PubCS: A Massive Dataset of Scientific Articles in Computer Science Domain

ABSTRACT

We introduce *PubCS*, a massive dataset of scientific articles in computer science domain. The dataset is curated from Microsoft Academic Search website and manually processed to serve as a computational resource. PubCS contains more that 1.5 million scientific papers along with a set of metadata information for each paper – the title of the paper, a unique index for the paper, its author(s), the affiliation of the author(s), the year of publication, the publication venue, the related field(s) of the paper, the abstract and the keywords of the papers. We also present a number of statistics about various networks generated from the dataset and several other statistics pertaining to the metadata information. The dataset is publicly available at `http://cnerg.org` to facilitate scientific research.

1. INTRODUCTION

Extraction and mining of academic social networks have, of late, become an extensively popular topic of research as it allows one to have a clear picture of the underlying principles of the dynamics of scientific research. In an academic social network, people are not only interested in searching for different types of information (such as authors, conferences, and papers), but are also interested in finding semantic information (such as structured researcher profiles). Most of the existing datasets provide only the network information of the publication dataset [4, 8]. However, these datasets are often insufficient for mining because of a couple of reasons as follows: (1) lack of semantic information (the information obtained from user-entered profiles or by extraction using heuristics is sometimes incomplete or inconsistent); 2) lack of an unified approach to efficiently model different aspects of the academic network [7].

In this direction, Microsoft Academic Search (MAS)[1] is one of the most successful initiatives to archive a huge volume of publication dataset of various domains including computer science, biology, chemistry etc. It is a free public search engine for academic papers and literature, developed by Microsoft Research for the purpose of object-level vertical search. As of February 2014, it has indexed over 39.9 million publications and 19.9 million authors [1].

One fundamental problem with the MAS, however, is the fact that it is only useful as a search engine, i.e., given a certain query, it returns a set of relevant papers. However, this rich source of information is inaccessible to the scientific community for any further research. We embarked on an ambitious initiative to collect the entire computer science dataset in order to facilitate future research on academic social network.

[1] `http://academic.research.microsoft.com/`

Table 1: Percentage of papers in various fields of computer science domain.

Fields	% of papers	Fields	% of papers
AI	12.64	Algorithm	9.89
Networking	9.41	Databases	5.18
Distributed Systems	4.66	Comp. Architecture	6.31
Software Engg.	6.26	Machine Learning	5.00
Scientific Computing	5.73	Bioinformatics	2.02
HCI	2.88	Multimedia	3.27
Graphics	2.20	Computer Vision	2.59
Data Mining	2.47	Programming Language	2.64
Security	2.25	Information Retrieval	1.96
NLP	5.91	World Wide Web	1.34
Education	1.45	Operating Systems	0.90
Embedded Systems	1.98	Simulation	1.04

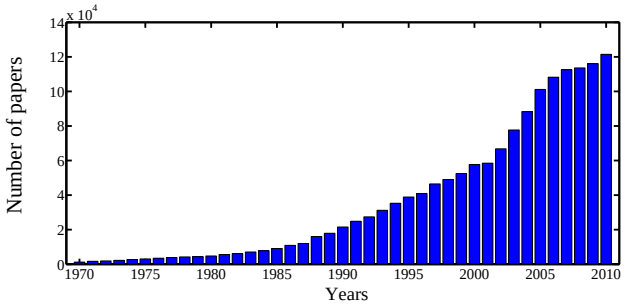

Figure 1: Year-wise growth of the number of publications from 1970 to 2010 that are present in the filtered dataset.

Table 2: General information of the raw and filtered dataset.

	Raw	Filtered
Number of valid entries	2,473,171	1,549,317
Number of entries with no venue	343,090	–
Number of entries with no author	45,551	–
Number of entries with no publication year	191,864	–
Partial data of the years before 1970 and 2011-2012	343,349	–
Number of authors	1,186,412	821,633
Avg. number of papers per author	5.18	5.04
Avg. number of authors per paper	2.49	2.67
Number of unique publication venues	6,143	5,938
Percentage of entries with multiple fields	9.08%	8.68%

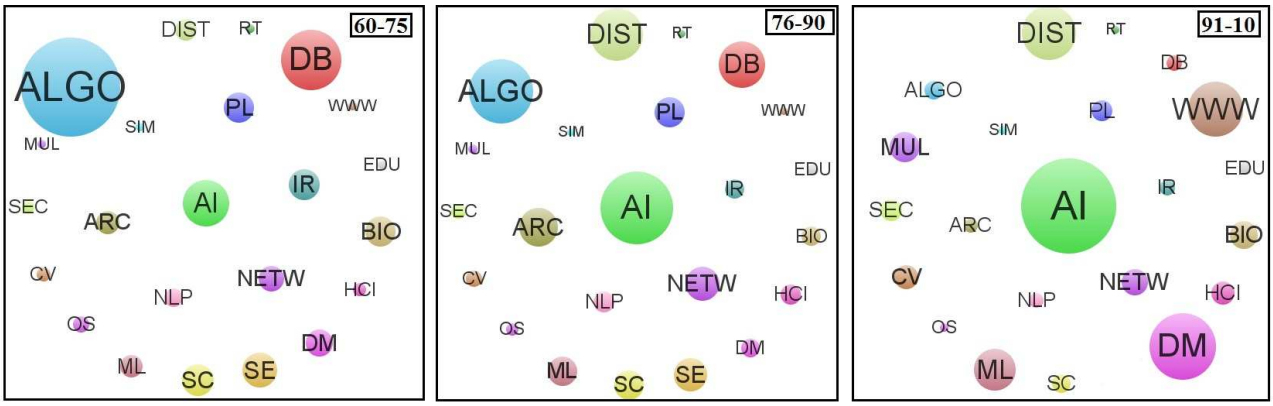

Figure 2: (color online) Number of papers in each field in three successive time periods (1960-1975, 1976-1990 and 1991-2010). The size of each circle denotes the number of papers published in a particular field.

2. CURATION OF THE DATASET

Crawling the papers of computer science domain present in MAS was started on March 2014 and took six weeks to complete. The automated crawler initially used the rank-list given by MAS for each field to obtain the list of unique paper IDs. The paper IDs were then used to fetch the metadata of the publications. We used Tor[2] to distribute our crawling to different systems in order to avoid overloading a particular server with bursty traffic. We employed random exponential back-off time whenever the server or the connection returned some error and sent the request again. We followed the robot restrictions imposed by the servers to ensure efficient crawling of data from both the client and the server perspective. The completely crawled dataset contained all the information related to around 2.5 million papers which are further distributed over 24 fields[3] of computer science domain as shown in Table 1.

3. PREPROCESSING OF THE CURATED DATASET

The crawled data had several inconsistencies that were removed through a series of steps. We first removed few forward citations which point to the papers published after the publication of the source paper. These forward citations appear because there are certain papers that are initially uploaded in public repositories (such as `http://arxiv.org/`) but accepted later in a publication venue. Further, we considered only those papers published in between 1970 and 2010 because this time period seemed to be most consistent since most of the articles published at that time period are available in the dataset. We only consider those papers that cite or are cited by at least one paper (i.e., we removed isolated nodes with zero in-degree or zero out-degree). An advantage of using this dataset is that the problem arising due to the ambiguity of named-entities (authors and publication venues) has been completely resolved by MAS itself, and a unique identity has been associated with each author, paper and publication venue [6]. Some of the authors were found missing in the information of the corresponding papers which were resolved by the DOI (Digital Object Identifier) of the publications. We double checked the filtered papers (around 1.5 million in number) having the author and metadata information from DOI and kept only the consistent ones. Some of the references that pointed to such papers absent in our dataset (i.e., dangling references) were also removed from the dataset. Unless otherwise stated, the filtered dataset is called ***PubCS***, a **Pub**lication dataset of **C**omputer **S**cience domain. Some of the general information pertaining to the raw and filtered dataset is noted in Table 2.

The year-wise growth in the number of overall publications and the field-wise publications from 1970 to 2010 are shown in Figure 1 and Figure 2 respectively. In Figure 2, we observe that in the initial years (1960-1975), the fields like Algorithms and Theory, Databases fully dominated the computer science research; the trend has gradually shifted with the appearance of the fields like Distributed Systems, Networking and Hardware & Architecture in the middle of 80's. In the recent decade, while the number of papers in the fields like Algorithms and Theory, Databases, Operating Systems seem to diminish significantly, the fields like WWW, Data Mining, Multimedia, Computer Vision indicate a larger production of new publications. This result presents a preliminary evidence of the increasing research interest in the integrative areas vis-a-vis a decreasing trend of research in the core fields.

4. METADATA INFORMATION

As mentioned earlier, each paper present in PubCS dataset is associated with a set of additional metadata information. A sample example of an entry is shown in Table 3. For each block, each line starting with a specific prefix indicates an attribute of the paper. In the rest of the section, we present a brief description of these additional attributes present in PubCS.

- **Paper index:** The block starts with a unique identification number of each paper which is marked by the tag "#index".
- **Paper title:** The title of each paper is marked by the

[2]http://torproject.org.in/

[3]Fields are the sub-areas of a research domain. For instance, Algorithms, Databases, Operating Systems are the fields of computer science domain.

Table 3: Sample example of an entry present in PubCS.

#index220
#*Porting the Galaxy System to Mandarin Chinese
#@Chao Wang[53042697]
#IMassachusetts Institute of Technology[568]
#!Galaxy is a human-computer conversational system that provides a spoken language interface for accessing on-line information. It was initially implemented for English in travel-related domains, including air travel, local city navigation, and weather. Efforts were started to develop multilingual systems within the framework of galaxy several years ago. This thesis focuses on developing the Mandarin Chinese version of the galaxy system, including speech recognition, language understanding and language generation components. Large amounts of Mandarin speech data have been collected from native speakers to derive linguistic rules, acoustic models, language models and vocabularies for Chinese. Comparisons between the Chinese and English languages have been made in the context of system implementation. Some issues that are specific for Chinese have been addressed, to make the system core more language independent. Overall, the system produced reasonable responses nearly 70% of the time for spontaneous Mandarin test data collected in a "wizard" mode, a performance that is comparable to that of its English counterpart. This demonstrates the feasibility of the design of galaxy aimed at accommodating multiple languages in a common framework.
#N12
#Y1997
#FNatural Language and Speech
#KLanguage Understanding[22054]
#KMandarin Chinese[24028]
#KSpeech Recognition[39372]
#%WHEELS: a conversational system in the automobile classifieds domain[70384]
#%A probabilistic framework for feature-based speech recognition[69445]
#%Speech Understanding and Dialogue over the telephone: an overview of the ESPRIT SUNDIAL project[1390135]

tag "#*".

- **Name of the author(s):** The list of author(s) is marked by the tag "#@". Multiple authors are separated by comma. Each author is associated with a unique identity. As mentioned earlier, the named-entities of the authors has been completely resolved by MAS itself [6].
- **Affiliation of the author(s):** Each author is also associated with the affiliation or the name of her institute marked by the tag "#I". Multiple authors are separated by comma. Each affiliation is associated with a unique identity.
- **Abstract:** The abstract of each paper is separated by the tag "#!". This attribute might be missing for few entities.
- **Page number:** Total number of pages of each paper is separated by a tag "#N".
- **Year of publication:** The year of publication of each paper is marked by the tag "#Y".
- **Field(s) of publication:** Each paper can belong to one or more number of its related fields denoted by the tag "#F". Note that, in the filtered dataset 8.68% papers belong to multiple fields (act as interdisciplinary papers).
- **Keyword(s) of publication:** MAS assigns keywords, from a global set of keywords, against each paper in order to characterize it properly. Total 39,645 unique keywords are present in our dataset. The keywords are marked by the tag "#K" with multiple lines indicating multiple keywords. Each keyword is also distinguished by a unique index.
- **Reference(s) of publication:** Both the title and the unique index of each paper which the current paper has cited are marked by the tag "#%". For instance, in Table 3, the paper with ID 220 has referred to the papers with indices 70384, 69445 and 1390135.

5. NETWORK CONSTRUCTION

Given the PubCS dataset, one can easily construct different networks such as paper-paper citation network, author-author citation network, author-author coauthorship network, author-paper bipartite network, keyword-keyword network. In this paper, we construct two networks namely, paper-paper citation network and author-author coauthorship network, and analyze their statistics. Since the temporal information is also available in PubCS, one can also associate time information to the construction of these networks.

5.1 Citation Network

A citation network is formally defined as a directed and acyclic graph $G^c =< V^c, E^c >$ where each node $v_i^c \in V^c$ represents a paper and a directed edge e_{ji}^c pointing from v_j^c to v_i^c indicates that the paper corresponding to v_j^c cites the paper corresponding to v_i^c in its references. From our dataset, we have constructed such a network where papers represent nodes and citations represent edges. Note that, at

Table 4: Statistics of the citation and the coauthorship networks.

Statistics	Description	Citation	Coauthor
Nodes	Number of nodes in the network	1549317	786573
Edges	Number of edges in the network	11776995	2838895
Nodes in largest WCC	Number of nodes in the largest weakly connected component	1531415	720647
Edges in largest WCC	Number of edges in the largest weakly connected component	11765128	2770755
Nodes in largest SCC	Number of nodes in the largest strongly connected component	21	720647
Edges in largest SCC	Number of edges in the largest strongly connected component	118	2770755
Clustering coefficient	Average clustering coefficient	0.1433	0.6028
Triangles	Number of triples of connected nodes (considering the network as undirected)	12889777	3823881
Frac. of closed triangles	Ratio of the number of connected triples of nodes and the number of (undirected) length 2 paths	0.010991	0.053984
Diameter	Maximum undirected shortest path length	19	13
90-percentile effective diameter	90-th percentile of undirected shortest path length distribution (sampled over 1,000 random nodes)	7.92173	8.19891

a higher level, one can again construct a field-field network where each field (i.e., a collection of papers) can be thought of as a single vertex and two vertices can again be linked based on an aggregate of citations [2]. We construct an aggregate static network of all the papers present in our filtered dataset. Few statistics pertaining to the constructed network is presented in the third column of Table 4.

The temporal information in our dataset further allows us to show the evolution of the indegree and the outdegree distributions of the filtered citation network in Fig. 3. We observe that indegree (inward citation) distributions follow a power-law behavior. On the other hand, outdegree (outward citation/reference) distributions follow a truncated power-law with a heavy tail at the end.

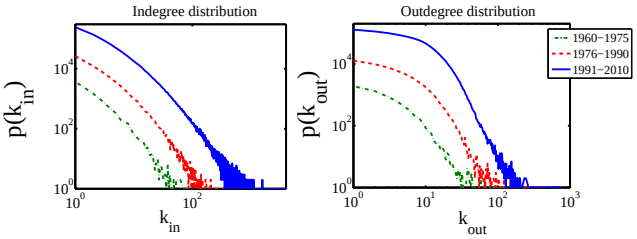

Figure 3: Evolution of (a) indegree (citation) and (b) outdegree (reference) distributions of the filtered citation network. Both axes are in logarithmic scale.

5.2 Coauthorship Network

In the coauthorship network, an edge is created for each collaboration. For example, if a paper is written by Franz Josef Och and Hermann Ney, then an edge is created between the two authors. Formally, a coauthorship network is defined as a graph $G^a =< V^a, E^a >$ where each node $v_i^a \in V^a$ represents a researcher and an undirected edge e_{ij}^a between v_i^a and v_j^a is drawn if the two researchers represented by v_i^a and v_j^a collaborate at least once via publishing a paper. From the above dataset, an overall collaboration network G^a has been constructed with researchers representing nodes and undirected edges representing collaborations between two researchers. Few statistics pertaining to the constructed network is presented in the fourth column of Table 4.

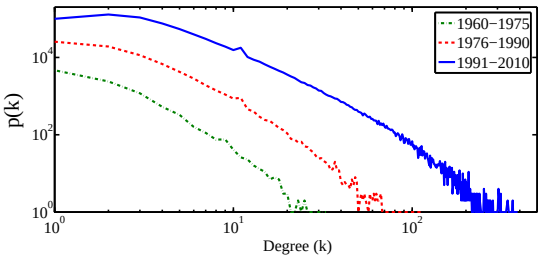

Figure 4: Evolution of degree distribution of the filtered coauthorship network. Both axes are in logarithmic scale.

Figure 4 shows the evolution of degree distribution for the coauthorship network. We observe that for all the time points, the distributions follow mostly a power-low after an initial plateau. We further observe that in the latest time point (1991–2010), a huge majority of authors have degree=2.

6. SAMPLE RANKINGS

This section shows some of the rankings that were computed using PubCS. Note that, it might be possible that the statistics reported in Tables 5, 6 and 7 are not exactly similar to the actual statistics, since we measure all the quantities based on the filtered dataset available in PubCS. However, we believe that the original and the reported statistics should have a high correlation. Table 5 shows top five papers having high in-citations individually for three time stamps: 1960 – 1975, 1976 – 1990, 1991 – 2010. Note that, the number of incitations for a paper is measured within its corresponding time stamp. Similarly, in Table 6 and Table 7 we present the top five authors with high incoming citations and high h-index [3] respectively. We notice that out of top five authors, most of them are common in the lists of incoming citation and h-index [5]. We also show top five keywords that appear in most of the papers for different time stamps in Table 8.

7. CONCLUSION

In this paper, we presented a novel and massive dataset of

Table 5: Sample ranking of papers present in our dataset in three different time periods.

(a) Top five papers with high in-citations published between 1960 – 1975

Sl. no	Name of the paper	# in-citations
1.	A Relational Model for Large Shared Data Banks, *Commu. of The ACM* , 13 (6), 1970.	104
2.	The structure of the "THE"-multiprogramming system, *Commu. of The ACM*, 11 (5), 1968	88
3.	Co-operating sequential processes, *J. of Prog. Lang.*, 1966	73
4.	The Programming Language Pascal, *Acta Informatica*, 1 (1), 1971	62
5.	Programming semantics for multiprogrammed computations, *Commu. of The ACM*, 9 (3),1966	

(a) Top five papers with high in-citations published between 1976 – 1990

Sl. no	Name of the paper	# in-citations
1.	Computers and Intractability: A Guide to the Theory of NP-Completeness, *Artificial Evolution*, 1979	669
2.	The entity-relationship model–toward a unified view of data, *TODS*, 1 (1), 1976	629
3.	Communicating sequential processes, *Commu. of The ACM*, 21 (8), 1978	613
4.	The art of computer programming, *Math. Comput.*, 1979	599
5.	The notions of consistency and predicate locks in a database system, *Commu. of The ACM*, 19 (11), 1976	443

(a) Top five papers with high in-citations published between 1991 – 2010

Sl. no	Name of the paper	# in-citations
1.	Distinctive Image Features from Scale-Invariant Keypoints, *IJCV*, 60 (2), 2004	4210
2.	Chord: A scalable peer-to-peer lookup service for internet applications, *SIGCOMM*, 2001	3474
3.	Mining association rules between sets of items in large databases, *Sigmod Record*, 22 (2), 1993	2971
4.	Fast Algorithms for Mining Association Rules, *VLDB*, 1994	2870
5.	A scalable content-addressable network, *CCR*, 31 (4), 2001	2865

Table 8: Top five keywords that appear in most of the papers in different time stamps.

	Sl. no	Keywords	Count
1960 – 1975	1.	Programming Language	2822
	2.	Operating System	2175
	3.	Time Sharing	1519
	4.	Data Structure	1488
	5.	Computer Program	1268
1976 – 1990	1.	Programming Language	24408
	2.	Data Structure	17159
	3.	Satisfiability	16479
	4.	Distributed System	14681
	5.	Database System	14106
1991 – 2010	1.	Real Time	35782
	2.	Indexing Terms	34102
	3.	Satisfiability	32487
	4.	Neural Network	28907
	5.	Case Study	26133

scientific articles, *PubCS* that apart from the network structures also provides a set of rich additional attributes for each article. The dataset is suitably structured for scientific computation with each attribute distinguished by a certain tag. To the best of our knowledge, this is the largest academic dataset that is publicly available to facilitate scientific research. PubCS is publicly available at `http://cnerg.org`.

As an immediate future work, we plan to crawl the datasets of other domains such as physics, chemistry, biology, mathematics available at Microsoft Academic Search and make them publicly available separately. We also wish to make an aggregated dataset by combining the datasets of all the domains so that the effect of inter-domain interactions can be systematically understood.

8. REFERENCES

[1] `http://datamarket.azure.com/dataset/mrc/microsoftacademic`.

[2] T. Chakraborty, S. Sikdar, V. Tammana, N. Ganguly, and A. Mukherjee. Computer science fields as ground-truth communities: their impact, rise and fall. In *ASONAM*, pages 426–433, 2013.

[3] J. E. Hirsch. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3):741–754, Dec. 2010.

[4] M. Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In *Proceedings of the 9th International Symposium on String Processing and Information Retrieval*, SPIRE 2002, pages 1–10, London, UK, UK, 2002. Springer-Verlag.

[5] L. I. Meho and Y. Rogers. Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of scopus and web of science. *J. Am. Soc. Inf. Sci. Technol.*, 59(11):1711–1726, Sept. 2008.

[6] S. B. Roy, M. De Cock, V. Mandava, S. Savanna, B. Dalessandro, C. Perlich, W. Cukierski, and B. Hamner. The microsoft academic search dataset and kdd cup 2013. In *Proceedings of the 2013 KDD Cup 2013 Workshop*, KDD Cup '13, pages 1:1–1:6, New York, NY, USA, 2013. ACM.

[7] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 990–998, New York, NY, USA, 2008. ACM.

[8] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, pages 3:1–3:8, New York, NY, USA, 2012. ACM.

Table 6: Sample ranking of authors present in our dataset in three different time periods based on total number of in-citations.

(b) Top five authors with high in-citations within 1960 – 1975

Sl. no	Name of the author	# in-citations
1.	Edsger Wybe Dijkstra (Eindhoven University of Technology)	405
2.	Niklaus Emil Wirth (Swiss Federal Institute of Technology Zurich)	311
3.	Donald E. Knuth (Stanford University)	273
4.	Peter Denning (George Mason University)	265
5.	Charles Antony Richard Hoare (Microsoft)	231

(b) Top five authors with high in-citations within 1976 – 1990

Sl. no	Name of the author	# in-citations
1.	Jeffrey D. Ullman (Stanford University)	2193
2.	Philip A Bernstein (Microsoft)	1797
3.	Robert Endre Tarjan (Princeton University)	1758
4.	Michael Stonebraker (Massachusetts Institute of Technology)	1741
5.	Raymond Lorie (IBM Research)	1705

(b) Top five authors with high in-citations within 1991 – 2010

1.	Rakesh Agrawal (Microsoft)	19242
2.	Scott J. Shenker (University of California Berkeley)	18819
3.	Ian T. Foster (Argonne National Laboratory)	17445
4.	Deborah Estrin (University of California Los Angeles)	17003
5.	David E. Culler (University of California Berkeley)	15475

Table 7: Sample ranking of authors present in our dataset in three different time periods based on h-index.

(c) Top five authors with high h-index within 1960 – 1975

Sl. no	Name of the author	# H-index
1.	Niklaus Emil Wirth (Swiss Federal Institute of Technology Zurich)	11
2.	Peter Denning (George Mason University)	10
3.	Edsger Wybe Dijkstra (Eindhoven University of Technology)	10
4.	Zohar Manna (Stanford University)	9
5.	Seymour Ginsburg University of Southern California)	8

(c) Top five authors with high h-index within 1976 – 1990

1.	Jeffrey D. Ullman (Stanford University)	26
2.	Robert Endre Tarjan (Princeton University)	25
3.	Philip A Bernstein (Microsoft)	22
4.	Michael Stonebraker (Massachusetts Institute of Technology)	22
5.	Leslie Lamport (Microsoft)	21

(c) Top five authors with high h-index within 1991 – 2010

1.	Scott J. Shenker (University of California Berkeley)	67
2.	Ian T. Foster (Argonne National Laboratory)	64
3.	Hector Garcia-Molina (Stanford University)	64
4.	Deborah Estrin (University of California Los Angeles)	63
5.	Anil K. Jain (Michigan State University)	62