DISERTASI

MODEL DALAM MENGATASI MASALAH DUPLIKASI PADA BASIS DATA RISET MENGGUNAKAN PENDEKATAN *THRESHOLD-BASED* DAN *RULE-BASED*

Diajukan Untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Doktor
Dalam Bidang Ilmu Teknik Informatika

M. MIFTAKUL AMIN
03013622025004

PROGRAM STUDI ILMU TEKNIK PROGRAM DOKTOR
FAKULTAS TEKNIK
UNIVERSITAS SRIWIJAYA

2024

HALAMAN PENGESAHAN

MODEL DALAM MENGATASI MASALAH DUPLIKASI PADA BASIS DATA RISET MENGGUNAKAN PENDEKATAN *THRESHOLD-BASED* DAN *RULE-BASED*

DISERTASI

Diajukan Untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Doktor dalam Bidang Ilmu Teknik Informatika
Fakultas Teknik Universitas Sriwijaya

Oleh:

M. MIFTAKUL AMIN
NIM. 03013622025004

Palembang, 19 Maret 2024

Promotor

Prof. Deris Stiawan, M.T., Ph.D.
NIP. 197806172006041002

Ko-Promotor

Dr. Ermatita, M. Kom.
NIP. 196709132006042001

Mengetahui,

Plh. Dekan Fakultas Teknik

Dr. Ir. Bhakti Yudho Suprapto, S.T., M.T., IPM.
NIP. 197502112003121002

Koordinator Program Studi

Prof. Dr. Ir. Nukman, M.T.
NIP. 195903211987031001

HALAMAN PERSETUJUAN

Karya tulis ilmiah berupa laporan disertasi dengan judul **"MODEL DALAM MENGATASI MASALAH DUPLIKASI PADA BASIS DATA RISET MENGGUNAKAN PENDEKATAN *THRESHOLD-BASED* DAN *RULE-BASED*"** telah dipertahankan di hadapan Tim Penguji Karya Tulis Ilmiah Program Studi Ilmu Teknik Program Doktor Fakultas Teknik Universitas Sriwijaya pada tanggal 19 Maret 2024.

Palembang, 19 Maret 2024

Tim Penguji Karya Tulis Ilmiah berupa Laporan Disertasi

Ketua:

Prof. Dr. Ir. Nukman, MT.
NIP. 195903211987031001 ( )

Promotor:

Prof. Deris Stiawan, M.T., Ph.D.
NIP. 197806172006041002 ( )

Ko-Promotor:

Dr. Ermatita, M.Kom.
NIP. 196709132006042001 ( )

Anggota:

1. Dr. Lukman, S.T., M.Hum.
   NIP. 197805112003121002 ( )
2. Prof. Dr. Yusuf Hartono, M.Sc.
   NIP. 196411161990031002 ( )
3. Dr. Abdiansah, S.Kom., M.CS.
   NIP. 198410012009121005 ( )

Mengetahui,

Plh. Dekan Fakultas Teknik

Dr. Ir. Bhakti Yudho Suprapto, S.T., M.T., IPM.
NIP. 197502112003121002

Koordinator Program Studi,

Prof. Dr. Ir. Nukman, M.T.
NIP. 195903211987031001

iii

HALAMAN PERNYATAAN INTEGRITAS

Yang bertanda tangan di bawah ini:

Nama	: M Miftakul Amin
NIM.	: 03013622025004
Tempat, Tanggal Lahir	: Raman Aji, 17 Desember 1979
Jurusan/Program Studi	: S3 Ilmu Teknik – Bidang Teknik Informatika
Fakultas	: Teknik Universitas Sriwijaya
Alamat Rumah	: Perumahan Kota Modern Sriwijaya Blok A2/52, Jalan Kadir TKR No. 99, Karanganyar, Gandus, Palembang, 30148
Alamat E-Mail	: miftakul_a@polsri.ac.id

Dengan ini menyatakan dengan sesungguhnya bahwa disertasi yang berjudul "MODEL DALAM MENGATASI MASALAH DUPLIKASI PADA BASIS DATA RISET MENGGUNAKAN PENDEKATAN THRESHOLD-BASED DAN RULE-BASED"

Bebas dari plagiarisme dan bukan hasil karya orang lain

Apabila di kemudian hari ditemukan seluruh atau sebagian dari disertasi tersebut terdapat indikasi plagiarisme, saya bersedia menerima sanksi dengan peraturan perundang-undangan yang berlaku.

Demikian pernyataan ini saya buat dengan sesungguhnya tanpa ada paksaan dari siapapun dan untuk dipergunakan sebagaimana mestinya.

Palembang, 19 Maret 2024
Yang membuat pernyataan,

M Miftakul Amin
NIM. 03013622025004

KATA PENGANTAR

Puji syukur dipanjatkan ke hadirat Allah S.W.T., atas limpahan rahmat dan hidayah-Nya dapat menyelesaikan penelitian Disertasi dengan judul "MODEL DALAM MENGATASI MASALAH DUPLIKASI PADA BASIS DATA RISET MENGGUNAKAN PENDEKATAN THRESHOLD-BASED DAN RULE-BASED". Penyelesaian penyusunan Disertasi ini dapat diselesaikan dengan dukungan, bimbingan, dan arahan dari berbagai pihak. Dalam kesempatan ini diucapkan terimakasih yang sedalam-dalamnya kepada Yth:

1. Prof. Dr. Taufiq Marwa, S.E., M.Si., selaku Rektor Universitas Sriwijaya.
2. Prof. Dr. Eng. Ir. H. Joni Arliansyah, M.T., selaku Dekan Fakultas Teknik Universitas Sriwijaya.
3. Prof. Dr. Ir. Nukman, M.T., selaku Koordinator Program Studi Ilmu Teknik Program Doktor, Fakultas Teknik Universitas Sriwijaya yang telah memberikan motivasi untuk menyelesaikan Disertasi.
4. Bapak Prof. Deris Stiawan, M.T., Ph.D., sebagai promotor Disertasi yang telah memberikan bimbingan, arahan, dan motivasi untuk menyelesaikan Disertasi.
5. Ibu Dr. Ermatita, M.Kom., sebagai ko-promotor atas bimbingan, arahan, dan motivasi yang diberikan dalam penyusunan Disertasi.
6. Dosen-dosen Program Studi Ilmu Teknik (S3) BKU Teknik Informatika beserta staf yang telah mengajarkan, berbagi ilmu, dan juga membantu penulis dalam pengurusan persyaratan administrasi.
7. Pimpinan, rekan kerja dan staf Politeknik Negeri Sriwijaya atas dukungannya selama menempuh pendidikan S3.
8. Rekan-rekan jurusan teknik komputer Politeknik Negeri Sriwijaya, terimakasih atas dukungannya.
9. Rekan-rekan di Program Studi Ilmu Teknik (S3) BKU Teknik Informatika khususnya angkatan 2019 Semester Genap yang telah bersama-sama berjuang menyelesaikan pendidikan S3.

10. Keluarga tercinta, Istri dan anak-anakku, yang telah memberikan dorongan, pengertian, dan pengorbanan.
11. Semua pihak yang telah memberikan bantuan sehingga selesainya disertasi ini.

Mudah-mudahan Allah S.W.T. membalas semua amal kebajikan tersebut dan menjadikannya sebagai amal shaleh.

Salam Hormat,

M Miftakul Amin
NIM. 03013622025004

RINGKASAN

Basis data riset dalam pangkalan data SINTA berasal dari beberapa sumber seperti *Google Scholar*, *Scopus*, dan *Web of Science*. Namun demikian masih terdapat duplikasi data yang secara logika merujuk pada entitas yang sama, sehingga mengurangi kualitas data yang terdapat di dalamnya. Tidak selalu adanya informasi yang unik seperti *primary key* atau *unique identifier* dalam basis data riset mengakibatkan terjadi duplikasi. Dalam basis data yang besar, duplikasi ini sulit untuk dideteksi. Duplikasi data ini menjadikan beberapa perhitungan produktifitas publikasi ilmiah menjadi kurang valid, seperti perhitungan *impact factor* (IF) pada level jurnal dan *h-index* pada level *author*, dan beberapa perhitungan sejenis dengan memanfaatkan basis data riset sebagai sumber datanya. Dalam penelitian ini konsep *entity matching* dapat dijadikan sebagai salah satu pendekatan untuk mengatasi terjadinya duplikasi pada basis data riset. Ditinjau dari aspek teknis, pemrosesan *entity matching* pada umumnya bersifat manual dan menyita waktu, rentan terjadi *error*, dan tidak relevan diaplikasikan pada jumlah data yang besar, sehingga diperlukan pendekatan yang bersifat semi-otomatis untuk meningkatkan kinerja *matching*. Pada penelitian ini menggunakan pendekatan metode *threshold-based* dan *rule-based*. Dengan menambahkan *rule* dapat meningkatkan hasil deteksi yang dilakukan oleh metode *threshold-based* dan memberikan hasil yang lebih optimal. Penelitian ini diawali dengan meninjau beberapa pendekatan yang sudah pernah dilakukan oleh para peneliti pada bidang *entity matching* dan deteksi duplikasi, kemudian dilanjutkan dengan mengidentifikasi beberapa karakteristik yang dapat meningkatkan kinerja *entity matching*, selanjutnya melakukan studi perbandingan dalam proses *entity matching* yang diaplikasikan pada basis data riset. Hasil penelitian ini menyajikan beberapa evaluasi kinerja model yang menunjukkan bahwa penggunaan *threshold* dalam pembentukan rule menghasilkan kinerja yang lebih baik, dibandingkan menggunakan *threshold* saja. Pada dataset Wos rule 4 dan rule 5 yang dibentuk memberikan hasil kinerja terbaik dengan nilai 100,00% untuk *accuracy*, *precision*, *recall*, dan *F1-measure*. Pada *dataset* scopus rule 4 dan rule 5 menghasilkan nilai 100,00% untuk *accuracy* dan *precision*, sedangkan nilai 96,00% dan 98,00% untuk nilai *recall* dan *F1-measure*. Pada *dataset google scholar* untuk rule 4 dan rule 5, nilai accuracy sebesar 100,00%, nilai precision 96,00%. Nilai recall untuk rule 4 sebesar 96,00% dan sebesar 97,00% untuk F1-measure. Nilai F1-measure untuk rule 4 sebesar 97,00% dan rule 5 sebesar 98,00%.

Kata kunci: deteksi duplikasi, duplikasi, basis data riset

SUMMARY

The research database in the SINTA database comes from several sources such as Google Scholar, Scopus, and Web of Science. However, there is still duplication of data that logically refers to the same entity, thus reducing the quality of the data contained therein. There is not always unique information such as primary keys or unique identifiers in research databases, resulting in duplication. In large databases, this duplication is difficult to detect. This data duplication makes some calculations of scientific publication productivity less valid, such as the calculation of impact factor (IF) at the journal level and h-index at the author level, and several similar calculations by utilizing research databases as data sources. In this research, the concept of entity matching can be used as an approach to overcome duplication in research databases. From a technical aspect, entity matching processing is generally manual and time-consuming, prone to errors, and irrelevant when applied to large amounts of data, so a semi-automatic approach is needed to improve matching performance. This research uses threshold-based and rule-based approaches. Adding rules can improve the detection results performed by the threshold-based method and provide more optimal results. This research begins by reviewing several approaches that have been carried out by researchers in the field of entity matching and duplication detection, then continues by identifying several characteristics that can improve entity matching performance, then conducting a comparative study in the entity matching process applied to research databases. The results of this study present several model performance evaluations that show that the use of thresholds in rule formation results in better performance, compared to using thresholds alone. In the Wos dataset, rule 4 and rule 5 formed provide the best performance results with a value of 100.00% for accuracy, precision, recall, and F1-measure. On the Scopus dataset rule 4 and rule 5 produce a value of 100.00% for accuracy and precision, while the value of 96.00% and 98.00% for the value of recall and F1-measure. On the google scholar dataset for rule 4 and rule 5, the accuracy value is 100.00%, the precision value is 96.00%. The recall value for rule 4 is 96.00% and 97.00% for F1-measure. The F1-measure value for rule 4 is 97.00% and rule 5 is 98.00%.

Keywords: duplication detection, duplication, research database

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1. 1. Latar Belakang

Interaksi dan penggunaan internet yang terus meningkat, membawa konsekuensi informasi disimpan dalam bentuk elektronik, sehingga dapat diakses dan dipertukarkan dengan mudah. Pengguna dapat melakukan pencarian informasi dari mana saja dan kapan saja untuk mengakses sumber informasi digital, dan mencari koleksi informasi sesuai dengan kebutuhan. Koleksi data elektronik ini mudah dipertukarkan, sehingga dapat menjadi salah satu sarana untuk diseminasi informasi di bidang pendidikan yang berkualitas. Pertukaran data dan informasi repositori ilmiah dapat menjadi salah satu cara mempersempit adanya jurang informasi yang masih menyimpan informasi digital ilmiah secara teritorial dan parsial di Indonesia. Terbukanya akses informasi ilmiah secara luas, dapat meningkatkan wawasan masyarakat dalam literasi ilmiah.

Menurut Amorim et al. (Amorim et al., 2017) menyatakan bahwa manajemen data penelitian dengan cepat menjadi perhatian penting bagi para peneliti, sehingga institusi dan lembaga penelitian perlu menyediakan separangkat *tools* untuk mendukung pengelolaan data publikasi ilmiah. Dalam aspek tata kelola menurut Heidorn (Heidorn, 2008) pengelolaan aset ilmiah ini menuntut informasi yang terkandung di dalamnya berisi informasi yang valid. Di samping itu menurut Subroto et al. (Subroto et al., 2014) salah satu kendala dalam pengumpulan data riset adalah jumlah artikel dan penerbit artikel yang terus bertambah seiring waktu. Di sisi lain, pengumpulan data dengan cara manual akan menyita waktu, sehingga diperlukan pendekatan yang bersifat semi-otomatis untuk mendapatkan basis data riset dari berbagai sumber yang terpercaya.

Saat ini terdapat beragam perangkat lunak yang digunakan dalam mengelola publikasi ilmiah dalam perangkat lunak *institutional repository* (IR)

seperti *open journal system* (OJS)[1], *eprints*[2], *dspace*[3], dan *digital library* yang bertujuan memudahkan dalam pengelolaan dokumen publikasi ilmiah. Hal ini menjadikan pengelolaan data dilakukan oleh masing-masing penerbit artikel maupun lembaga penelitian secara terpisah. Pernyataan yang disampaikan oleh Lynch (Lynch, 2003) menegaskan bahwa repositori ilmiah ini dikelola oleh Peguruan Tinggi dan lembaga penelitian untuk pengelolaan dan penyebaran informasi digital publikasi ilmiah yang telah dihasilkan oleh institusi dan para peneliti.

Dalam konteks di Indonesia dari hulu ke hilir seperti dapat dilihat pada Gambar 1.1 memberikan informasi bahwa pangkalan data GARUDA merupakan salah satu pangkalan data yang berperan melakukan integrasi data publikasi ilmiah sebagai basis data riset. Sedangkan pangkalan data SINTA menurut Lukman et al. (Lukman et al., 2018) berperan sebagai pengindeks dan pemberian skor untuk melakukan pengukuran produktifitas penelitian, yang tidak hanya mengambil data dari basis data riset di Indonesia, tetapi juga dari basis data ilmiah internasional seperti *Web of Science* (WOS), *Google Scholar* (GS), dan *SCOPUS*.

Gambar 1.1 Pemetaan Basis Data Riset di Indonesia

Lebih lanjut menurut Caragea et al. (Caragea et al., 2014) penemuan entitas yang sama dalam beberapa domain yang berasal dari sumber database yang berbeda merupakan salah satu fungsi yang penting dalam dunia digital. Sebagai contoh seorang pakar di bidang paten yang akan menilai sebuah usulan paten dapat

[1] https://pkp.sfu.ca/ojs/
[2] https://www.eprints.org/uk/
[3] https://duraspace.org/dspace/

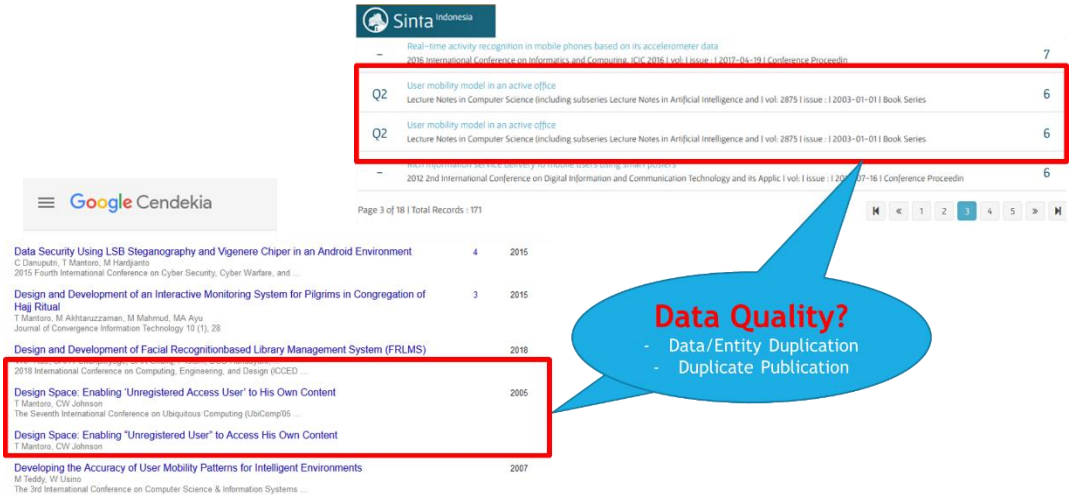
menemukan sejumlah paten yang mirip atau yang memiliki irisan dengan paten yang sedang diusulkan. Contoh lain adalah seorang dokter dapat menemukan resep obat yang tepat untuk diberikan kepada pasien dengan kasus penyakit serupa yang pernah ditemui sebelumnya.

Gambar 1.2 Konsep Entitas dalam Artikel Ilmiah

Dalam konteks publikasi ilmiah, sebuah artikel dapat dipandang sebagai sebuah entitas. Gambar 1.2 memperlihatkan bahwa sebuah dokumen publikasi ilmiah dapat dianggap sebagai sebuah entitas yang unik dengan melihat beberapa informasi yang terkandung di dalamnya. Informasi seperti judul publikasi, penulis, *venue* yang merupakan informasi nama jurnal atau prosiding, dan periode publikasi (tahun, volume, nomor, halaman) merupakan informasi penting sebagai penanda sebuah artikel ilmiah. Sebuah artikel ilmiah dapat dibedakan satu dengan yang lainnya berdasarkan informasi yang terkandung dalam entitas tersebut.

Seiring waktu informasi yang ada dalam pangkalan data SINTA akan terus berkembang dengan bertambahnya jumlah publikasi ilmiah yang dihasilkan oleh para Dosen dan peneliti. Kualitas data yang ada dalam repositori ilmiah dan pengindeks diharapkan memiliki validitas dan kualitas yang tinggi, sehingga dapat dijadikan acuan berintegritas bagi masyarakat ilmiah di Indonesia. Melihat pada Gambar 1.3 terlihat indikasi adanya duplikasi data yang muncul di *Google Scholar* dan pangkalan data SINTA. Hal ini dapat dijadikan sebagai satu buah fakta

perlunya meninjau kembali aspek kualitas data terkait dengan basis data riset yang terkandung didalamnya.



Gambar 1.3 Duplikasi Data pada Basis Data Riset

Menurut Christen (Christen, 2012b) salah satu faktor penyebab terjadinya duplikasi pada basis data riset adalah adanya nama penulis yang sama dan disimpan dengan inisial yang sama pada basis data riset. Demikian juga disebabkan adanya nama penulis yang sama pada domain penelitian yang sama. Penyebab lain adalah adanya penyingkatan *venue* yang berisi informasi nama jurnal atau nama prosiding dengan tidak menuliskan secara lengkap.

Berdasarkan hasil penelitian dari Mishra et al. (Mishra et al., 2016) terkait dengan informasi afiliasi yang melekat pada seorang penulis juga mengakibatkan terjadinya duplikasi. Hal ini dapat ditinjau pada kasus di mana seorang penulis bertugas pada afiliasi yang berbeda pada saat melakukan publikasi. Factor lain dapat juga dikarenakan perpindahan tugas ataupun karena mewakili afiliasi yang berbeda.

Tabel 1.1 Contoh Duplikasi pada Basis Data *Google Scholar*

AM Miftakul. Image Steganography Dengan Metode Least Significant Bit (LSB). *journal CSRID* 6, 81-87, 2014.
MM Amin. Image steganography dengan metode least significant bit (SLB). *Jurnal Computer Science Research and Its Development (CSRID)* 6 (1), 2014.

Mengacu pada Tabel 1.1 dapat dilihat bahwa terjadi duplikasi pada basis data riset *Google Scholar* yang dijadikan sebagai salah satu sumber data rujukan pada pangkalan data SINTA. Duplikasi ini terjadi karena tersimpan dari sumber penyimpanan yang berbeda, sementara dari item datanya juga terdapat adanya perbedaan penulisan nama penulis dan penulisan nama *venue* dalam hal ini adalah penulisan nama jurnal secara lengkap dan salah satunya menggunakan singkatan.

Tabel 1.2 Contoh Duplikasi pada Basis Data *Scopus*

M Miftakul Amin, Andino Maseleno, K Shankar, Eswaran Perumal, RM Vidhyavathi, SK Lakshmanaprabu. (2018).Active database system approach and rule based in the development of academic information system. *International Journal of Engineering & Technology*, 7(2.26), 95-101. DOI: 10.26594/register.v4i1.1129.
M Miftakul Amin, Andino Maseleno, K Shankar, Eswaran Perumal, RM Vidhyavathi, SK Lakshmanaprabu. (2018).Active database system approach and rule based in the development of academic information system. *International Journal of Engineering & Technology*, 7(2.26), 95-101. DOI: 10.14419/ijet.v7i3.9789.

Observasi yang dilakukan pada salah seorang penulis pada pangkalan data SINTA dan menelusuri publikasi yang telah dilakukan pada basis data *SCOPUS* juga terdapat duplikasi seperti dilihat pada Tabel 1.2. perbedaan ini terletak pada informasi DOI dari artikel ilmiah yang terekam dalam basis data *SCOPUS* pada pangkalan data SINTA. Informasi DOI ini dapat terjadi karena sumber pengambilan digital berbeda dari masing-masing DOI.

Menurut Naumann & Herschel (Naumann & Herschel, 2010) besarnya volume basis data riset, sehingga basis data ilmiah ini biasanya tidak terintegrasi ke dalam satu buah sistem basis data tunggal, tetapi akan diberikan tautan ke representasi lain dalam basis data. Dapat juga digunakan strategi dengan melakukan integrasi perintah query yang berasal dari sumber basis data riset yang berbeda. Berbagi basis data riset dapat meningkatkan dampak dalam iklim ilmiah di Indonesia dan meningkatkan transparansi dalam pelaksanaan riset. Transparansi ini

penting dalam upaya meningkatkan kualitas penelitian, karena hasil penelitian yang telah dihasilkan oleh para peneliti dapat dikembangkan lagi oleh berbagai pihak yang memiliki bidang penelitian yang sama, dan juga dapat dijadikan sebagai kontrol terhadap masyarakat ilmiah untuk menghindari terjadinya praktek plagiasi dan pelanggaran etika publikasi ilmiah. Bahkan Irawan et al. (Irawan et al., 2019) mengungkapkan bahwa dalam konteks di Indonesia, terdapat beberapa masalah utama dalam berbagi basis data riset, diantaranya disebabkan faktor siklus data yang cukup pendek dan kurangnya inisiatif dalam aspek pengelolaan basis data riset.

Sesuai saran yang disampaikan oleh Trippel & Zinn (Trippel & Zinn, 2021) basis data riset sebaiknya dapat dihimpun, dievaluasi, diinventarisasi, dan dapat diakses dengan cara yang mudah bagi yang membutuhkan. Basis data riset dihimpun dari berbagai sumber yang terpercaya dan valid, untuk menyediakan sumber daya informasi riset yang berkualitas. Demikian juga yang disampaikan oleh Lynch (Lynch, 2003) menyatakan bahwa pengelolaan aset ilmiah ini menjadi perhatian para peneliti dan afiliasi yang harus mengelola basis data riset dengan baik dan memastikan bahwa informasi yang terkandung di dalamnya merupakan informasi yang benar dan valid.

Kualitas data dalam basis data riset menjadi salah satu perhatian penting untuk menyajikan data yang handal. Aspek duplikasi yang muncul dalam basis data riset menjadi salah satu faktor cerminan kualitas data yang perlu diperbaiki yang salah satu faktor penyebabnya karena input data yang kurang sempurna. Sebagaimana diungkapkan oleh Kilkenny & Robinson (Kilkenny & Robinson, 2018) yang menyatakan istilah sampah masuk sampah keluar, merupakan pertimbangan penting terkait kualitas data, dan kebutuhan tingkat akurasi informasi yang tinggi. Akurasi data sangat dibutuhkan untuk melakukan analisis data dalam jumlah yang besar. Terkait dengan akurasi ini, maka menurut Ling et al. (Ling et al., 2013) akurasi data dapat dicapai dengan adanya basis data dengan jumlah eror yang minimal. Sebagaimana diungkapkan oleh Liu et al. (Liu et al., 2011) bahwa salah satu bentuk eror dalam sebuah basis data adalah terjadinya duplikasi input data. Hal senada juga diungkapkan oleh Wei et al. (Wei et al., 2006) bahwa

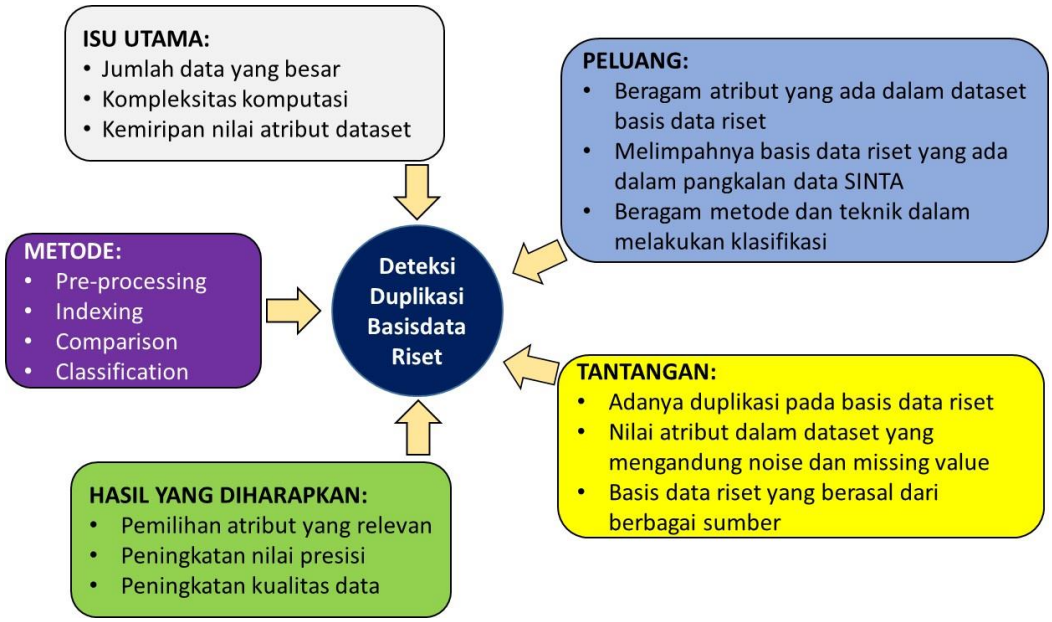
duplikasi data merupakan salah satu masalah yang sering ditemui dalam manajemen basis data. Deteksi duplikasi juga merupakan aspek yang sangat penting dalam tahapan pra-pemrosesan dan sebagai sebuah usaha untuk mengurangi jumlah redudansi dalam proses integrasi data, sebagaimana diungkapkan oleh Alenazi & Kamsuriah (Alenazi & Kamsuriah, 2016).

Secara prinsip dalam konteks basis data bahwa sebuah record dapat diidentifikasi sebagai duplikasi atau tidak adalah dengan melihat informasi penanda unik seperti kunci utama dalam definisi sebuah struktur tabel sebagaimana diungkapkan oleh Akel (Akel, 2012). Hal senada juga diungkapkan oleh Murray (Murray, 2015) yang menyatakan bahwa tidak adanya pendanda unik merupakan penyebab terjadinya duplikasi dalam sebuah basis data. Dalam konteks basis data riset, dimana sebuah artikel ilmiah disimpan secara digital, salah satu informasi yang dapat dijadikan sebagai penanda unik adalah DOI. Tetapi secara faktual, masih banyak artikel ilmiah yang tidak memiliki informasi DOI, sehingga perlu adanya pendekatan dan alternatif lain untuk melakukan identifikasi terjadinya duplikasi pada basis data riset. Bahkan menurut penelitian yang dilakukan oleh Gyawali et al. (Gyawali et al., 2020) mencatat dari 8,500 dokumen dari berbagai repositori ilmiah, terdapat 82% data yang berhasil dikimpulkan tidak memiliki DOI. Sebagian besar DOI direpresentasikan dalam bentuk yang umum.

Deteksi duplikasi merupakan masalah untuk mengidentifikasi record dalam basis data yang mewakili entitas dunia nyata yang sama. Menurut Akel (Akel, 2012) pada saat sebuah record tidak memiliki informasi kolom yang menandai record yang unik, maka hal ini akan membuat deteksi duplikasi menjadi sesuatu yang sulit. Dalam melakukan integrasi basis data yang berasal dari sumber yang berbeda, akan muncul masalah heterogenitas yang secara umum dapat dikelompokkan menjadi 2 bagian, yaitu 1) heterogenitas struktural dan 2) heterogenitas leksikal. Heterogenitas struktural terjadi ketika dua sumber data memiliki struktur kolom yang berbeda, sebagai contoh pada salah satu sumber data sebuah informasi alamat disimpan dalam satu field tunggal, sementara pada sumber data yang lain dipecah dalam beberapa kolom seperti jalan, kota, dan kode pos. Sedangkan heterogenitas leksikal terjadi jika sumber data yang berbeda memiliki

struktur kolom yang sama, sedangkan memiliki representasi data yang berbeda. Sebagai contoh data nama pada salah satu sumber basis data adalah M. M. Amin, sedangkan pada sumber data yang lain adalah M. Miftakul Amin.

Seperti dapat dilihat pada Gambar 1.4 bahwa deteksi duplikasi pada basis data riset ini memiliki isu utama berupa jumlah data basis data riset yang besar pada pangakalan data SINTA dari waktu ke waktu, sehingga membawa konsekuensi pada kompleksitas komputasi, dan banyaknya informasi dengan tingkat kemiripan yang tinggi. Pendekatan pencocokan entitas dapat digunakan sebagai tahapan baku dimulai dari tahapan *pre-processing*, *indexing*, *comparison*, dan *classification*. Beragam atribut dataset dalam basis data riset, melimpahnya basis data riset pada pangkalan data SINTA, dan beragam metode dan teknik dalam melakukan klasifikasi merupakan peluang yang membuka pintu lebar dalam penelitian ini. Namun demikian adanya tantangan berupa duplikasi pada basis data riset, adanya inormasi missing value dan noise, serta basis data riset yang berasal dari berbagai sumber merupakan tantangan yang perlu mendapatkan perhatian yang cukup dalam penelitian ini. Selanjutnya hasil yang diharapkan adalah pemilihan atribut yang relevan, peningkatan nilai presisi, dan peningkatan kualitas data pada basis data riset.



Gambar 1.4 Isu dan Tantangan Penelitian

Penelitian deteksi duplikasi menggunakan pendekatan *threshold* telah dilakukan oleh Mishra et al. (Mishra et al., 2013) dengan melibatkan atribut temporer dan atribut non-temporer. Penelitian ini menghasilkan pengujian model dengan nilai rata-rata *f-measure* sebesar 97,83% dari 7 buah dataset yang diuji. Penelitian lain yang dilakukan oleh Ektefa et al. (Ektefa et al., 2011) telah mendapatkan nilai *f-measure* sebesar 99,1% menggunakan pendekatan *threshold-based* pada dataset *restaurant*. Pendekatan *machine learning* telah dilakukan oleh Alifikri & Bijaksana untuk pembentukan dataset *Indonesian name matching* (Alifikri & Bijaksana, 2018). Dataset diperoleh dari data nama siswa dari data pokok pendidikan SMA/SMK dan data pilkada 2017 yang dikeluarkan oleh KPU. Hasil pengukuran menunjukkan nilai *f-measure* sebesar 97,20%. Pendekatan *machine learning* juga dilakukan oleh Sefid et al. (Sefid et al., 2019a) untuk melakukan deteksi duplikasi menggunakan dataset yang berasal dari *citeseerx, wos, pubmed*, dan *DBLP*. Penelitian ini menghasilkan nilai *precision* sebesar 98,10%, nilai *recall* 86,90%, dan nilai *f-measure* sebesar 92,20%. Penelitian yang dilakukan oleh López-Cuadrado et al. (López-Cuadrado et al., 2020) melakukan kombinasi antara *machine learning* dan pembentukan *rule* untuk melakukan deteksi duplikasi pada dataset obat. Penelitian ini menghasilkan nilai *accuration* sebesar 85%. Penelitian pada pembentukan *rule* dalam melakukan deteksi duplikasi dilakukan oleh Jiang et al. (Jiang et al., 2014). Sejumlah 7 buah rule dibuat untuk melakukan deteksi duplikasi. Penelitian ini hanya melakukan perbandingan antara waktu eksekusi dari setiap dataset dan jumlah record yang diindikasikan sebagai duplikat.

Beragam metode telah digunakan dalam tugas deteksi duplikasi, sehingga memberikan peluang dan tantangan. Dalam penelitian ini pendekatan *threshold-based* dan *rule-based* akan dikombinasikan untuk meningkatkan hasil yang lebih optimal untuk melakukan deteksi duplikasi. Sesuai dengan penelitian yang dilakukan oleh Sefid et al. (Sefid et al., 2019b), merekomendasikan 4 buah atribut yang melekat pada artikel ilmiah untuk pembentukan pasangan record dalam penentuan duplikasi. Seperti juga yang disampaikan oleh Draisbach & Naumann (Draisbach & Naumann, 2013) bahwa faktor pemilihan nilai *threshold* yang optimal merupakan salah satu kesulitan utama dalam konfigurasi untuk menentukan sistem

duplikasi data. Kesederhanaan komputasi dari pendekatan *threshold-based* dan *rule-based* merupakan salah satu faktor yang dipertimbangkan, sehingga dapat diimplementasikan secara mudah dalam berbagai Bahasa pemroraman. Pangkalan data SINTA sebagai pangkalan data penelitian di Indonesia saat ini masih sedikit yang melakukan eksplorasi terkait dengan adanya fakta duplikasi, sehingga membuka peluang penelitian untuk melakukan deteksi duplikasi. Informasi struktur dataset telah tersedia dan akan dilakukan tahapan pemrosesan lebih lanjut untuk melakukan tahapan deteksi duplikasi pada pangkalan data SINTA, dimana sumber datanya berasal dari *Google Scholar*, *Scopus*, dan *Web of Science*.

1. 2. Rumusan Masalah

Dari uraian yang telah dipaparkan pada bagian latar belakang, maka dapat dirumuskan beberapa permasalahan sebagai berikut:

1. Bagaimana model pendekatan *threshold-based* dan *rule-based* dapat digunakan untuk mengatasi masalah duplikasi pada pangkalan data SINTA?
2. Bagaimana mengukur kinerja serta efektivitas dari model yang dikembangkan, sehingga dapat melakukan validasi dan verifikasi terhadap kinerja model?
3. Bagaimana mengembangkan model deteksi duplikasi untuk meningkatkan kinerja dalam mengatasi masalah duplikasi pada basis data riset?

1. 3. Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah:

1. Membangun sebuah model deteksi duplikasi yang dapat digunakan untuk mengatasi masalah duplikasi data pada basis data riset secara simultan.
2. Melakukan evaluasi dan validasi terhadap model yang telah dikembangkan, sehingga memiliki kinerja yang memadai untuk dijadikan sebagai rujukan dalam proses deteksi duplikasi.
3. Meningkatkan kinerja model yang dikembangkan sehingga dapat diaplikasikan pada basis data riset.

1. 4. Batasan Masalah

Dengan beragam jenis basis data riset dan kompleksitas informasi yang menjadi karakteristik dari basis data riset, maka penelitian ini dibatasi pada:

1. Basis data riset yang digunakan dalam penelitian ini adalah *Google Scholar* (GS), *Scopus*, dan *Web of Science (WOS)*.
2. Lingkup publikasi mengacu pada bidang informatika, sistem informasi, elektro, dan multimedia.
3. Penelitian ini tidak membahas secara spesifik faktor-faktor yang menyebabkan terjadinya duplikasi, maupun validitas informasi yang terekam pada basis data riset.

1. 5. Kontribusi Penelitian

Kontribusi penelitian ini adalah menyediakan alternatif model *entity matching* baru menggunakan gabungan pendekatan *threshold-based* dan *rule-based*. Model ini nantinya berperan untuk melakukan deteksi duplikasi yang ada pada basis data riset pada pangkalan data SINTA, sehingga meningkatkan kualitas data yang ada di dalamnya.

Aspek keterbaruan (*novelty*) dari penelitian ini adalah dengan melihat fakta bahwa basis data riset dalam pangkalan data SINTA masih terdapat duplikasi, sehingga perlu adanya upaya untuk mengurangi terjadinya hal tersebut dan menjadikan informasi yang terkandung di dalamnya menjadi lebih valid. Pendekatan metode *rule-based* dapat digunakan untuk meningkatkan hasil deteksi duplikasi yang diperoleh, jika hanya menerapkan pendekatan tunggal menggunakan *threshold-based* saja. Investigasi karakteristik yang melekat pada pangkalan data SINTA dan parameter *threshold* dan pendefinisian *rule* merupakan salah satu aspek penting sebagai *novelty* dari penelitian ini.

1. 6. Sistematika Penulisan

Disertasi ini terdiri dari 5 (lima) bab. Bab 1 menjelaskan tentang pendahuluan dari penelitian. Bab 2 menjelaskan tentang beberapa teori dasar yang menjadi fondasi dalam penelitian ini seperti konsep interoperabilitas, konsep

pencocokan entitas, penelitian terkait dengan topik penelitian, dan model evaluasi yang dapat diterapkan dalam penelitian tentang pencocokan entitas. Bab 3 menjelaskan tentang metode penelitian, dan Bab 4 menjelaskan hasil dan pembahasan penelitian, dan bab 5 berisi kesimpulan dan saran.

DAFTAR PUSTAKA

Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., Papotti, P., Stonebraker, M., & Tang, N. (2016). Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*, *9*(12), 993–1004. https://doi.org/10.14778/2994509.2994518

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Akel, O. H. (2012). A Comparative Study of Duplicate Record Detection Techniques. *Master Degree in Computer Science, Middle East University*.

Alemu, G., Stevens, B., & Ross, P. (2012). Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries: A social constructivist approach. *New Library World*, *113*(1), 38–54. https://doi.org/10.1108/03074801211199031

Alenazi, S. R., & Kamsuriah. (2016). Record duplication detection in database: A review. *International Journal on Advanced Science, Engineering and Information Technology*, *6*(6), 838–845. https://doi.org/10.18517/ijaseit.6.6.1368

Ali, A., Emran, N. A., & Asmai, S. A. (2021). Missing values compensation in duplicates detection using hot deck method. *Journal of Big Data*, *8*(1), 1–17. https://doi.org/10.1186/s40537-021-00502-1

Alifikri, M., & Bijaksana, M. A. (2018). Indonesian name matching using machine learning supervised approach. *Journal of Physics: Conference Series*, *971*(1). https://doi.org/10.1088/1742-6596/971/1/012038

Amorim, R. C., Castro, J. A., Rocha da Silva, J., & Ribeiro, C. (2017). A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, *16*(4), 851–862. https://doi.org/10.1007/s10209-016-0475-y

Barlaug, N., & Gulla, J. A. (2021). Neural Networks for Entity Matching: A Survey. *ACM Transactions on Knowledge Discovery from Data*, *15*(3), 1–36. https://doi.org/10.1145/3442200

Batini, C., & Scannapieco, M. (2006). Web services. In *Data Quality, Concepts, Methodologies and Techniques*. Springer. https://doi.org/10.1007/978-1-4020-4749-5_4

Bharambe, D., Jain, S., & Jain, A. (2012). A Survey : Detection of Duplicate Record. *International Journal of Emerging Technology and Advanced Engineering*, *2*(11), 298–307.

Blakely, T., & Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology*, *31*(6), 1246–1252.

Caragea, C., Wu, J., Ciobanu, A., & Williams, K. (2014). CiteSeer x : A Scholarly Big Dataset. *Advances in Information Retrieval: 36th European Confer- Ence on IR Research, ECIR 2014, Amsterdam, The Nether- Lands, April 13-16, 2014*, *1*, 311–322.

Christen, P. (2012a). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, *24*(9), 1537–1555. https://doi.org/10.1109/TKDE.2011.127

Christen, P. (2012b). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. In *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. https://doi.org/10.1007/978-3-642-31164-2

Christen, P., & Goiser, K. (2007). Quality and complexity measures for data linkage and deduplication. *Studies in Computational Intelligence*, *43*, 127–151. https://doi.org/10.1007/978-3-540-44918-8_6

Churches, T., Christen, P., Lim, K., & Zhu, J. X. (2002). Preparation of name and address data for record linkage using hiddenMarkov models. *BMC Medical Informatics and Decision Making*, *2*, 1–16. https://doi.org/10.1186/1472-6947-2-1

Direktorat E-Government Kementerian Komunikasi dan Informatika RI. (2013). *Kerangka Kerja Interoperabilitas e-Government Indonesia*.

Draisbach, U., & Naumann, F. (2013). On choosing thresholds for duplicate detection. *Proceedings of the 18th International Conference on Information Quality, ICIQ 2013*.

Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., & Tang, N. (2018). Distributed representations of tuples for entity resolution. *Proceedings of the VLDB Endowment*, *11*(11), 1454–1467. https://doi.org/10.14778/3236187.3269461

Ektefa, M., Jabar, M. A., Sidi, F., Memar, S., Ibrahim, H., & Ramli, A. (2011). A threshold-based similarity measure for duplicate detection. *2011 IEEE Conference on Open Systems, ICOS 2011*, 37–41. https://doi.org/10.1109/ICOS.2011.6079233

Elezaj, O., & Tuxhari, G. (2017). Record Linkage using Probabilistic Methods and Data Mining Techniques. *Mediterranean Journal of Social Sciences*, *8*(3), 203–207. https://doi.org/10.5901/mjss.2017.v8n3p203

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 1–16. https://doi.org/10.1109/TKDE.2007.250581

Elziky, M. A., Ibrahim, D. M., & Sarhan, A. M. (2018). Improved Duplicate Record Detection Using ASCII Code Q-gram Indexing Technique. *Arabian Journal for Science and Engineering*, *43*(12), 7409–7420. https://doi.org/10.1007/s13369-018-3105-6

Graf, M., Sold, F., Laskowski, L., Panse, F., Papsdorf, F., Naumann, F., & Gremmelspacher, R. (2022). Frost: A Platform for Benchmarking and Exploring Data Matching Results. *Proceedings of the VLDB Endowment*, *15*(12), 3292–3305. https://doi.org/10.14778/3554821.3554823

Grannis, S. J., Overhage, J. M., & McDonald, C. J. (2002). Analysis of identifier performance using a deterministic linkage algorithm. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 305–309.

Gschwind, T., Miksovic, C., Minder, J., Mirylenka, K., & Scotton, P. (2019). Fast Record Linkage for Company Entities. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 623–630. https://doi.org/10.1109/BigData47090.2019.9006095

Gu, B., Li, Z., Zhang, X., Liu, A., Liu, G., Zheng, K., Zhao, L., & Zhou, X. (2017). The Interaction Between Schema Matching and Record Matching in Data Integration. *IEEE Transactions on Knowledge and Data Engineering*, *29*(1), 186–199. https://doi.org/10.1109/TKDE.2016.2611577

Gu, L., & Baxter, R. (2004). Adaptive filtering for efficient record linkage. *SIAM Proceedings Series*, 477–481. https://doi.org/10.1137/1.9781611972740.50

Gyawali, B., Anastasiou, L., & Knoth, P. (2020). Deduplication of scholarly documents using locality sensitive hashing and word embeddings. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, *May*, 901–910.

Hadzic, D., & Sarajlic, N. (2020). Methodology for fuzzy duplicate record identification based on the semantic-syntactic information of similarity. *Journal of King Saud University - Computer and Information Sciences*, *32*(1), 126–136. https://doi.org/10.1016/j.jksuci.2018.05.001

Haslhofer, B., & Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, *42*(2), 1–37. https://doi.org/10.1145/1667062.1667064

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280–299.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Data quality and record linkage techniques. In *Data Quality and Record Linkage Techniques*. https://doi.org/10.1007/0-387-69505-2

Irawan, D. E., Darujati, C., Soebandhi, S., Hayati, F., & Ayu Puspito Sari, D. (2019). How to Extend your Data Lifetime: Research Data Management in Indonesia's Context. *Advances in Social Science, Education and Humanities*

Research, *203*(Iclick 2018), 162–165. https://doi.org/10.2991/iclick-18.2019.33

Jiang, Y., Lin, C., Meng, W., Yu, C., Cohen, A. M., & Smalheiser, N. R. (2014). Rule-based deduplication of article records from bibliographic databases. *Database*, *2014*, 1–8. https://doi.org/10.1093/database/bat086

Joffe, E., Byrne, M. J., Reeder, P., Herskovic, J. R., Johnson, C. W., McCoy, A. B., Sittig, D. F., & Bernstam, E. V. (2014). A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. *Journal of the American Medical Informatics Association*, *21*(1), 97–104. https://doi.org/10.1136/amiajnl-2013-001744

Kilkenny, M. F., & Robinson, K. M. (2018). Data quality: "Garbage in – garbage out." *Health Information Management Journal*, *47*(3), 103–105. https://doi.org/10.1177/1833358318774357

Kim, H., Choo, C. Y., & Chen, S. S. (2010). Generating a meta-DL by federating search on OAI and non-OAI servers. *Journal of Intelligent Information Systems*, *34*(2), 177–191. https://doi.org/10.1007/s10844-009-0084-9

Koumarelas, I., Jiang, L., & Naumann, F. (2020). Data Preparation for Duplicate Detection. *Journal of Data and Information Quality*, *12*(3). https://doi.org/10.1145/3377878

Koumarelas, I., Papenbrock, T., & Naumann, F. (2020). MDedup: Duplicate detection with matching dependencies. *Proceedings of the VLDB Endowment*, *13*(5), 712–725. https://doi.org/10.14778/3377369.3377379

Lattar, H., Ben Salem, A., & Ben Ghezala, H. H. (2020). Duplicate record detection approach based on sentence embeddings. *Proceedings of the Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE*, *2020-Septe*, 269–274. https://doi.org/10.1109/WETICE49692.2020.00059

Lattar, H., & Salem, A. Ben. (2020). *Duplicate record detection approach based on sentence embeddings*. *April 2021*. https://doi.org/10.1109/WETICE49692.2020.00059

Lehti, P., & Fankhauser, P. (2006). Unsupervised duplicate detection using sample non-duplicates. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4244 LNCS*(July), 136–164. https://doi.org/10.1007/11890591_5

Ling, Y., An, Y., Liu, M., & Hu, X. (2013). An error detecting and tagging framework for reducing data entry errors in electronic medical records (EMR) system. *Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013*, 249–254.

https://doi.org/10.1109/BIBM.2013.6732498

Liu, S. N., Li, X. Y., Lu, C. J., Wen, Z. H., & Guo, X. F. (2011). An accuracy comparison between two methods of double data entry in Chinese medicine research. *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2011*, *1*, 749–751. https://doi.org/10.1109/BIBMW.2011.6112464

López-Cuadrado, J. L., González-Carrasco, I., López-Hernández, J. L., Martínez-Fernández, P., & Martínez-Fernández, J. L. (2020). Automatic learning framework for pharmaceutical record matching. *IEEE Access*, *8*, 171754–171770. https://doi.org/10.1109/ACCESS.2020.3024558

Lukman, L., Dimyati, M., Rianto, Y., Subroto, I. M. I., Sutikno, T., Hidayat, D. S., Nadhiroh, I. M., Stiawan, D., Haviana, S. F. C., Heryanto, A., & Yuliansyah, H. (2018). Proposal of the S-score for measuring the performance of researchers, institutions, and journals in Indonesia. *Science Editing*, *5*(2), 135–141. https://doi.org/10.6087/KCSE.138

Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age. *Portal: Libraries and the Academy*, *3*(2), 327–336. https://doi.org/10.1353/pla.2003.0039

Maidasani, H., Namata, G., Huang, B., & Getoor, L. (2012). Entity Resolution Evaluation Measures. *University of Maryland, Tech. Rep.*, 1–24. http://www.cs.umd.edu/Honors/reports/hitesh.pdf

Minton, S. N., Nanjo, C., Knoblock, C. A., Michalowski, M., & Michelson, M. (2005). A heterogeneous field matching method for record linkage. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 314–321. https://doi.org/10.1109/ICDM.2005.7

Mishra, S., Mondal, S., & Saha, S. (2013). Entity matching technique for bibliographic database. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8056 LNCS*(PART 2), 34–41. https://doi.org/10.1007/978-3-642-40173-2_5

Mishra, S., Saha, S., & Mondal, S. (2016). An automatic framework for entity matching in bibliographic databases. *2016 IEEE Congress on Evolutionary Computation, CEC 2016*, 271–278. https://doi.org/10.1109/CEC.2016.7743805

Moretti, A., & Shlomo, N. (2023). Improving Probabilistic Record Linkage Using Statistical Prediction Models. *International Statistical Review*, *91*(3), 368–394. https://doi.org/10.1111/insr.12535

Murray, J. S. (2015). Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering. *Journal of Privacy and Confidentiality*, *7*(1), 1–22. https://doi.org/10.29012/jpc.v7i1.643

Naumann, F., & Herschel, M. (2010). An Introduction to Duplicate Detection. In *Synthesis Lectures on Data Management* (Vol. 2, Issue 1). https://doi.org/10.2200/s00262ed1v01y201003dtm003

Paganelli, M., Del Buono, F., Guerra, F., Pevarello, M., & Vincini, M. (2021). Automated machine learning for entity matching tasks. *Advances in Database Technology - EDBT*, *2021-March*, 325–330. https://doi.org/10.5441/002/edbt.2021.29

Paganelli, M., Guerra, F., Sottovia, P., & Velegrakis, Y. (2019). TuneR: Fine tuning of rule-based entity matchers. *International Conference on Information and Knowledge Management, Proceedings*, 2945–2948. https://doi.org/10.1145/3357384.3357854

Panse, F., & Naumann, F. (2021). Evaluation of duplicate detection algorithms: From quality measures to test data generation. *Proceedings - International Conference on Data Engineering*, *2021-April*(5), 2373–2376. https://doi.org/10.1109/ICDE51399.2021.00269

Papadakis, G., Skoutas, D., Thanos, E., & Palpanas, T. (2020). Blocking and Filtering Techniques for Entity Resolution: A Survey. *ACM Computing Surveys*, *53*(2). https://doi.org/10.1145/3377455

Pei, S. O. A. (2008). A Comparative Study of Record Matching Algorithms. *European Master in Informatics (EuMI)*, *9001*(276905), 94. https://www.inf.ed.ac.uk/publications/thesis/online/IM080663.pdf%0Awww.irjet.net

Ranjana, G., & Thippeswamy, K. (2017). A Brief Survey on Record Linkage Techniques. *3rd National Conference On Emerging Trends In Computer Science & Engineering (NCETCSE-2017)*, 1–4.

Ren, J., Xia, F., Chen, X., Liu, J., Hou, M., Shehzad, A., Sultanova, N., & Kong, X. (2021). Matching Algorithms: Fundamentals, Applications and Challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *5*(3), 332–350. https://doi.org/10.1109/TETCI.2021.3067655

Samiei, A., & Naumann, F. (2016). Cluster-Based Sorted Neighborhood for Efficient Duplicate Detection. *IEEE International Conference on Data Mining Workshops, ICDMW*, *0*(December), 202–209. https://doi.org/10.1109/ICDMW.2016.0036

Sayers, A., Ben-Shlomo, Y., Blom, A. W., & Steele, F. (2016). Probabilistic record linkage. *International Journal of Epidemiology*, *45*(3), 954–964. https://doi.org/10.1093/ije/dyv322

Sefid, A., Wu, J., C.Ge, A., Zhao, J., Liu, L., Carage, C., Mitra, P., & Giles, C. L. (2014). 2019-Cleaning noisy and heterogeneous metadata for record linking across scholarly big datasets.pdf. *Advances in Information Retrieval: 36th European Confer- Ence on IR Research, ECIR 2014, Amsterdam, The*

Nether- Lands, April 13-16, 2014. Proceedings, 311–322.

Sefid, A., Wu, J., Ge, A. C., Zhao, J., Liu, L., Caragea, C., Mitra, P., & Lee Giles, C. (2019a). Cleaning noisy and heterogeneous metadata for record linking across scholarly big datasets. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 9601–9606. https://doi.org/10.1609/aaai.v33i01.33019601

Sefid, A., Wu, J., Ge, A. C., Zhao, J., Liu, L., Caragea, C., Mitra, P., & Lee Giles, C. (2019b). Cleaning noisy and heterogeneous metadata for record linking across scholarly big datasets. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 9601–9606. https://doi.org/10.1609/aaai.v33i01.33019601

Sinta Wahyuni, N. M., & Sanjaya ER, N. A. (2021). Rule-based Named Entity Recognition (NER) to Determine Time Expression for Balinese Text Document. *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, *9*(4), 555. https://doi.org/10.24843/jlk.2021.v09.i04.p14

Song, S., & Chen, L. (2009). Discovering matching dependencies. *International Conference on Information and Knowledge Management, Proceedings*, 1421–1424. https://doi.org/10.1145/1645953.1646135

Steorts, R. C., Ventura, S. L., Sadinle, M., & Fienberg, S. E. (2014). A comparison of blocking methods for record linkage. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8744*, 253–268. https://doi.org/10.1007/978-3-319-11257-2_20

Subramaniyaswamy, V., & Pandian, S. C. (2012). A complete survey of duplicate record detection using data mining techniques. *Information Technology Journal*, *11*(8), 941–945. https://doi.org/10.3923/itj.2012.941.945

Subroto, I. M. I., Sutikno, T., & Stiawan, D. (2014). The architecture of indonesian publication index: A major indonesian academic database. *Telkomnika (Telecommunication Computing Electronics and Control)*, *12*(1), 1–5. https://doi.org/10.12928/TELKOMNIKA.v12i1.1790

Thirumuruganathan, S., Li, H., Tang, N., Ouzzani, M., Govind, Y., Paulsen, D., Fung, G., & Doan, A. (2021). Deep learning for blocking in entity matching: A design space exploration. *Proceedings of the VLDB Endowment*, *14*(11), 2459–2472. https://doi.org/10.14778/3476249.3476294

Trippel, T., & Zinn, C. (2021). Lessons learned: on the challenges of migrating a research data repository from a research institution to a university library. *Language Resources and Evaluation*, *55*(1), 191–207.

https://doi.org/10.1007/s10579-019-09474-4

Valstar, N., Frasincar, F., Brauwers, G., Valstar, N., Frasincar, F., & Brauwers, G. (2021). APFA : Automated Product Feature Alignment for Duplicate Detection. *Expert Systems With Applications*, 114759. https://doi.org/10.1016/j.eswa.2021.114759

Vogel, T., Heise, A., Draisbach, U., Lange, D., & Naumann, F. (2014). Reach for gold: An annealing standard to evaluate duplicate detection results. *Journal of Data and Information Quality*, *5*(1–2). https://doi.org/10.1145/2629687

Wandhekar, V., & Mohanpurkar, A. (2015). Validation of Deduplication in Data using Similarity Measure. *International Journal of Computer Applications*, *116*(21), 18–22. https://doi.org/10.5120/20460-2819

Wang, Y., Zhang, H., Li, Y., Wang, D., Ma, Y., Zhou, T., & Lu, J. (2016). A data cleaning method for citeseer dataset. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10041 LNCS*(November), 35–49. https://doi.org/10.1007/978-3-319-48740-3_3

Warke, Y. (2017). Suffix array blocking for efficient record linkage and de-duplication in sliding window fashion. *Advances in Intelligent Systems and Computing*, *468*, 57–65. https://doi.org/10.1007/978-981-10-1675-2_7

Wei, M., Sung, A. H., & Cather, M. E. (2006). Improving database quality through eliminating duplicate records. *Data Science Journal*, *5*(December 2015), 127–142. https://doi.org/10.2481/dsj.5.127

Wu, J., Sefid, A., Ge, A. C., & Giles, C. L. (2017). A supervised learning approach to entity matching between scholarly big datasets. *Proceedings of the Knowledge Capture Conference, K-CAP 2017*, *May*, 2–6. https://doi.org/10.1145/3148011.3154470

Yang, Y., Sun, Y., Tang, J., Ma, B., & Li, J. (2015). Entity Matching across Heterogeneous Sources Categories and Subject Descriptors. *Kdd*.

Ye, Y., Zhong, B., Srimal, S., Alsarkhi, A., & Talburt, J. (2018). A study on the impact of missing values in probabilistic matching. *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018*, 158–163. https://doi.org/10.1109/CSCI46756.2018.00038

Zingmond, D. S., Ye, Z., Ettner, S. L., & Liu, H. (2004). Linking hospital discharge and death records — accuracy and sources of bias. *Journal of Clinical Epidemiology*, *57*, 21–29. https://doi.org/10.1016/S0895-4356(03)00250-6

Zuccala, A., & Cornacchia, R. (2016). Data matching, integration, and interoperability for a metric assessment of monographs. *Scientometrics*, *108*(1), 465–484. https://doi.org/10.1007/s11192-016-1911-8