Scientists Social Network dataset for Computer Science

Adam Matusiak
Institute of Computing Science
Poznan University of Technology
Piotrowo 2, 60-965 Poznan, Poland
Adam.Matusiak@cs.put.poznan.pl

Mikołaj Morzy
Institute of Computing Science
Poznan University of Technology
Piotrowo 2, 60-965 Poznan, Poland
Mikolaj.Morzy@put.poznan.pl

DRAFT

Abstract

Introduction

Social network of scientists is linked by various activities. Being publication co-author or committee co-members is clear indication of link in such a network. Currently there are many separate sources of information on those activities. There are few in public domain that may constitute a base for further analysis. None of them consists of large number of data. Aim of this research is to provide graph of scientists linked by events with specific dates to enable statistical analysis with respect to time of bound creation. Some data may be retrieved from existing open databases, while other must be crawled from WWW pages. Information on publications is used for citation and therefore is stored, catalogued and may be easily searched. Worse is the case of conferences, which pages are often archived or even deleted and no accessible, so link information is lost. Other goal of this research is to preserve this information so in the future long range analysis may be performed. In the future other types of links may be added to enrich dataset and allow for even better analysis.

Data acquisition

World Wide Web is a huge source of documents that follow structural rules for formatting purposes, but unfortunately not for data meta tagging. One page differs from another in structure even if are finally rendered in similar way. HTML tags structure may be rendered in very different way basing on visual styles applied. To shorten data acquisition time and lower software development effort only HTML was analysed and visual effects were omitted.

It's obvious that analysis of already parsed data is much easier. Therefore first number of datasets available to public domain was researched. Two covering biggest amount of data in discipline of computer science were selected – DBLP and WikiCFP.

DBLP dataset was used for scientists' dictionary creation consisting of:

- Name (for some records none or more than one name was supplied),
- Initial (for some records none or more than one initial was supplied),
- Surname (always at least one surname for each record was supplied).

This dataset was used as a dictionary for further data collection. In further process of analysis conference pages were searched for names retrieved from DBLP. It was assumed that people not existing in DBLP database have insufficient contribution in science to be member of conference committee. This assumption is clearly wrong as DBLP is not a complete set of data and manual data check seemed to confirm that some links have been missed. On the other hand attempt to supplement this name dictionary may lead to creation of many false entries due to web sites structure irregularity.

Publications stored in DBLP database were used to create links between scientists. Each author and editor is linked to all publications found. Publications are additionally grouped by books or journals containing given publication where applicable. For each publication issue date was stored as link creation date. When both year and month were supplied, first day of given month was assumed to be the issue day. When only year was present, first day of year was stored as link date.

Secondly WikiCFP has been scanned for conferences URLs. Only conference URL and name were retrieved to keep the process uniform for all the conferences and allow to use different or other sources in the future.

Crawler uses multi-level text analysis using some of the Web Data Extraction Techniques identified by Laender et al. (2002):

- HTML-Aware – analysis of tag structure of website's source code and operation on the tag tree,
- NLP-Based – filtering, part-of-speech and lexical semantic tagging is used to build relationships between phrases and sequence elements,
- Wrapper Induction – generation of delimiter-based extraction rules,
- Modelling-Based – location of the website parts conforming with a given structure,
- Ontology-Based – using an ontology to locate constants - relying directly on the data.

HTML–Aware analysis is used to identify links that may be followed to scan other pages and site section boundaries. Wrapper Induction is helpful to identify committee and its members. Other techniques are mixed to e.g. qualify links that are worthy following.

First for each identified conference past and future events are searched. This is done in two step process. First if URL contains year algorithm replaces it with other values and tries to retrieve pages from newly formed URLs. Later all pages from initial scan and found in step one are searched for links to other conferences. Crawler follows page links to browse subpages and find possible links to other conferences. All those links are used as root conference links in next steps.

For each conference committees and theirs members are identified as well as conference date. Members are mapped to entries of scientists dictionary created on base of DBLP in process described below. Committees are used to link them and conferences constitutes Link Groups.

Each text string that was possibly a scientist name *n* was passed through matching procedure that assigned similarity s_k to each name form dictionary m_k:

$$s_k = \sum_i k_i f_i(m_k, n)$$

where:

$f_i(x,y)$ – matching functions described below,

k_i – empirical weighting coefficients for above functions retrieved in manual process on data sample subsets.

Function	**Condition**
	Condition is run against all words in parameters, only best match is used for each pair x_i word from parameter *x* y_i word from parameter *y*
Equality matcher	x_i is equal to y_i
Substring matcher	x_i starts with y_i or y_i starts with x_i
Initial matcher	x_i is initial (single letter followed by a dot) and y_i starts with letter from x_i or y_i is initial (single letter followed by a dot) and x_i starts with letter from y_i
Similarity matcher	x_i and y_i has no more than one different character

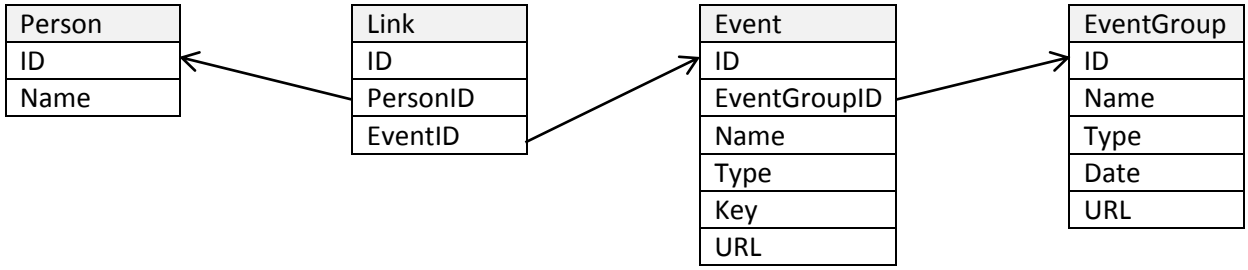
For each input string best match with highest similarity is found:

$$s = \max(s_k)$$

When similarity is higher than value if higher than s_{min} threshold it is accepted as match. Otherwise input string is not recognized as person name. Threshold value was retrieved in same manual process on same data sample subsets that k_i coefficients.

The dataset

Data is stored in four tables:

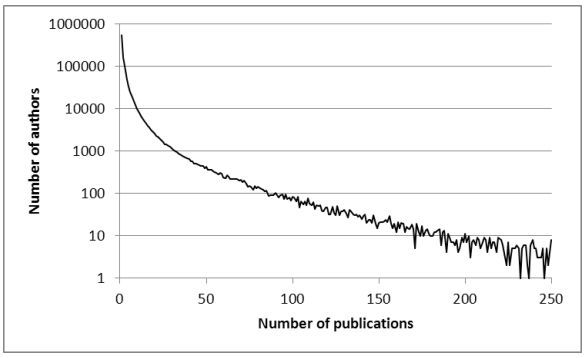


- Person – list of all indexed scientists,
- Link – many-to-many connection between scientists and events,
- Event – publication or committee (which may be determined by Type column),
    - Key column may be used to refer back to DBLP,
    - URL may be used to get conference source,
- EventGroup is:
    - book or journal, when grouped events are publications,
    - conference, when grouped events are committees.

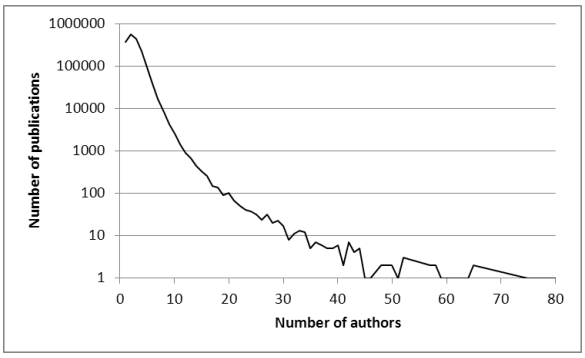
In the process of data acquisition dataset has been collected:

Entity	**Quantity**
Scientist	1025692
Publication – author link	4722415
Committee – member link	410044
Event (total)	1773983
Publication	1759452
Book / journal	32450
Committee	14531
Conference	6146

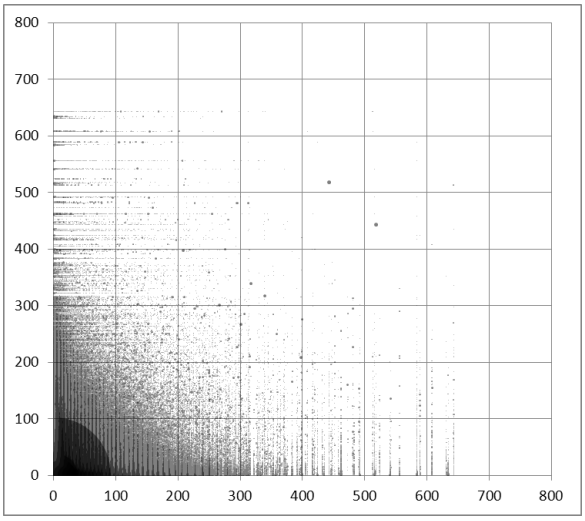
Extended statistics of dataset are presented on pictures below. Please note that part of the long tail have been truncated for chart essential part clarity.



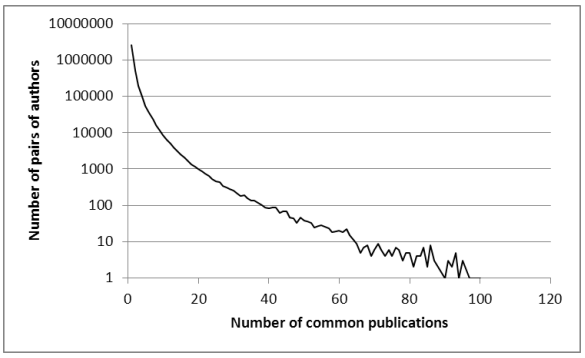
Pic 1 Distribution of number of authors per publication



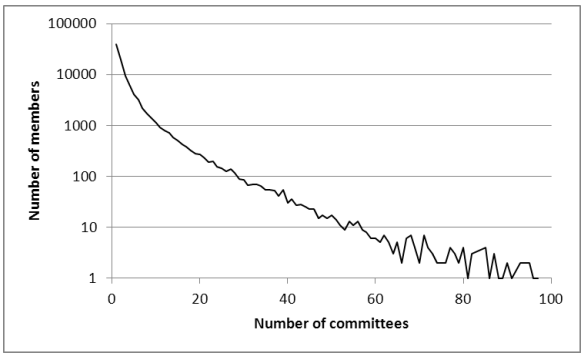
Pic 2 Distribution of number of publications per author



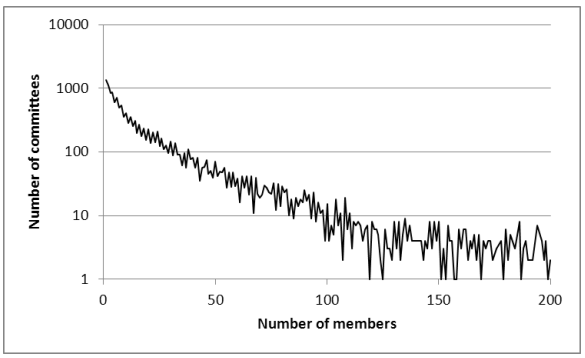
Pic 3 Distribution of probability that two auhors of given number of publications are co-authors



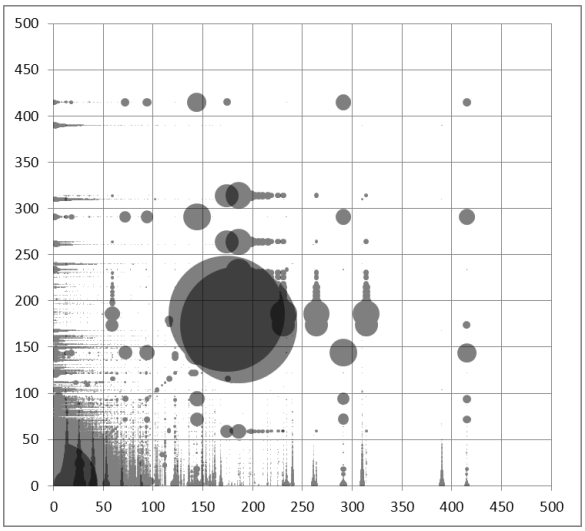
Pic 4 Distribution of number of common publications per co-author pair



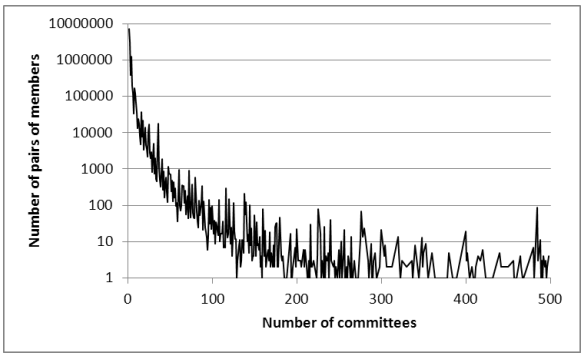
Pic 5 Distribution of number of members per committee



Pic 6 Distribution of number of committees per member



Pic 7 Distribution of probability that two members of given number of committees are members of one committee



Pic 8 Distribution of number of common committees per co-member pair

Conclusion

Collected dataset may be useful base for social network analysis. May on one hand be used for validation of social network analysis algorithms that were developed for other types of networks – i.e. Facebook or Twitter. May help to create new approaches and analysis techniques. On the other hand main goal is to allow the computer science society analysis. Multiple techniques may be used to analyse how this network has grown to point where it's now.

This dataset is far from being complete and will evolve in time by adding new information or updating existing. This network is growing and changing as growing and changing is the community that is constituting it by its scientific effort. In the same time it covers great part of major events so may be treated as a close representation of the whole network.

Dataset access

The dataset is maintained under the IQSS Dataverse Network at Harvard University, and can be located at the following URL: http://dvn.iq.harvard.edu/dvn/dv/SSN . All researchers are required to sign a user agreement as well as to submit a brief statement of their intended research. Additional information on the Dataverse Network Project can be found in King (2007) and at the project's homepage: http://thedata.org/.

References

King, G., 2007. An introduction to the Dataverse Network as an infrastructure for data sharing. Sociological Methods and Research 36, 173–199.

DBLP, Computer Science Bibliography. http://www.informatik.uni-trier.de/~ley/

WikiCfP, http://www.wikicfp.com/, Last recall 26.10.2011.

A. H. Laender, B. A. Riberio-Neto, A. S. da Silva, and J. S. Texeira, 2002. A brief survey of web data extraction tools.