

Project Report: Author Identifiers

Fakhri Momeni, Philipp Mayr (GESIS)

October 16, 2015

This project report describes the work undertaken in the project “Author Identifiers for WoS” at GESIS (from May to October 2015). The report describes an approach to build a gold standard for author name disambiguation on the basis of the DBLP corpus and a simple algorithm basing on co-authorship analysis to disambiguate author names. In a second step (October 2015), we have worked on the adaptation of our disambiguation approach for WoS data. Fakhri Momeni has worked on the project starting in May 2015, she worked in part time.

1 Gold standard

In this project we created a gold dataset for author name disambiguation on the basis of DBLP¹. For this goal we searched for homonym authors in DBLP. In DBLP homonym authors are distinguished with an id after their name. We used this type of names to build our gold dataset.

Totally 1.578.316 unique names exist in DBLP. There are 5.408 authors, who have a number at the end of their name and altogether 1.932 homonym names (we mention them as disambiguated authors). We got these numbers from xml-file of DBLP, downloaded 01.05.2015 from <http://dblp.uni-trier.de/xml/>.

We converted this xml-file to a mysql table. Here are some tables with data from xml-file:

- ‘dblparticle’: list of all documents in DBLP (entirely 2.912.888)
- ‘author’: list of all author’s names in DBLP
- ‘disambauthor’: list of all disambiguated author’s names in DBLP
- ‘rand_author’: list of 1.000 randomly selected homonym names from ‘disambauthor’
- ‘rand_publications’: list of documents for 1.000 randomly selected homonym names (32.273 documents in total)

The following tables are exported as csv-files are appended to this report:

- goldstandard.csv: All documents of selected gold standard. It is contain these columns:
 - docId: the given Id for each document
 - key: the key of the document obtained from xml file.
 - author: the list of all authors of the document
 - title: the document’ title.
- goldstandard_ambiguous.csv: same as goldstandard.csv, just identity number of homonym names is removed.
- cluster.csv: contains all documents (32.273 in total). Below is a description of its fields:
 - docId: the given Id for each document
 - authored: the given Id for the author
 - block: the number of block for each selected author name

¹ <http://dblp.uni-trier.de/>

- person: the author's name with the identity number
- name: the author's name without the identity number
- cluster2: the cluster's number detected by our method with threshold 2
- precision2: the BCubed precision calculated with threshold 2
- recall2: the BCubed recall calculated with threshold 2
- recall2: the BCubed recall calculated with threshold 2
- fmeasure2: the BCubed F calculated with threshold 2
- cluster4: the cluster's number detected by our method with threshold 4
- precision4: the BCubed precision calculated with threshold 4
- recall4: the BCubed recall calculated with threshold 4
- fmeasure4: the BCubed F calculated with threshold 4
- author_#documents.csv: a list of all random authors (rand_author) and all their homonym names
 - person: disambiguated author name
 - documentsNumber: frequency of paper in DBLP
 - block: number of block
- problematicNames.csv: the list of 50 homonym names that have most documents in dblp and contains:
 - name: the author's name
 - document_count: number of document belongs to the name.

2 The GESIS disambiguation approach in DBLP

Considering the lack of author's information for stored documents in DBLP, we used a co-author communities approach. In [1] [2] state the co-author analysis is a good solution to identify homonymous authors automatically.

Therefore we built a social network of authors and documents of DBLP, which authors would be connected together through documents in different depths [2].

1.000 homonym names are selected randomly and publications of these researchers were retrieved using a clustering algorithm [3] to detect publications per researcher.

For this purpose we built 1.000 blocks and each block contains documents with the homonym name that are categorized by different researchers.

In the next step the homonym names were removed from the list of authors for all documents in each block and the social network of entire DBLP was built with this status.

The structure of the social network in neo4j

The network contains two node-sets and a relationship.

Author: node-set contains all authors in DBLP

Doc: node-set contains all documents in DBLP

WRITTEN_BY: to identify who wrote each document (with removing 1,000 selected author's names from the list of documents in mentioned block).

So, the relationship between a pair of documents is defined based on connections between authors and documents with the type of 'WRITTEN_BY' relation. For example, in block 1 we have two documents with id 131.367 and 135.587.

Both of them are written by 'Abbas Mohammadi 0002'. The document with id 131.367 is written by 'Abdolali Abdipour' with id '19.039' too. Also, the document with id 135.587 is written as well by 'Mahdi Majidi' with id 864.084. An article with id 135.357 belongs to 'Abdolali Abdipour' and Mahdi Majidi'. In this way two documents are connected to each other. Their connection is defined as the shortest path between them (without considering 'Abbas Mohammadi 0002' as one of the authors). Figure 1 shows the shortest path between these two documents (with length 4).

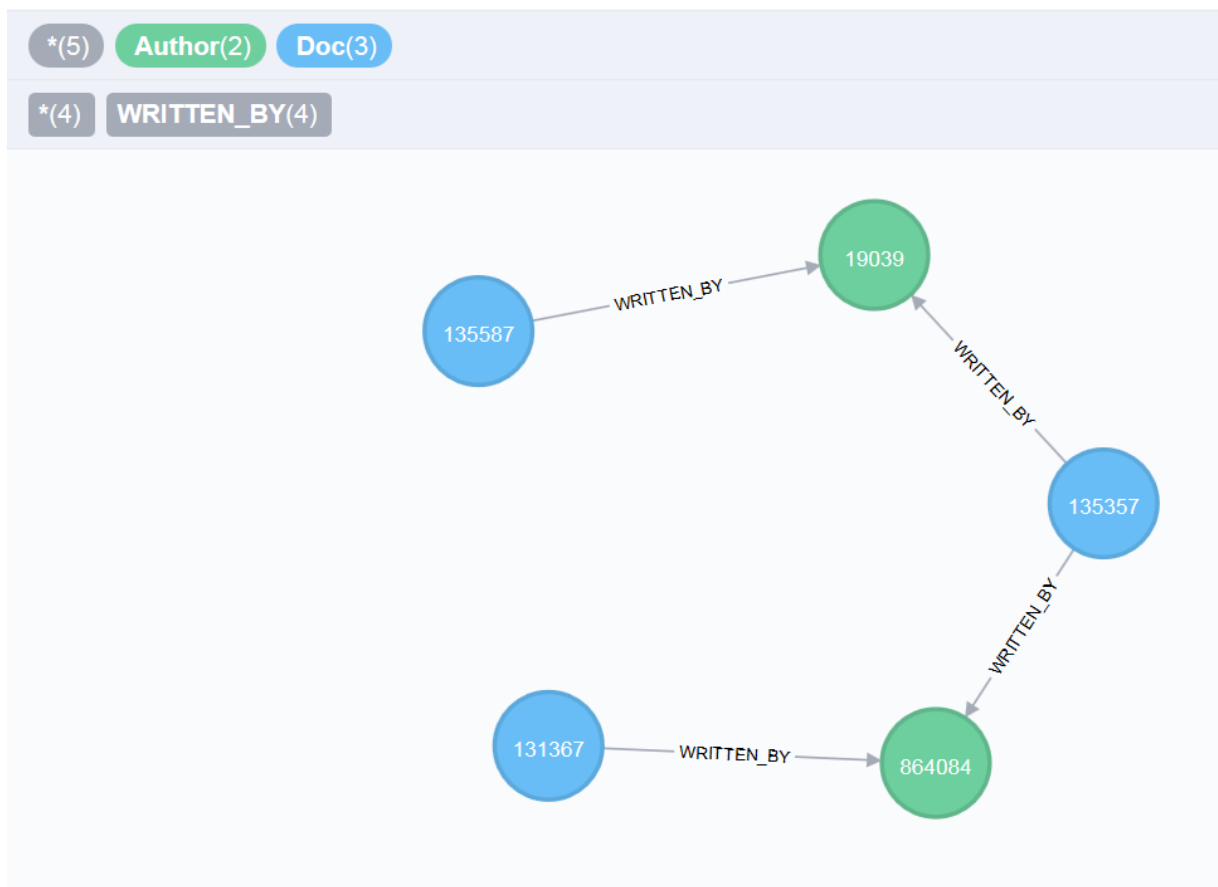


Figure 1 Shortest path between two documents with ids 131367 and 135587

Then, with finding a shortest path between any pair of documents in each block, we estimated that they are written by the same author or just have homonym authors. The length of the shortest path is considered as the threshold for connectivity between two documents. We tested it with two thresholds (length=2 and length=4).

At the end, for every block we have a graph with documents as nodes and connection between them. Any connected subgraph is considered as a cluster which contains a list of documents for a researcher and the number of these cluster are the number of researchers in this block.

Evaluation

The results of the clustering are evaluated with metrics as BCubed precision, BCubed recall and BCubed F described in [4]. BCubed precision of an item is the proportion of items in its cluster which have the item's category (including itself). The overall BCubed precision is the averaged precision of all items in the distribution. The BCubed recall is analogous, replacing "cluster" with "category" in definition of BCubed precision. Figure 2 illustrates how the precision and recall of one item is computed by BCubed metrics. Assume that circles (items) with same colors are the documents belong to a person and the clustered items are the documents that our method detects as the publications of a certain person.

BCubed F is computed as follows:

$$F(R, P) = \frac{1}{\alpha \left(\frac{1}{P}\right) + (1 - \alpha) \left(\frac{1}{R}\right)}$$

being R and P two evaluation metrics and being α and $(1 - \alpha)$ the relative weight of each metric ($\alpha = 0.5$ leads to the harmonic average of P, R).

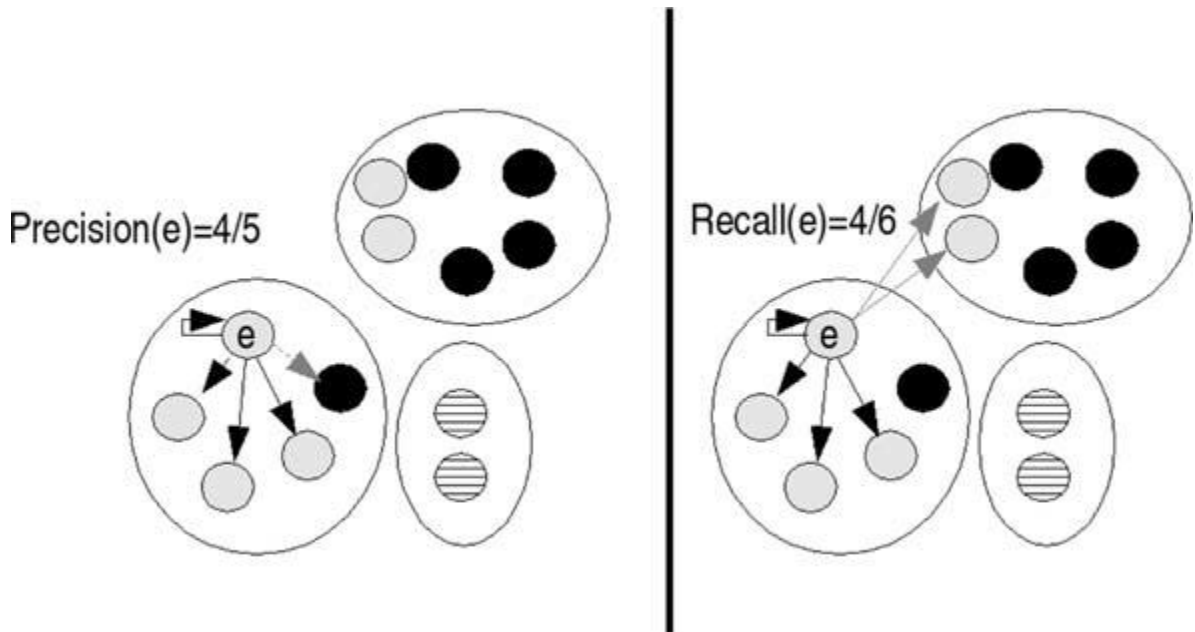


Figure 2 Example of computing the BCubed precision and recall for one item

We calculated the BCubed precision, BCubed recall and BCubed F for each document. Then their average for each block is obtained as block's BCubed precision, recall and F.

Table 1 shows the mean values of evaluation metrics for 1.000 selected names against the gold standard.

Table 1 The mean values of BCubed metrics for 1,000 blocks measured against the gold standard.

	BCubed precision	BCubed recall	BCubed F
Threshold=4	0.946	0.8133	0.82
Threshold= 2	0.988	0.742	0.79

Figure 3 and Figure 4 show the relationship between the number of documents of each author name and the obtained BCubed F through the linear regression analysis. We can see, with increasing the number of documents in the blocks our algorithm is less able to disambiguate the documents. It would be worse for threshold=4. Also, Figure 5 indicates that the top one percent author names in our selection are Chinese (or west Asia). It means that Chinese names are most problematic author names and need to be more verified.

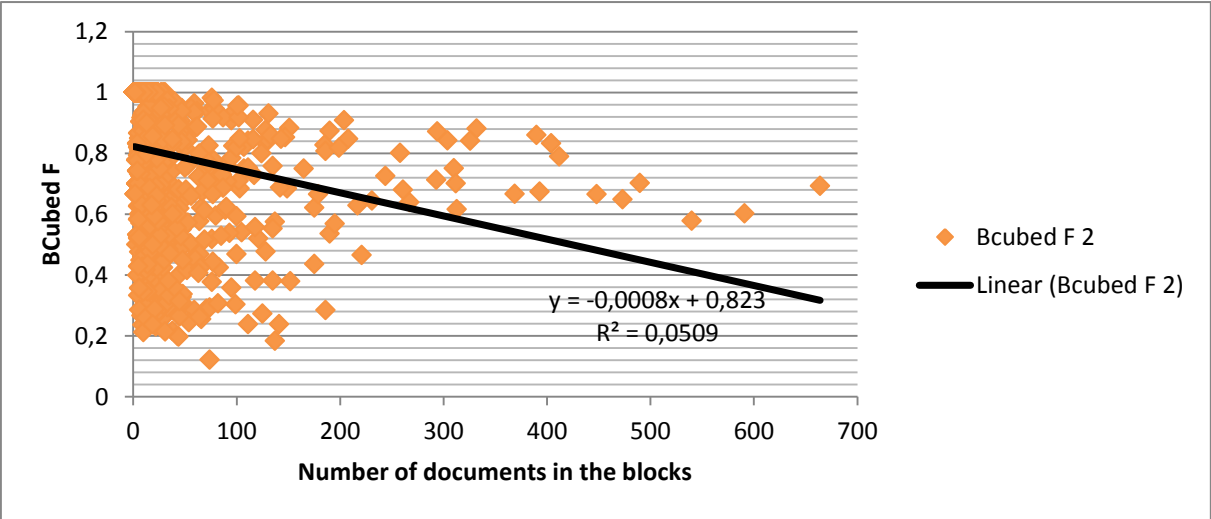


Figure 3 The correlation between BCubed F and the number of documents in the blocks based on linear regression Analysis with Threshold 2

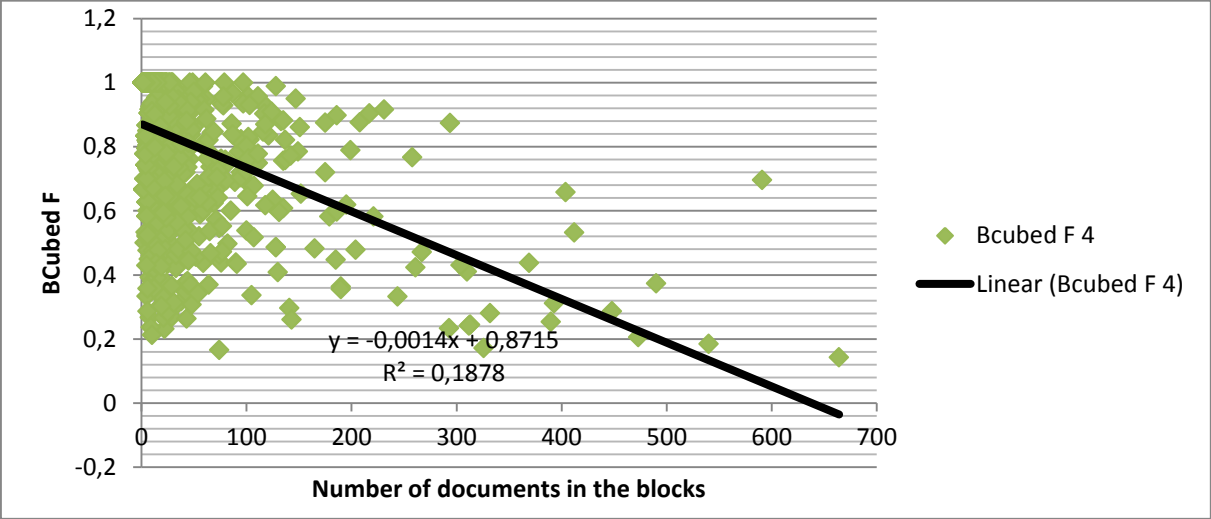


Figure 4 The correlation between BCubed F and the number of documents in the blocks based on linear regression Analysis with Threshold 4

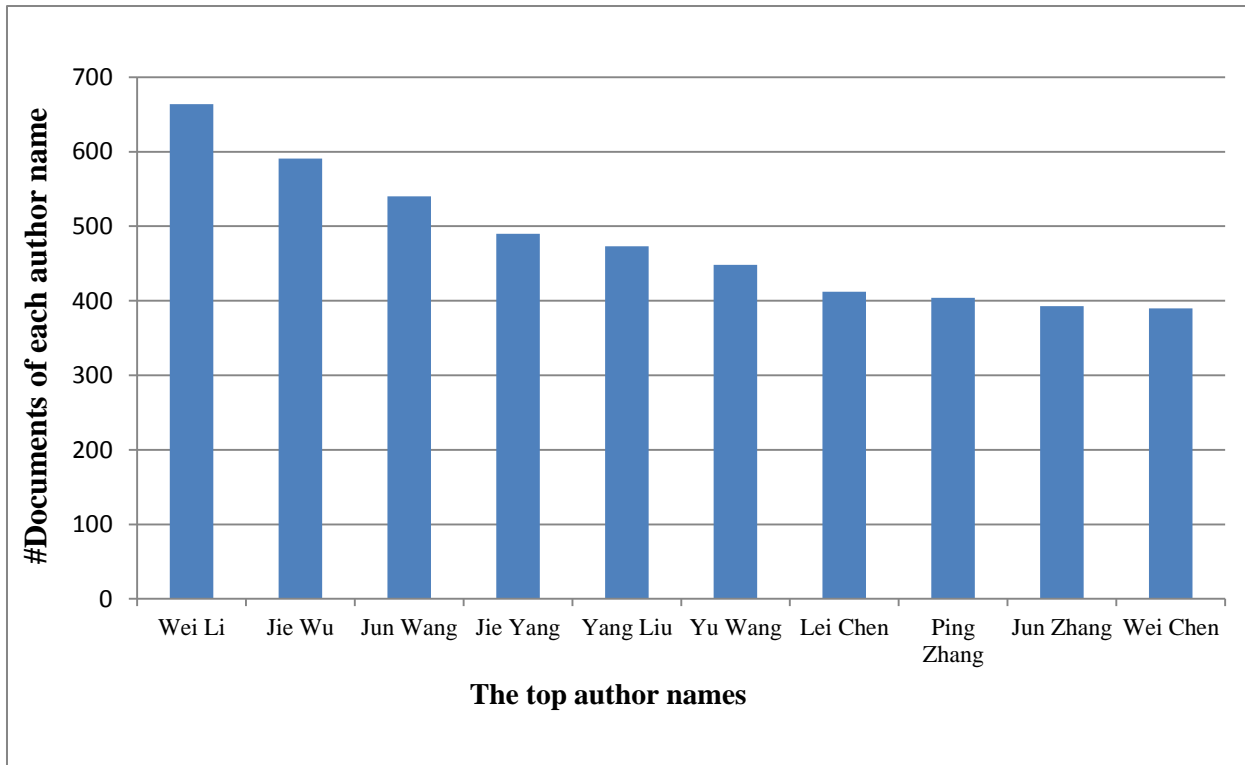


Figure 5 The top 1% author names in 1.000 random selected names

We used the shortest path to detect the similarity between documents. For the future work we can compute all paths between two documents and count the number of paths as a weight for their connection. Also we can compare titles to get the better result. For this purpose, keywords could be extracted from titles to discover a similarity between each pair documents and estimate their likely relation.

Using HCS (Highly Connected Subgraphs) Clustering Algorithm would be a solution to optimize our clustering algorithm, if the size of blocks were too large.

3 Adaptation on WoS data

In the previous section we defined a gold standard for homonym names in DBLP and tried to distinguish them via co-author's community.

The situation is different for WOS (Web of Science), because we have more information about authors and documents. So, these bibliographic elements are applied to identify the authors.

The WOS's gold standard

We know that some authors in WOS have researcher Ids and used these IDs as our gold standard. We choose the authors who have different Ids and the same name (same first name, same last name and same standard name) from WOS_R (unprocessed database). There are 7.084 author names with these conditions. **1000 random names, which have more than five documents, are chosen.** The amount of documents for these authors is 11.867.

The disambiguation approach

The same clustering method used in DBLP is used for WOS. Homonym names are categorized into 1000 blocks. We utilized two databases of 'KB-Datenbanken' to get the bibliographic information:

- WOS_R: the unprocessed database
- WOS12B: the bibliometric database

First of all we extract the factors that would be suitable as the similarity measures between documents. Here are the bibliographic elements applied as the similarity measures for each pair documents and their source tables:

- Email : WOS_R. PERSONS_EMAILADDR
- Address of author: WOS12B. ITEMS_AUTHORS_INSTITUTIONS
- Grant number: WOS12B. GRANTNUMBERS
- The source (**publisher**) of items: WOS12B.ITEMS_AUTHORS_INSTITUTIONS
- Subjects of source (**publisher**): WOS12B. SOURCES_CLASSIFICATIONS
- Subjects of items: WOS12B. ITEMS_CLASSIFICATIONS
- Keywords: WOS12B. ITEMS_KEYWORDS
- Self-citation: WOS12B. CITINGITEMS
- Bibliographic coupling: WOS12B. CITINGITEMS
- Co-citation: WOS12B. CITINGITEMS

In the next step, we tried to find pairs of documents within blocks that are probable written by the same author based on a set of scoring similarity measure. So, we gave for each similarity measure a score (inspired by the work of Caron, E., Van Eck and N. J. 2014). This means different shared elements have different impact on detecting the probability of same authors. For example two publications with the same email address are highly probable to be written by the same author. So, email has the highest score comparing to other factors. Table 2 shows the score assigned to each element.

Table 2 elements, scores and threshold for similarity measures

category	element	Criterion	Score
Author	email		100
	address		7
article	Shared co-authors	one	4
		two	7
		More than two	10
	grant number		10
keyword	one	2	
	two	4	
	more than two	8	
classification		3	
source	classification		3
	name of source		6
citation	self-citation		10
	bibliographic coupling	one	2
		two	4
		three	6
		four	8
		more than four	10
co-citation	one	2	
	two	3	
	three	4	
	four	5	
	more than four	6	
	Threshold		10

The sums of scores with the threshold ≥ 10 were considered as the meter of connectivity between the pairs. Like previous section with having connection between each two documents we built a set of graphs and each connected graphs are defied as oeuvres of an author.

Evaluation

The same as the previous part we evaluated the output with BCubed metrics. Table shows the evaluation result.

Table 3 The mean values of BCubed metrics for 1,000 blocks measured against the gold standard.

	BCubed precision	BCubed recall	BCubed F
Threshold= 10	0.96	0.86	0.879

With comparing the result of the evaluation in Table to Table 1 from DBLP we can conclude that the more number of shared bibliographic elements used as the similarity measures for the pair documents, the higher the likelihood of detecting the publications written by the same authors. On the other hand, just using co-author networks for the author disambiguation and getting a good result for DBLP reveals that co-author networks have an important role to author identifier.

4 Outlook

In the next step we are planning to implement our disambiguation algorithm for a larger set of the WOS database. In consideration of more information of documents such as the author's affiliation and address, citation and the document's details, we should regard them as similarity measures for the comparisons.

In a follow-up project (see separate proposal) we plan to make our approach scalable for larger data sets.

5 References

1. D. Shin, T. Kim, H. Jung, J. Choi.: Automatic Method for Author Name Disambiguation using Social Networks. In AINA, pp. 1263-1270. IEEE Computer Society, Los Alamitos (2010)
2. Levin F. H., Heuser C. A.: Evaluating the use of social networks in author name disambiguation in digital libraries. *Journal of Information and Data Management*, 1(2):183-197 (2010)
3. Caron, E., Van Eck, N. J.: Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), *19th International Conference on Science and Technology Indicators*. Leiden: CWTS-Leiden University (2014)
4. E Amigó, J Gonzalo, J Artiles, F Verdejo.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval* 12,461-486 (2009).

6 Appendix 1 for DBLP

- goldstandard.csv
- goldstandard_ambiguous.csv
- problematicNames.csv
- cluster.csv
- author_#documents.csv

7 Appendix 2 for WoS

The following tables are exported as csv-files are attached to this report:

- Goldstandard.csv: All documents of selected 1000 names (11.867 in total). It contains these columns:
 - FIRSTNAME: first name or initial
 - LASTNAME: last name
 - STANDARDNAME: the first name and the last name separated by ‘,’
 - R_ID: researcher id of the author
 - BLOCK: the block number of each group of homonym names
 - WOS12B_AUTHORID: intern Id of the author in the table WOS12B.AUTHORS
 - WOSR_AUTHORID: intern Id of the author in the table WOS12B.PERSONS
 - WOS_UID: unique id of the document in WOS
 - WOS12B_ITEMID: intern id of the document in the table WOS12B.ITEMS
- Cluster.csv: contains the documents and their cluster that our disambiguation algorithm has detected. The evaluation results are in this file too. Below is a description of its fields:
 - FIRSTNAME: first name or initial
 - LASTNAME: last name
 - STANDARDNAME: the first name and the last name separated by ‘,’
 - R_ID: researcher id of the author
 - BLOCK: the block number of each group of homonym names
 - WOS12B_AUTHORID: intern Id of the author in the table WOS12B.AUTHORS
 - WOSR_AUTHORID: intern Id of the author in the table WOS12B.PERSONS
 - WOS_UID: unique id of the document in WOS
 - WOS12B_ITEMID: intern id of the document in the table WOS12B.ITEMS
 - CLUSTER: the cluster’s number detected by the algorithm
 - PRECISION: the BCubed precision
 - RECALL: the BCubed recall
 - FMEASURE: the BCubed F
- Combination.csv: contains the pairs of documents of homonym names inside the blocks that should be compared:
 - PK_COMBINATION: intern Id of the pair documents.
 - WOS_UID1: unique id of the first document in WOS
 - WOSR_ITEMID1: intern id of the first document in WOS_R.ITEMS
 - WOS12B_ITEMID1: intern id of the first document in WOS12B.ITEMS
 - WOSR_AUTHORID1: intern Id of the first author in the table WOS12B.PERSONS

- WOS12B_AUTHORID1: intern Id of the first author in the table WOS12B.AUTHORS
- R_ID1: researcher id of the first document's author
- WOS_UID2: unique id of the second document in WOS
- WOSR_ITEMID2: intern id of the second document in WOS_R.ITEMS
- WOS12B_ITEMID2: intern id of the second document in WOS12B.ITEMS
- WOSR_AUTHORID2: intern Id of the second author in the table WOS12B.PERSONS
- WOS12B_AUTHORID2: intern Id of the second author in the table WOS12B.AUTHORS
- R_ID2: researcher id of the second document's author
- similarity_measure_score.csv: In this file there are the similarity score given to each pair and sum of them:
 - FK_COMBINATION: intern Id of the pair documents.
 - BLOCK
 - R_ID1
 - R_ID2
 - EMAIL
 - CO_AUTHOR
 - INSTITUTION
 - GRANTMUNBER
 - CLASSIFICATION
 - KEYWORD
 - SOURCE
 - SOURCE_CLASSIFICATION
 - BIB_COUPLING
 - CO_CITATION
 - SELF_CITATION
 - SUM: sum of above scores
- author_#documents.csv: a list of all selected authors and the number of their documents:
 - FIRSTNAME: first name or initial
 - LASTNAME: last name
 - STANDARDNAME: the first name and the last name separated by ‘ ’
 - R_ID: researcher id of the author
 - DOCUMENT_COUNT: the number of documents belong to the author
 - block: number of block
- problematicNames.csv: the list of 50 homonym names that have most documents in dblp and contains:
 - FIRSTNAME: first name or initial
 - LASTNAME: last name
 - STANDARDNAME: the first name and the last name separated by ‘ ’
 - DOCUMENT_COUNT: number of document belongs to the name.