Conquering DBLP – A short study

Usama Khalid and Shah Saood Khan
Department of Computer Science
National University of Computer and Emerging Sciences

Abstract— the dblp data records of 7.1 million publications are read line by line and processed via parsing for the proposed ERD of the dblp dataset. The Focus of Research (FoR) of each author was found from FoR of the journal/conference in which it was published. If an author published more papers in journal rather than conference, then focus of research of author will be declared as Focus of Research of J1 because of greater number of publications. A GUI application was designed to visualize the results/statistics. Then all the authors having same Focus of Research have been found and the output was shown in GUI in the form of list. Then all the authors having same Focus of Research and have published at least 'x' papers together, are shown. The output was in the form of graph (nodes and edges) where each node is attached with name of the author and each edge displays number of published papers. using two classification algorithms such as Naive Bayes and Decision tree the number of papers to be published in each Focus of Research were predicted and the numbers of papers published in a given conference or journal in a given were predicted.

***Index Terms*—Focus of Research, author and conference.**

I. Introduction

DBLP is a computer science bibliography website. Starting in 1993 at the University of Trier, it grew from a small collection of HTML files [1] and became an organization hosting a database and logic programming bibliography site. DBLP listed more than 4.86 million journal articles, conference papers, and other publications on computer science in December 2019. [2]

DBLP originally stood for *Database systems and Logic Programming*. As a backronym, it has been taken to stand for *Digital Bibliography & Library Project*;[3] however, it is now preferred that the acronym be simply a name, hence the new title "The DBLP Computer Science Bibliography"[4]
[1] The dblp XML format is modeled after the BibTeX *.bib file format. The format is defined in the document type definition (DTD) file in the same directory. By design our DTD is not very strict, as it makes no restriction to element order or multiplicity, and even allows nonsensical elements (e.g., ‹school› in <article> elements, <editor> and <author> elements at the same time that won't be found in the actual

dblp data set. Our priority was to keep the definition clean and simple, and not to model every aspect of the publication landscape.

The main background of the problem is the lacking citation information and the varying coverage for different subfields of computer science. In DBLP many ad hoc solutions are poorly designed. If a person has several names (synonyms) or if there are several persons with the same name (homonyms) then the mapping becomes tricky. The main obstacles are to abbreviate given names beyond recognition and spelling errors. The main algorithmic idea is of the use to identify names we should check more precisely, is to look at person pairs which have the distance two in the co-author graph and have a "similar" name.
Having all the dblp data organized in a relational database we identified the FoR (Focus of Research) of each author. If an author named A1 published 3 papers in journal named J1 and 1 paper in conference named C1, then FoR of A1 will be declared as FoR of J1 because A1 has published more papers in J1.
The data about FoR is not present in dblp but can be obtained from http://portal.core.edu.au. Each author can publish in a journal or conference and each of them has FoR. The most number of publications of an author of FoR is their main Field of Research. A GUI application was then designed to visualize the results like total FoRs, total authors and total publications.
Next the authors having same Focus of Research (FoR) and having published the same paper together (co authorship) were identified i.e. all the authors having same FoR and have published at least 'x' papers together. Then the co authorship between them was displayed via graph. The nodes in the graph represent authors while the edges represent number of publications between them.

II. Literature Review

For computer science researchers the DBLP web site is a popular tool to trace the work of colleagues and to retrieve bibliographic details when composing the lists of references for new papers. Ranking and profiling of persons, institutions, journals, or conferences is another sometimes controversial usage of DBLP. The DBLP data may be downloaded. The bibliographic records are contained in a huge XML file[5]

The main disadvantages of DBLP for this purpose are the lacking citation information and the varying coverage for different subfields of computer science [6]. The main advantages are the free availability and the inclusion of many conference proceedings which play an essential role for many branches of CS and are poorly covered by other bibliographic data bases.

The DBLP data set is available from the location http://dblp.uni-trier.de/xml/. Which are published in journals, transactions, magazines, or newsletters. The second part of a DBLP key typically designates the conference series or periodical the papers appeared in. The last part of the key may be any sequence of alphanumerical characters, in most cases these IDs are derived from the authors' names and the year of publication, sometimes a letter is appended to make this key part unique. The file dblp.xml contains all bibliographic records which make DBLP. It is accompanied by the data type definition file dblp.dtd. You need this auxiliary file to read the XML file with a standard parser [7].

III. Proposed Methodology

A. *Overview*

We have proposed four solutions corresponding to the problems targeted in this research. The proposed algorithms are listed below. The first problem was to parse and load the dblp.xml file into a database. So we used python for parsing and used MySQL database for storing all the data. The next problem was to find the Focus of Research (FoR) of each author so we proposed that we will consider the maximum number of publications the author has done in a Field and make that their FoR.

The third problem we targeted was to build a graph to display the number of publications each author has with every other author for a given FoR. We combined each author with every other author for a given FoR and counted their publications then we displayed this data in a graph. We used visJS a JavaScript library to display graphs.

The last problem targeted was to predict the number of publications in the future. We used naïve Bayes and decision tree classifiers and the data from the database to train our models.

B. *Algorithms*

1. Parsing and loading dblp.xml –

```
elements ← ['article','proceeding','book'...]
for tag in dblp.xml
if (tag in elements)
        data ← parse(tag)
        insert_into_database(data)
```

2. Finding Focus of Research –

```
journals ← get_author's_journals(author_id)
conferences ← get_author's_conferences(author_id)
union ← conferences ∪ journals
max_publications(union)
```

3. Finding coauthors that have X papers together –

```
authors ← get_author's_in_FoR(FoR)
coauthors ← authors ∪ authors
coauthors_X ← atleast(X,coauthors)
show_graph(coauthors_X)
```

4. Predict Publications in FoR,Journal/Conference –

```
model ← train_data(csv)
predict(model,(year,id))
```

C. *Figures*

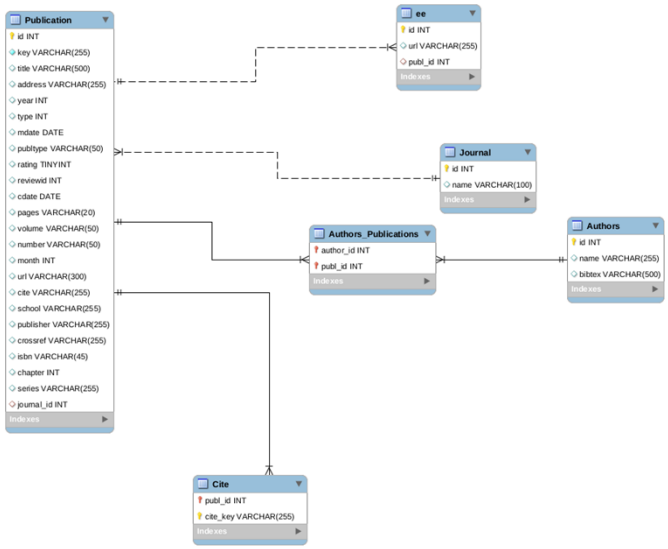

Figure 1: Proposed ERD for loading dblp data

IV. Results / Experiments

We made a comprehensive GUI for all the solutions we proposed. Figure 2 shows the main page of the GUI which can be used to view all Authors/ Publications/FoRs, Search for an Author, Make graphs for Coauthors, Predict publications in Journal/Conference/FoR by year.

Below are all the screenshots showing different use cases of the comprehensive GUI.

A. *Figures*

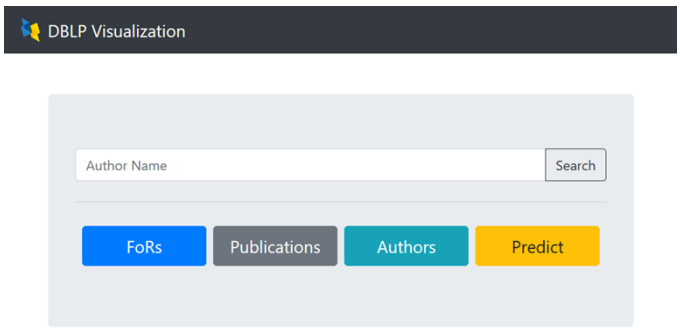

Figure 2: Main page of GUI application

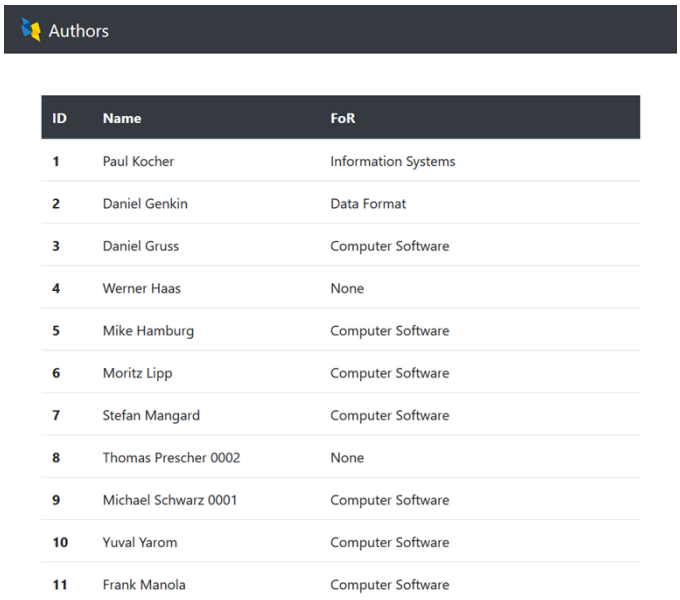

Figure 3: List of all authors in GUI

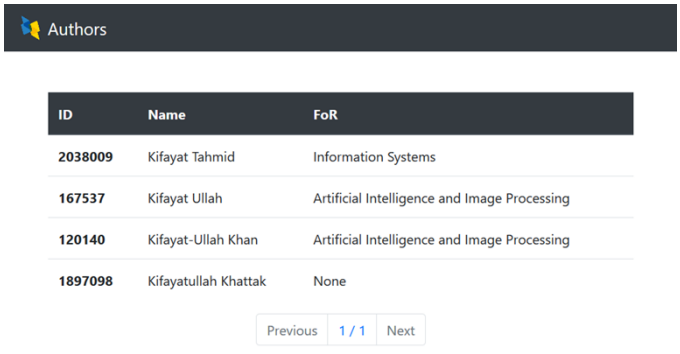

Figure 4: result of a search query

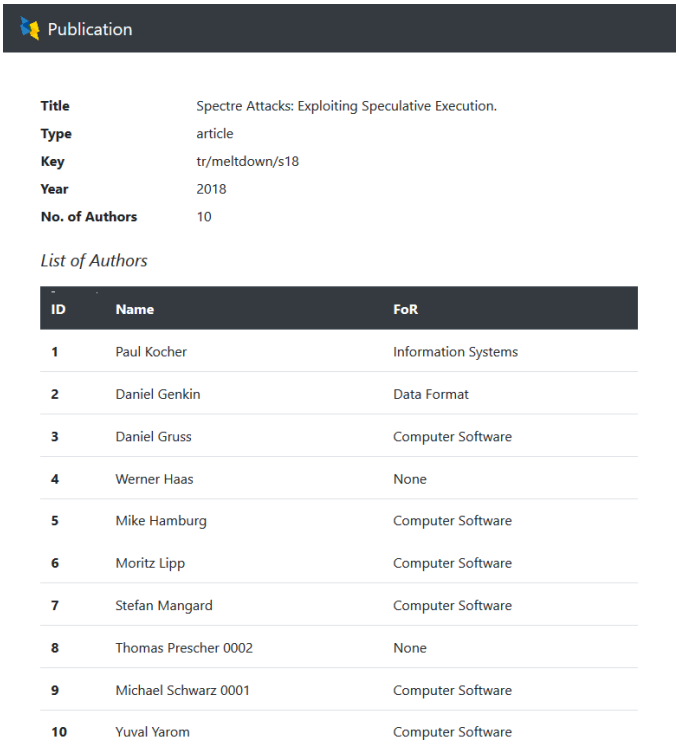

Figure 5: Details page for a particular publication

Author

Name	Kifayat-Ullah Khan
Focus of Research	Artificial Intelligence and Image Processing
No. of publications in FoR	4
Total no. of publications	27

List of Publications

ID	Key	Title	Type	Year
3881169	homepages/131/6850	Home Page	www	None
1399486	journals/corr /NawazHKL13	Personalized Email Community Detection using Collaborative Similarity Measure.	article	2013
5013649	conf/bdcloud /KhanNL14	Set-Based Unified Approach for Attributed Graph Summarization.	inproceedings	2014
5013587	conf/bdcloud /NawazKL14	CORE Analysis for Efficient Shortest Path Traversal Queries in Social Graphs.	inproceedings	2014
1394429	journals/corr /NawazKL14	Shortest Path Analysis in Social Graphs.	article	2014
1319423	journals/corr /KhanNNL14	OLAP on Structurally Significant Data in Graphs.	article	2014
1352117	journals/corr /NajeebullahKNL14	BPP: Large Graph Storage for Efficient Disk Based Processing.	article	2014
1495652	journals/dpd /NawazKLL15	Intra graph clustering using collaborative similarity measure.	article	2015
1676094	journals/apin /NawazKL15	SPORE: shortest path overlapped regions and confined traversals towards graph clustering.	article	2015
6499684	conf/icuimc/KhanNL14	Lossless graph summarization using dense subgraphs discovery.	inproceedings	2015
2116916	journals/computing /KhanNL15	Set-based approximate approach for lossless graph summarization.	article	2015
4815414	conf/icde/Khan15	Set-based approach for lossless graph summarization using Locality Sensitive Hashing.	inproceedings	2015
5615750	conf/edb/DuongKJL16	Top-k frequent induced subgraph mining using sampling.	inproceedings	2016

Figure 6: Details page of a particular author

Publications

ID	Key	Title	Type	Year
1	tr/meltdown/s18	Spectre Attacks: Exploiting Speculative Execution.	article	2018
2	tr/meltdown/m18	Meltdown	article	2018
3	tr/acm/CS2013	Computer Science Curricula 2013	book	2013
4	tr/gte /TR-0263-08-94-165	An Evaluation of Object-Oriented DBMS Developments: 1994 Edition.	article	1994
5	tr/gte /TR-0222-10-92-165	DARWIN: On the Incremental Migration of Legacy Information Systems	article	1993
6	tr/gte /TR-0174-12-91-165	Integrating Heterogeneous, Autonomous, Distributed Applications Using the DOM Prototype.	article	1991
7	tr/gte /TM-0149-06-89-165	Object Model Capabilities For Distributed Object Management.	article	1989
8	tr/gte /TR-0310-11-95-165	Integrating Object-Oriented Applications and Middleware with Relational Databases.	article	1995

Figure 7: List of some publications in GUI ordered by ID

FoRs

#	ID	Focus Of Research
1	8	Information and Computing Sciences
2	801	Artificial Intelligence and Image Processing
3	802	Computation Theory and Mathematics
4	803	Computer Software
5	804	Data Format
6	805	Distributed Computing
7	806	Information Systems
8	807	Library and Information Studies
9	899	Other Information and Computing Sciences
10	902	Automotive Engineering
11	905	Civil Engineering
12	906	Electrical and Electronic Engineering
13	909	Geomatic Engineering
14	910	Manufacturing Engineering
15	913	Mechanical Engineering
16	999	Other Engineering
17	1005	Communications Technologies
18	1006	Computer Hardware
19	1117	Public Health and Health Services
20	1203	Design Practice and Management
21	1503	Business and Management

No. of papers published together

2

Figure 8: List of all Focus of Rsearches in GUI

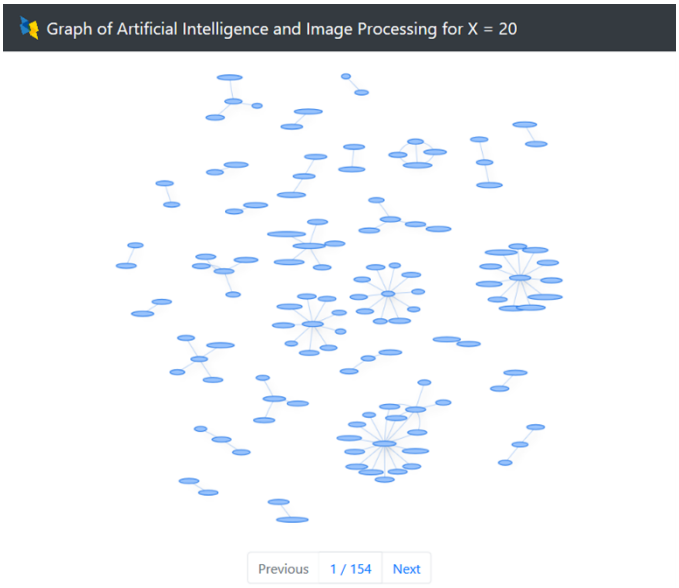

Figure 9: Graph of Authors that have published at least 20 papers in Artificial Intelligence and Image Processing

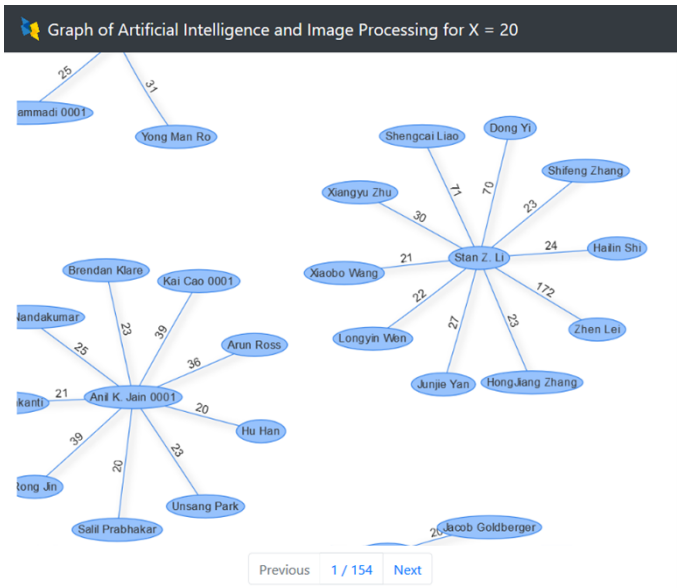

Figure 10: Graph of Figure 9: zoomed to display edge weights

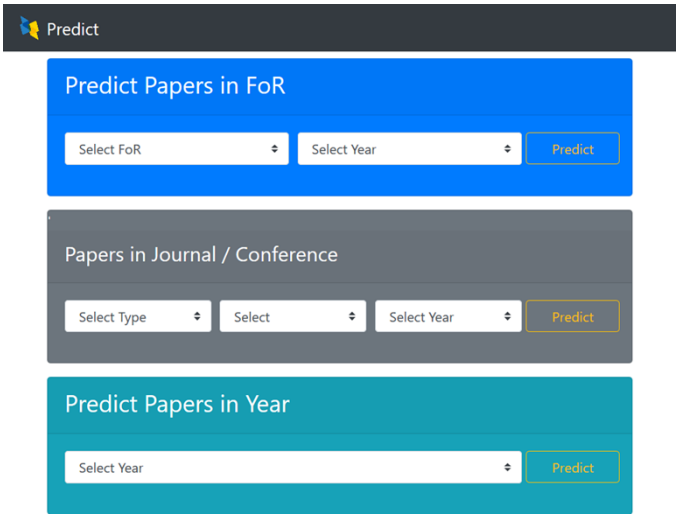

Figure 11: Interface to perform predictions based on papers published in FoR, Journal/Conference or simply by year

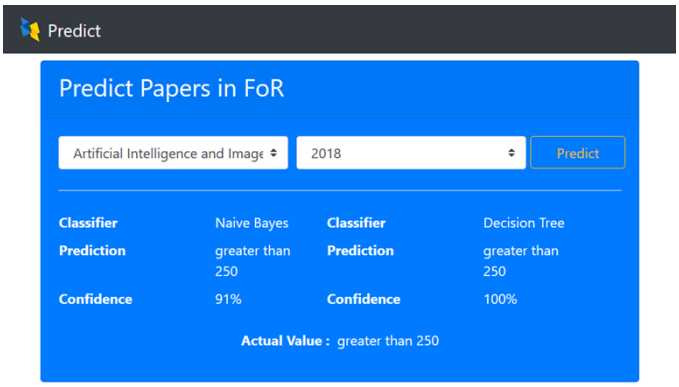

Figure 12: An example prediction for the number of papers published in Artificial Intelligence and Image Processing in the year 2018. Both the Decision Tree as well as Naive Bayes predict correctly but the Decision Tree is more confident in its prediction.

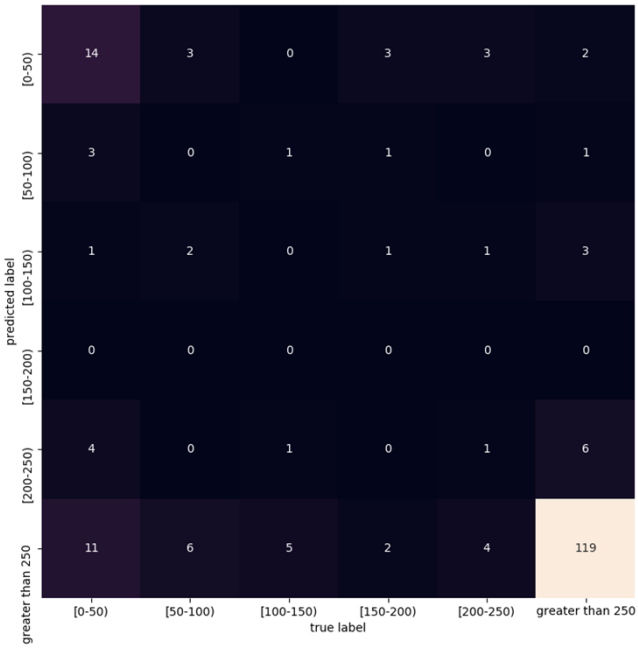

Figure 13: Confusion matrix for the Naive Bayes Classifier that predicts the number of papers to be published in an FoR given the year.

As you can see from the confuxion matrix the probabilty of the number of publications being greater than 250 is significantly large so predictions can be wrogly assigned this label. This decreases our accuracy of prediction.

V. Conclusion

We wrote a python script to scrape data of both Journals and conferences. We modified our scrapping script to go on each journal and conference page and extract the title. Because the FoR title was mentioned in the same tag as the FoR ID and we had the FoR ID so we could search for the tag. The 'key' attribute of each publication contained the acronym of the journal or conference it belonged to. So we extracted those acronyms and we also had acronyms from core's data of conferences. So we were able to link up publications with their conferences and find their FoRs.

We used SQL's 'LIKE' operator that can match words of variable length to match up the titles. e.g. A title of a journal in dblp was "Inf. Sci." and real title was "Information Sciences". So we collected all the matching titles and selected the title with the least difference in length between those two. We had to host the server as well as the 4 GB dblp data somewhere online to ensure 24 hours access for the Android App. So we used Google Cloud SQL to host our database and Heroku to host our python web server.

We uploaded mysql server to Google cloud for faster query processing. We also optimized the query further and we were able to find co-authors of a single FoR in a few hours. We designed web based GUI and used vis-js to draw graphs. We Pre- computed and stored information about FoR and the number of papers each author has published with every other author. This happened because there were thousands of authors for a particular FoR with 0.4 million being the highest. We set a limit on the number of nodes that will be displayed at a time. We made an edge list for all graphs and stored in our database so that graph plotting becomes fast.

References

*[1] Ley, Michael (2009). DBLP: Some Lessons Learned (PDF). VLDB. Proceedings of the VLDB Endowment. **2** (2). pp. 1493-1500. CiteSeerX 10.1.1.151.3018. doi:10.14778/1687553.1687577. ISSN 2150-8097. Retrieved March 13, 2018.*

[2] *"Records in DBLP". Statistics. DBLP. Retrieved December 1, 2019.*

[3] *Ley, Michael; Reuther, Patrick (2006). "Maintaining an Online Bibliographical Database: the Problem of Data Quality" (PDF). EGC, ser. Revue des Nouvelles Technologies de l'Information. RNTI-E-6: 5–10. CiteSeerX 10.1.1.67.6180. Retrieved May 1, 2018.*

[4] *"What is the meaning of the acronym dblp?". FAQ. DBLP. Retrieved March 13, 2018.*

[5] Personal name. Wikipedia, 2009.

[6] [6] A. H. F. Laender, C. J. P. de Lucena, J. C. Maldonado, E. de Souza e Silva, and N. Ziviani. Assessing the research and education quality of the top Brazilian Computer Science graduate programs. SIGCSE Bulletin, 40(2):135–145, 2008.

[7] [7] T. C. Lam, J. J. Ding, and J.-C. Liu. XML document parsing: Operational and performance characteristics. IEEE Computer, 41(9):30–37, September 2008