32
IBM Thomas J. Watson Research Center	University of Illinois at Urbana-Champaign	32
International Business Machines Corporation	IBM Almaden Research Center	32
Microsoft Research	U. C. Berkeley	32
U. C. Berkeley	Microsoft Research	32
Microsoft Research	Microsoft Research Ltd. United Kingdom	27
Microsoft Research Ltd. United Kingdom	Microsoft Research	27
Georgia Institute of Technology	Microsoft Research	26
Microsoft Research	Georgia Institute of Technology	26

3.2 Similarity

In this section we describe how we measured the similarity between different organizations. For each paper we have a list of keywords and for each affiliation we have a list of papers that they have published. For each affiliation we generate a vector of keywords. We use TF-IDF to generate the proper value for each keyword.

We define the following variables:

Keyword frequency: number of the papers containing this keyword that is published by organization X.

Total Organizations: total number of the organizations (we have 3371 unique organizations)

Related Organizations: number of the organizations that have published at least one paper containing the keyword

We calculated the TF-IDF value of a specific keyword for a specific organization as following:

Keyword TF-IDF value = log (1+ Keyword frequency) X log(Total Organizations/ Related Organizations)

In this idea we look at the organizations as documents and keywords as terms.

Following is a sample of keyword vector that we have generated for each of the 3371 organizations in our data set.

{'learning': 3.4167411292366348, 'database applications': 3.316248743082616, 'data mining': 1.2943521879888784, 'selection process': 4.926078230015925, 'concept learning': 13.009898915170263, 'design': 1.076110408677922, 'software engineering education': 21.590896071984695, 'measurement': 1.3023285422348032, 'management': 1.3913285031660745, 'biology and genetics': 5.899837880135352, 'uml': 7.5168304843205975, 'languages': 1.6753914058958146, 'unified modeling language': 63.42325721145504, 'object-oriented programming': 9.142091129579105, 'performance': 1.501738613522447, 'life and medical sciences': 59.692477375487094, 'theory': 1.4868455903051732, 'curriculum': 13.009898915170263, 'case tools': 112.75245726480895, 'query formulation': 2.8910886478156144, 'human factors': 1.5230830962689226, 'learning styles': 507.3860576916403, 'algorithms': 0.6264025403600497, 'computer-aided software engineering': 13.713136694368657}

After generating this vector for every organization, we calculate the cosine similarity between every pair of organizations.

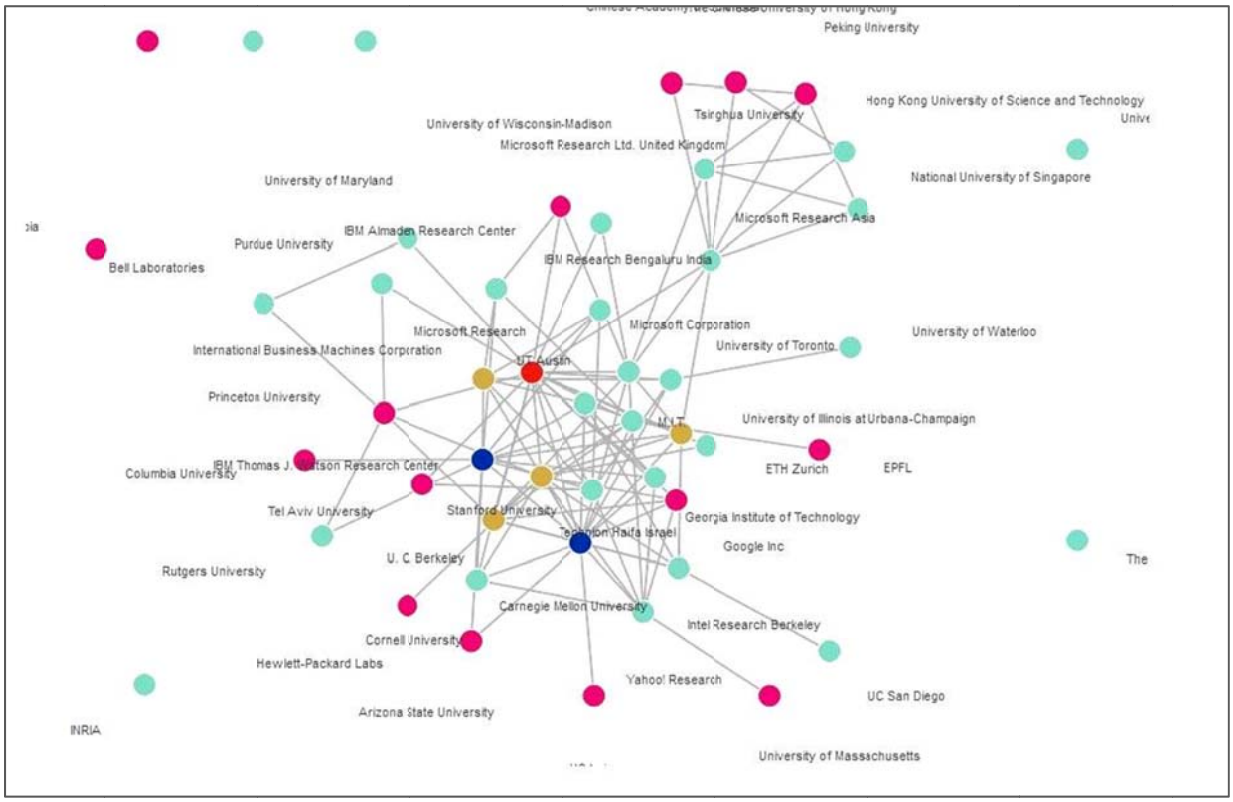
$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

The following table show a sample of the similarity between a pair of organizations.

Organization 1	Organization2	Similarity %
University of Pennsylvania	The Smith-Kettlewell Eye Research Institute	0.833013584
University of Pennsylvania	Yahoo! Research	0.754596108
University of Pennsylvania	Microsoft Research Ltd. United Kingdom	0.63891852
University of Pennsylvania	University of Toronto	0.567118533
University of Pennsylvania	Ludwig-Maximilians-Universitaet Germany	0.538554267
University of Pennsylvania	Google Inc.	0.460532083
University of Pennsylvania	Tel Aviv University	0.412254675
University of Pennsylvania	New York University	0.409646849
University of Pennsylvania	Microsoft Research	0.363367128
University of Pennsylvania	University of Virginia	0.354731421
University of Pennsylvania	University of Illinois at Urbana-Champaign	0.331467921
University of Pennsylvania	IBM Almaden Research Center	0.32282723
University of Pennsylvania	U. C. Berkeley	0.287373546
University of Pennsylvania	Dow Jones Co. USA	0.285225933
University of Pennsylvania	W3C INRIA RhÃ´ne-Alpes France	0.26824029
University of Pennsylvania	Ã‰cole Polytechnique France	0.266634753
University of Pennsylvania	Carnegie Mellon University	0.255097411
University of Pennsylvania	IBM Research Bengaluru India	0.233136436
University of Pennsylvania	Sprint Advanced Technology Labs	0.228299989

4. Visualization

We used D3 to visualize our results. For visualization we only show the top 50 most collaborative affiliations. We count the number of the papers that the organizations have published in collaboration with other organizations. We generate a graph with 50 vertices. Each vertex is an organization. We create an edge between two organizations only if the number of the papers that they have published together is at least 10. On each edge we show the number of the papers that those two organizations have published together. When you click on one node, the top twenty most similar organizations to that university are shown on the right side of the window. If any of those similar universities are among the collaborators of this organization, they will be highlighted in the graph.



5. Results and Discussions:

Our Results show that the organizations with higher amount of collaboration have strong collaborations with each other and also they have higher number of publications in total. Since we measure the similarity between different organizations considering all of the available keywords, the similarity value is very low. For our future work we plan to measure the similarity for different research topics.

6. Future work

We would like to continue this works in the future and achieve the following objectives.

- Study the specialized research collaboration on different research areas.
- Study the collaboration pattern between individuals.
- Study the relationship between the collaboration and physical distance of the organizations
- Enhance the interactivity of our website in order to 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 narrowed down by research.
- Find the connected components or semi connected components in the graph to find the group of organizations that are most likely to collaborated with each other than other organizations.

7. References:

http://d3js.org/
http://dl.acm.org/
https://pypi.python.org/pypi/beautifulsoup4