

# Project Report: Author Identifiers

## Part II

---

**Fakhri Momeni, Philipp Mayr (GESIS)**

October 31, 2016

The goal of this second part of the Author Name Disambiguation (AND) project was to disambiguate a large number of homonym names in ‘Web of Science’. In the other words we want to classify the publications written by the *same name* to get the different persons with their publications.

### 1. Overview of the disambiguation approach

We use the author grouping method [1] in order to assign all publications of each person to a certain group. For this purpose all publications belonging to the same author name are categorized into one block. In a next step we compare any pair of publications in each block with each other to find a similarity between them. If we have  $n$  blocks and  $m_i$  publications in the block  $i$ , the number of comparisons for all blocks is:

$$\sum_{i=1}^n \frac{m_i(m_i - 1)}{2}$$

In WoS  $n = 20,255,538$  and the number of comparisons = 19,426,538,865 .

The result of each comparison is true or false. The true result means that two publications belong to one person and the same cluster. If one of them was compared with some other publications before and assigned to a cluster, the other one is added to that cluster too. If both of them were compared before and belong to different clusters, two clusters are rebuilt to one cluster. Otherwise a new cluster will be created and two publications are put in new cluster. We use similarity measures listed here to compare the pair which exist in the bibliometric database ‘WOS12B’ and also email address from xml files:

- Email : XML file
- Address of author: WOS12B. ITEMS\_AUTHORS\_INSTITUTIONS
- Co- authors: WOS12B. ITEMS\_AUTHORS\_INSTITUTIONS
- Grant number: WOS12B. GRANTNUMBERS

- The source (publisher) of items: WOS12B.ITEMS
- 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

## 2. Data selection

In this project part we selected the names with the number of publications between 100 and 200 to disambiguate. The number of names is 140,049 and totally they have 19,033,499 publications.

## 3. Status of project

Because we should compare a large set of items, we should optimize our query to get a result in a reasonable time. We checked some solutions and now we have a stored procedure that we can run for selected publications. In the appendix you can find the code of stored procedure written in oracle.

## 4. Output of the project

At the end of project we will have a table consists of these items:

**ID\_AUTHOR:** The Id of name (came from WOS12B.AUTHORS)

**ID\_ITEM:** the id of publication for the ID\_AUTHOR (came from WOS12B.ITEMS)

**CLASS:** the number in this field is the group number name indicated by ID\_AUTHOR with the Publications indicated by ID\_ITEM. In fact all publications with the same name and the same class belong to one person.

## 5. References

1. F. Momeni, P. Mayr, Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names, in: 20th International Conference on Theory and Practice of Digital Libraries (TPDL 2016), 2016, pp. 386-391 doi: 10.1007/978-3-319-43997-6\_31.

## 6. Appendix

COMBINATION\_COMPARISON.sql