Improving Data Integrity of Individual-based Bibliographic Repository Using Clustering Techniques

Firdaus[1], Oky Budiyarti[2], Muhammad Anshori[1], Mira Afrina[2], Siti Nurmaini[1]

[1]Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya
[2]Department of Information System, Faculty of Computer Science, Universitas Sriwijaya
virdauz@gmail.com

ABSTRACT

This paper presents a method to improve data integrity of individual-based bibliographic repository. Integrity improvement is done by comparing individual-based publication raw data with individual-based clustered publication data. Hierarchical Agglomerative Clustering is used to cluster the publication data with similar author names. Clustering is done by two steps of clustering. The first clustering is based on the co-author relationship and the second is by title similarity and year difference. The two-step hierarchical clustering technique for name disambiguation has been applied to Universitas Sriwijaya Publication Data Center with good accuracy.

Keywords*:* Author Name Disambiguation; Bibliographic Repository; Hierarchical Agglomerative Clustering.

1. INTRODUCTION

Universitas Sriwijaya Publication Data Center (USPDC) is a bibliographic repository that objectives to collect publication metadata from various sources stored into a database. Publication metadata obtained from various online sources is directly associated with a particular author. It can be used for individual based bibliometric analysis. Bibliometric analysis plays a very important role in making a successful analysis of a research [1].

To associate the document with the author, the most important thing to do is to identify who the author is on a document. Identification of documents by author is still difficult to do. Two main challenges in recognizing the author in a document are [2][3]; 1) In some cases, the author writes a name in a different form. This difference can be attributed to synonyms, typos, name changes due to marriage, religion conversion and others. 2) Some authors have the same name.

Document association to the author on the Publication Data Portal is done directly when the document has been identified who the author is. Documents obtained by Author ID are associated to the author. While documents obtained from Affiliation ID is done manually based on the knowledge of system operators. Therefore, the identification of documents to the author becomes very important to do to get accurate and reliable results.

Author name disambiguation (AND) is an activity performed to associate a document to the appropriate author [4][5]. The main problem with the name ambiguity is that the same author uses a different name (synonym), or a different author has the same name (polysemy)[2][6].Succeeds AND will greatly assist the author in finding his academic information from the right source [7]. There are many AND techniques that have been proposed before. One of them is clustering technique [6][7][8][9][10].

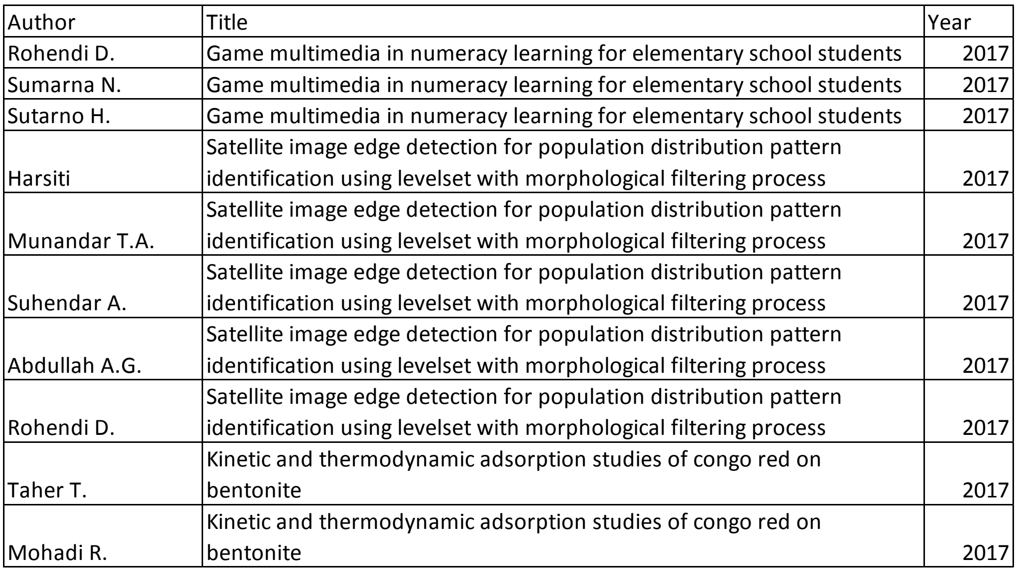
Various approaches have been widely applied to solving the name disambiguity problem. The most widely developed approach is hierarchical method using Hierarchical Agglomerative Clustering (HAC) technique. This technique is widely used in the settlement of name ambiguity because it has a better result than using other clustering method. The HAC technique has a better average percentage of results when compared to using the K-means algorithm for the author's name and article title attributes. In the previous research, using a modified HAC algorithm by applying a confidence ranking, obtained an average of 95.25% percent for precision, 84.11 for recall, and 88.98% for F1 Score [11].

This paper presents the implementation of polysemy AND technique into documents search application. This application is used to search for documents that are not included in author ID based searches and search for document that shouldn’t belong to one author ID based searches results. The HAC approach for grouping publication data with two ways of clustering is used to solve the problem.

2. METHODOLOGY

2.1 DATA SET

The publication data used in this research was obtained from Scopus by using Scopus Application Programming Interface (API). Scopus API provides document metadata based on first name and last name and document detail metadata based on document ID (eid). The first request, based on first name and last name provided by USPDC, generates document metadata with title, years, and eid attributes of the authors with the same name. With the eid obtained, a second request was made to obtain the document authors. The second request is repeated until the entire document gets its authors. The data collection process generates a set of data containing author, title and year of author with the same author name (Figure 1). The process repeated until all of the authors registered in USPDC is processed.



Author	Title	Year
Rohendi D.	Game multimedia in numeracy learning for elementary school students	2017
Sumarna N.	Game multimedia in numeracy learning for elementary school students	2017
Sutarno H.	Game multimedia in numeracy learning for elementary school students	2017
Harsiti	Satellite image edge detection for population distribution pattern identification using levelset with morphological filtering process	2017
Munandar T.A.	Satellite image edge detection for population distribution pattern identification using levelset with morphological filtering process	2017
Suhendar A.	Satellite image edge detection for population distribution pattern identification using levelset with morphological filtering process	2017
Abdullah A.G.	Satellite image edge detection for population distribution pattern identification using levelset with morphological filtering process	2017
Rohendi D.	Satellite image edge detection for population distribution pattern identification using levelset with morphological filtering process	2017
Taher T.	Kinetic and thermodynamic adsorption studies of congo red on bentonite	2017
Mohadi R.	Kinetic and thermodynamic adsorption studies of congo red on bentonite	2017

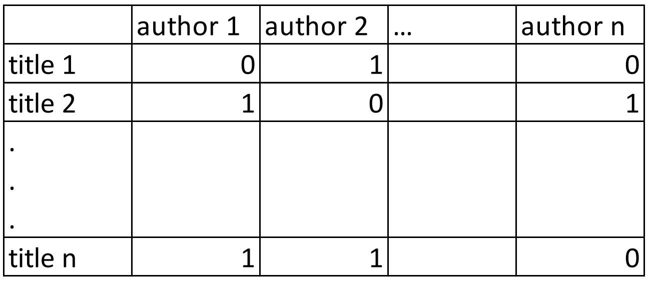
FIGURE 1. The data collection

2.2 DATA PREPROCESSING

In a set of author data, groupings are performed for each of the same author names, as well as for the titles. Grouping results are transformed into an adjacency matrix (Figure 2). For each aij valued with 1 or 0, 1 if title authored by author and 0 if it is not (Equation 1)

$$a_{ij} = \begin{cases} 1, title_i\ are\ connected \\ 0, otherwise \end{cases} \qquad (1)$$

The process repeated until all of the authors registered in USPDC is processed.



	author 1	author 2	…	author n
title 1	0	1		0
title 2	1	0		1
. . .				
title n	1	1		0

FIGURE 2. Co-author adjacency matrix

2.3 DATA ANALYSIS

To search for document conformity to an author, three level clustering should be performed (Figure 3)[12]; 1) titles clustering based on co-authorship; 2) titles clustering based on title similarity; 3) titles clustering based on publication year similarity.

First level clustering is done by grouping titles based on co-authorship. This level clustering uses a connected component (graph theory). Cluster search with connected component is done by using co-author adjacency matrix which has been prepared before. Each document (vertices) connected to each other by co-author (path) is grouped into one cluster. This step resulting clusters and its members.

The second level clustering is done by title similarity between cluster resulted in first level clustering. If the title similarity between two clusters is greater than 60%, it means that the two clusters are the same cluster, so it must be merged. If the similarity is less than 50%, it means that the two clusters are different clusters.

The third clustering is done only if the title similarity in second clustering is greater than 50% and less than 60%. The clustering is done by find publication year similarity.

There are two scenarios to calculate the publication year similarity, act span overlap and act span distance. Act span overlap is searched based on the similarity of the activity span (actspan) of the fragments to be compared. Actspan is an interval of the year from the start of the year to the end of the publication of (a). Results from actspan will be compared with Thoverlap value. Actspan distance is sought based on equality of activity peak (actpeak) which is the median of number of publications per year (a). Actspan distance is searched based on equality of activity peak (actpeak) which is the median of number of publications per year (a), then searched the difference between two fragments. The result of actpeak will be compared with Thdistance value. The threshold of Thoverlap and Thdistance is empirically determined.

The co-author matrix derived from the data preprocessing is used as an input to the clustering process of interconnecting authors. Clustering process is done using connected component. The connected component process resulted the number of title cluster and its members. Member title on one cluster indicates that the title has a relationship between authors.

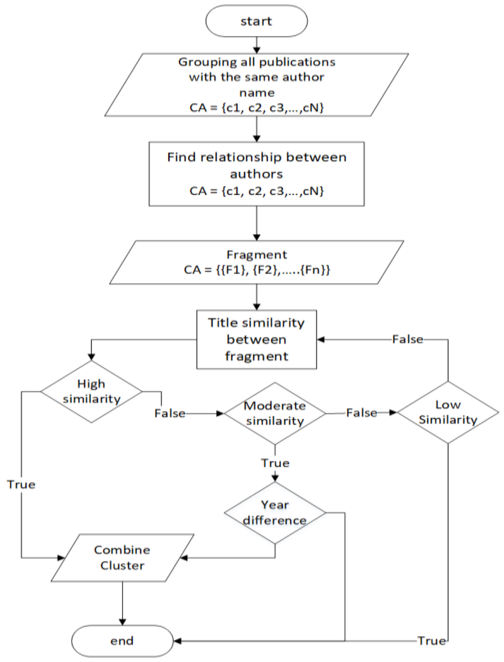


FIGURE 3. Two ways clustering process

3. RESULTS AND DISCUSSION

This section describes the implementation of a two-step hierarchical clustering technique at UPDC and a two-step hierarchical clustering technique evaluation. At the time of the experiment UPDC had 338 registered authors who had Scopus author ID and 966 documents consisting of 663 Universitas Sriwjaya affiliated documents and 303 which were not. System data processing resulting the system recommends 13 new documents belonging to registered author, and 15 documents listed do not belong to some registered author.

To evaluate the two-step hierarchical clustering technique three author names are used, namely Sri Haryati (Haryati, S) 100%, Dedi Rohendi (Rohendi, D) 83.33% and Budhi Setiawan (Setiawan, B). The name of the author was chosen by the name criterion has more than 4 documents in the UPDC, has more than 1 name in the Scopus data, and the remaining criterion is random. The results are Sri Haryati have 11documents, Dedi Rohendi have 12 documents and Budhi Setiawan have 410 documents.

TABLE 1.
Precision, Recall and F1 Score for three ambiguous names

Author	Author Number	Article Number	Precision	Recall	F1 Score
Sri Haryati	2	11	100%	100%	100%
Dedi Rohendi	2	12	83.33%	100%	90.91%
Budhi Setiawan	4	9	100%	100%	100%

There are 3 different authors with the same name which has 11 articles for Sri Haryati, 12 articles for Dedi Rohendi, and 9 articles for Budhi Setiawan (table 1). Each name has a different F1 Score value, 100% for Sri Haryati, 90.91 for Dedi Rohendi and 100% for Budhi Setiawan.

4. CONCLUSION

The two-step hierarchical clustering technique for name disambiguation has been applied to USPDC. From the application of this method to UPDPC, the system recommend 14 document in USPDC are not belong to authors registered in USPDC and 15 document outside USPDC belong to authors registered in USPDC. From the experiment of 3 author names, the two-step hierarchical clustering technique shows a good result in polysemy author name disambiguation.

ACKNOWLEDGEMENT

We thank the Universitas Sriwjaya which providing financial research Hibah Penelitian Dosen Muda SATEKs 2017 contract No. 1013/UN9.3.1/PP2017

REFERENCES

[1] C. A. D. Angelo, “A Heuristic Approach to Author Name Disambiguation in Bibliometrics Databases for Large-Scale Research Assessments,” vol. 62, no. 2, pp. 257–269, 2011.

[2] A. A. Ferreira and M. A. Gonçalves, “A Brief Survey of Automatic Methods for Author Name Disambiguation,” vol. 41, no. 2, 2012.

[3] N. R. Smalheiser and V. I. Torvik, “Author name disambiguation,” Annu. Rev. Inf. Sci. Technol., vol. 43, no. 1, pp. 1–43, 2009.

[4] H. Pasula, B. Marthi, B. Milch, S. J. Russell, and I. Shpitser, “Identity uncertainty and citation matching,” in Advances in neural information processing systems, 2003, pp. 1425–1432.

[5] J. Zhu, G. Cheong Fung, and X. Zhou, “Anddy: a system for author name disambiguation in digital library,” in Database Systems for Advanced Applications, 2010, pp. 444–447.

[6] M. Ali and A. Bart, “Use of ResearchGate and Google CSE for author name disambiguation,” Scientometrics, 2017.

[7] A. Khormali and J. Addeh, “A novel approach for recognition of control chart patterns: Type-2 fuzzy clustering optimized support vector machine,” ISA Trans., vol. 63, pp. 256–264, 2016.

[8] A. F. Santana, M. Andr, A. H. F. Laender, and A. A. Ferreira, “Incremental Author Name Disambiguation by Exploiting Domain-Specific Heuristics,” vol. 0, no. 0, 2016.

[9] J. Schulz, “Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses,” Scientometrics, 2016.

[10] P. Andruszkiewicz and S. Szepietowski, “Person Name Disambiguation for Building,” no. Ii, pp. 270–279, 2016.

[11] X. Lin, J. Zhu, Y. T. B, F. Yang, B. Peng, and W. Li, “A Novel Approach for Author Name,” pp. 169–182, 2017.

[12] R. Hazra, A. Saha, S. B. Deb, and D. Mitra, “An Efficient Technique for Author Name Disambiguation,” 2016.