# RESOLUSI ENTITAS DALAM JEJARING SOSIAL MENGGUNAKAN METODE PROPAGASI ATRIBUT YANG TIDAK IDENTIK

**Yesi Novia[1)], Arif Djunaidy[2)]**
[1]Prodi Sistem Informasi, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo, Surabaya, Jawa Timur, 60111
E-mail : novia_yesi@yahoo.com[1)]

***Abstract***

*An entity has its own character. And each character will be different according to the context. Needed a way, in order to identify character entities spread across a wide range of sources, that belong to the same entity. This method is known as Entity Resolution. Along with the development of social media, entity resolution research leads to the integration of two accounts, from two different social network but are owned by the same person.*

*From both these accounts, accounts similarities identified through profiles and network friendships. The approach used is attribute matching to match the profile data and graph matching to match friendships network.*

*Differs from previous studies, this study uses data with unidentical identifier (unidentical surface form) as a dataset. In addition, datasets are taken from Facebook and Google+, where users number were much different. So there are some accounts, has a disproportionate number of nodes. In order to be well compared, it is proposed to add methods Friend Recommender for the two graphs. the highest level accuracy using profile data and network of friends is 66% with average of 50%. However, the accuracy rises to 83% with average of 60% when using personal information and friend networks with Friend Recommender addition.*

***Abstrak***

*Sebuah entitas mempunyai karakternya sendiri. Dan setiap karakter akan berbeda sesuai dengan konteksnya. Untuk itu diperlukan sebuah cara, agar karakter entitas yang tersebar di berbagai sumber, dapat dikenali sebagai milik entitas yang sama. Cara ini dikenal dengan Entity Resolution (Resolusi Entitas). Seiring dengan berkembangnya media sosial, penelitian resolusi entitas mengarah kepada integrasi dua akun dari dua jejaring sosial yang berbeda tapi dimiliki oleh orang yang sama.*

*Dari kedua akun tersebut, kemiripan akun diidentifikasi melalui data profil dan jaringan pertemanannya. Pendekatan yang digunakan adalah pemadanan atribut (attribute matching) untuk memadankan data profil dan pemadanan graph (graph matching) untuk memadankan jaringan pertemanan.*

*Berbeda dari penelitian sebelumnya, penelitian ini menggunakan data pengenal yang tidak identik (unidentical surface form) sebagai dataset. Sehingga untuk melakukan pemadanan graf diperlukan suatu cara, agar node yang berbeda namun merujuk kepada entitas yang sama dapat dikenali. Selain itu, dataset diambil dari Facebook dan Google+, dimana jumlah penggunanya berbeda. Sehingga ada beberapa akun, mempunyai jumlah node yang tidak seimbang. Agar kedua node dari kedua graf yang dihasilkan dari dua jejaring sosial yang berbeda, dapat dibandingkan dengan baik, maka diusulkan penambahan metode Friend Recommender. Nilai kemiripan gabungan dari profil dan jejaring pertemanan mempunyai tingkat akurasi tertinggi rata-rata berturut-turut sebesar 66% dan 50%. Tetapi, jika jejaring pertemanan diperluas dengan penambahan Friend Recommender, maka diperoleh tingkat akurasi tertinggi dan rata-rata berturut-turut sebesar 83% dan 60%.*

***Kata kunci:*** *Resolusi entitas, attribute matching, graph matching*

## *PENDAHULUAN*

Resolusi entitas merupakan suatu cara untuk menentukan apakah dua reference merujuk kepada satu entitas yang sama di dunia nyata[1]. Dengan berkembangnya media sosial sebagai wadah pengguna untuk berinteraksi dengan orang lain, semakin banyak pula pengguna yang menggunakan lebih dari satu media sosial.

Sehingga dimungkinkan seseorang mempunyai akun di masing-masing media sosial yang berbeda.

Menyatukan profil user yang tersebar di berbagai jejaring sosial bertujuan untuk membangun pandangan tunggal dan komprehensif terhadap data user [2], sehingga dihasilkan graf sosial yang lebih lengkap. Dengan pandangan tunggal dan komprehensif tersebut dihasilkan beberapa manfaat di beberapa bidang. Seperti di **bidang pemasaran**, para pemasar atau penjual dapat mengirimkan iklan berupa pesan kepada calon pelanggan potensial sekali saja, dan calon pelanggan tidak akan menerima pesan yang berulang-ulang dari jejaring sosial yang berbeda. Di **bidang bisnis**, graf sosial yang lebih lengkap dapat memberikan manfaat kepada pengembang dan pengusaha untuk memperluas lingkaran jaringan pertemanan dengan sistem rekomendasi teman yang efektif [3],[4], mendukung penggunaan fungsionalitas servis yang memungkinkan user untuk menyampaikan *feed*/pesan dari satu media sosial dan menyebarkannya media sosial lain seperti Twitterfeed [5], kesempatan untuk menawarkan bisnis[6], serta mempermudah *social media recruiting*, yaitu menyaring kandidat pegawai berdasarkan profil yang ditulisnya di media sosial[7]. Akun yang terintegrasi juga bermanfaat di **bidang hukum** seperti pencarian teroris[8] yang dapat digunakan sebagai alat bukti di pengadilan[9], dan mendeteksi pengguna jahat yang berkedok pengguna nyata [10].

Penelitian [6] melakukan pemadanan profil sederhana dan menambahkan jaringan untuk menentukan apakah dua akun dari dua jejaring sosial yang berbeda merujuk kepada orang yang sama di dunia nyata. Penelitian ini memperbaiki penelitian sebelumnya, yang hanya memperhatikan atribut profil saja, dengan menambahkan jaringan pertemanan dalam perhitungan tingkat kemiripan dua akun milik orang yang sama dari dua jejaring sosial yang berbeda. Masukan dari penelitian ini adalah dua profil dan daftar teman dari dua jejaring sosial Hyves dan LinkedIn. Kedua profil ini dibandingkan atributnya untuk mendapatkan nila keserupaan atribut dengan menggunakan *approximate string matcher*. Luaran dari penelitian ini adalah pasangan profil yang dianggap yang paling cocok karena mempunyai nilai kemiripan gabungan yang paling tinggi.

Dalam penelitian [11] melakukan penelitian yang berawal dari kesulitan untuk melakukan pencarian orang karena banyaknya profil yang terdapat di sosial media dan kecenderungan orang untuk mendaftar di lebih dari satu sosial media. Tujuan dari penelitian ini adalah memfasilitasi pencarian orang di beberapa sosial media dan menggabungkan profil yang duplikat menjadi satu dengan pendekatan *machine learning*. Masukan yang digunakan dalam penelitian ini adalah dua profil yang direpresentasikan sebagai vektor. Untuk membandingkan setiap *field* dari dua profil maka digunakan beberapa fungsi matching seperti *exact matching*, *partial matching*, dan penelitian ini merancang Algoritma MN (MatchName) untuk membandingkan dua nama user. Untuk pengklasifikasi, dibandingkan 4 teknik yaitu Simple Weights, MLP, LogitBoost dan AdaBoost. Penelitian dilakukan dengan menggunakan data dari hasil *crawl* terhadap dua media sosial yaitu Facebook dan StudyVZ. Luaran dari penelitian ini adalah semua profil yang teridentifikasi identik.

Penelitian [12] melakukan pemadanan profil dengan membangun sebuah framework yang bisa memperhitungkan semua atribut profil dari user. User dapat memberikan bobot terhadap atribut tertentu untuk menunjukkan pentingnya sebuah atribut. Masukan dari sistem yang dibangun adalah 2 profil user lengkap dengan IFP (*Inverse Function Property*).

Penelitian dari [13] juga menggunakan metode *supervised* untuk menyelesaikan permasalahan pemadanan profil antar dua media sosial. Algoritma dibangun dengan memanfaatkan teknik *machine learning* dan 27 fitur yang diekstrak dari profil. Inputan dari penelitian ini adalah dua profil user dari dua media sosial yang berbeda. Peneliti membangun model dengan me;akukan uji coba fitur dengan 6 algoritma pembuat keputusan yang berbeda, yaitu AdaBoost, Rotation Forest, Random Forest, Logistics Model Tree, LogitBoost, dan Artficial Neural Network. Luaran dari algoritma ini adalah probabilitas apakah 2 profil pengguna yang menjadi masukan merupakan milik orang yang sama.

Penelitian [3] melakukan pemadanan profil dengan menggunakan fungsi string yang berbeda-beda. Untuk kemiripan jaringan pertemanan, dikembangkan algoritma yang didasarkan pada algoritma PageRank [14]. Inputan dari penelitian ini adalah dua profil dari dua jejaring sosial yang berbeda yaitu Facebook dan LinkedIn. Untuk mendapatkan kandidat yang akan dicocokkan, dicari profil yang mempunyai nama yang identik dengan profil pertama. Kemudian dilakukan perhitungan kemiripan untuk setiap nilai atribut. Penelitian

ini menggunakan fungsi yang berbeda-beda, yaitu Binary Similarity, Cosine Similarity dan SoftTFIDF untuk mendapatkan nilai kemiripan yang optimal. Atribut yang digunakan adalah tanggal lahir, jenis kelamin, lokasi, latar belakang pendidikan, pengalaman kerja dan minat. Kemiripan atribut $S_A(m,n)$ antara akun m si media sosial F dan akun n di media sosial Y dihitung berdasarkan rumus

$$S_A(m,n) = \sum_i w_i s(P^F_{m\backslash}.A_i, P^Y_n.A_i) \quad .....(1)$$

Dimana :

$s(P^F_m.A_i, P^Y_n.A_i)$ = kemiripan atribut $A_i$ antara akun m dan n

$w_i$ = adalah bobot kontribusi atribut ke-I terhadap jumlah keseluruhan.

Rumus (1) dapat langsung digunakan untuk menghitung kemiripan atribut dari profil kedua akun. Untuk bobot didapatkan dari jumlah nilai unik untuk setiap atribut dibagi dengan jumlah nilai atribut itu sendiri. Cara pembobotan yang sama juga digunakan oleh [15].

Penelitian [3] mempunyai beberapa kelemahan. Pertama, digunakan nama yang identik sebagai kunci untuk mencari kandidat dan mengukur kemiripan graf pertemanannya. Pada kenyataanya, cukup banyak akun yang menggunakan nama yang tidak identik. Kedua, rendahnya nilai F1-measure mungkin dikarenakan penggunaan algoritma yang didasarkan pada algoritma PageRank. Algoritma PageRank adalah algoritma yang memperhitungkan pentingnya (*importance*) sebuah website. Maka algoritma ini lebih cocok untuk digunakan untuk mengukur *centrality* sebuah node pada graf [16]. Ketiga, rendahnya nilai F1-measure mungkin karena nilai populasi pengguna dua jejaring sosial ini sangat berbeda jauh. Sehingga graf jaringan pertemanan menjadi tidak sebanding.

Karena itu permasalahan utama yang akan diteliti adalah ***bagaimana mengenali dua akun dengan pengenal yang tidak identik (unidentical surface form) adalah milik orang yang sama***. Adapun yang menjadi sub permasalahan dari permasalahan utama di penelitian ini adalah :

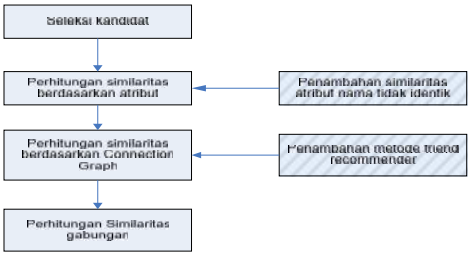
a. Bagaimana mengenali kandidat potensial, yang mungkin mempunyai dua akun di kedua jejaring sosial yang diteliti
b. Bagaimana menghitung keserupaan kedua profil berdasarkan atribut dan jaringan pertemanannya.
c. Bagaimana penambahan metode *Friend Recommender* dapat meningkatkan nilai kemiripan jaringan pertemanan.

## *METODOLOGI PENELITIAN*

Dalam bab ini dijelaskan mengenai metode penelitian yang diaplikasikan pada penelitian ini.

### Desain Model

Pada tahap ini dirancang dan dibuat model yang sesuai untuk menyelesaikan permasalahan yang dijelaskan sebelumnya. Tahapan yang terdapat pada bagian ini diperlihatkan gambar 1.




*Gambar 1. Desain Model*

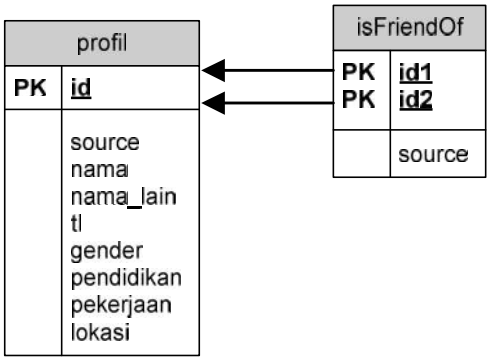
### Konstruksi Dataset

Data yang sudah dipraproses, akan disimpan di dalam database. Database yang digunakan terdiri dari dua tabel, yaitu tabel Profil dan tabel isFriendOf.

### Desain Algoritma

Tahapan pertama dari model yang akan dibangun adalah seleksi kandidat. Seleksi kandidat dilakukan untuk mendapatkan kandidat akun yang merupakan milik satu orang yang sama dari Facebook dan Google+ berdasarkan nama akun. Untuk setiap pasang nama akun, dihitung kemiripannya dengan fungsi string Jaro.

Tahapan kedua adalah perhitungan kemiripan atribut. Atribut yang diukur kemiripannya adalah nama, tanggal lahir, jenis kelamin, nama instansi pendidikan (pernah bersekolah di/alumni dari), nama instansi tempat bekerja (bekerja di), dan lokasi (daerah yang pernah ditinggali atau daerah tinggal saat ini). Tabel 1 menunjukkan fungsi string yang digunakan pada perhitungan kemiripan atribut.




*Gambar 2. Desain Dataset*

Tabel 1. Fungsi String yang digunakan untuk masing-masing atribut

| No | Atribut | Fungsi String |
|---|---|---|
| 1 | Nama, nama lain | Jaro, Soft TFIDF |
| 2 | Tanggal Lahir | Exact Match |
| 3 | Gender | Exact Match |
| 4 | Pendidikan | Soft TFIDF |
| 5 | Pekerjaan | Soft TFIDF |
| 6 | Lokasi | Soft TFIDF |

Pada penjumlahan, setiap atribut diberi bobot yang bernilai 0 hingga 1 dan jumlah seluruh bobot tersebut adalah 1. Menurut penelitian [13] dan [17], nama adalah pengidentifikasi yang unik dari sebuah akun. Sehingga nama sangat berkontribusi besar untuk menentukan sama atau tidaknya sebuah akun. Hampir senada dengan dua penelitian diatas [6] menyimpulkan bahwa nama, email dan tanggal lahir merupakan atribut paling penting sebagai penentu mirip atau tidaknya sebuah akun. Berdasarkan ketiga penelitian diatas, maka dalam peneltian ini tgl lahir diberi bobot lebih besar, dan atribut lain memperoleh bobot sama.

Tahapan ketiga adalah perhitungan kemiripan jaringan pertemanan. Untuk setiap kandidat yang didapat dari hasil seleksi kandidat dilakukan pemadanan graf teman. Hal ini didasarkan bahwa dua akun bisa dikatakan sama, jika akun tersebut terhubung dengan entitas yang sama. Kemiripan jaringan pertemanan dihitung dengan metode berdasarkan rumus :

$$Sim_{jaccard} = \frac{\left| N(v_i) \cap N(v_j) \right|}{\left| N(v_i) \cup N(v_j) \right|} \quad \ldots\ldots\ldots(2)$$

Tahapan keempat adalah perhitungan kemiripan jaringan pertemanan dengan penambahan *Friend Recommender.* Karena tidak seimbangnya jaringan pertemanan antara akun yang satu dengan yang lain, maka diusulkan penambahan nilai jaringan pertemanan dengan menambahkan usulan teman (*Friend Recommender*) untuk memastikan bahwa kedua akun kandidat dari Google+ dan Facebook adalah sama. Pada bagian ini diterapkan konsep *Friend Recommender* dengan didasarkan konsep *Trust Transitivity* yang merupakan bagian dari *Trust Propagation*[18]. Namun implementasinya hanya mengambil sebagian konsep dimana kemiripan antara user digantikan dengan kemiripan antara usulan teman di jaringan Google+ dengan teman Facebook kandidat. Dan jarak (*distance*) dipatok bahwa teman yang diusulkan didapat dari teman dari teman (*friends of friends*) dengan jumlah teman bersama minimal lima.

## *UJI COBA DAN ANALISIS*

Pada bagian ini akan dijelaskan rangkaian uji coba model yang telah dibangun. Selanjutnya juga diberikan analisis mengenai hasil uji coba yang dilakukan terhadap sampel data.

### Dataset

Model akan diuji dengan menggunakan data yang diambil dari Facebook dan Google+ dengan jumlah 1088 profil yang terdiri dari 855 data Facebook dan 236 data dari Google+. Dari keseluruhan profil terdapat 140 pasang kandidat yang cocok *(match)*. 140 pasangan kandidat tersebut dibagi ke dalam tiga kategori kandidat berdasarkan nama :

a. Kandidat dengan nama kedua akun yang sama persis dan atribut serupa.
   Contoh : Yesi Novia dan Yesi Novia
b. Kandidat dengan nama kedua akun dan atribut yang serupa. Diasumsikan, bahwa akun minimal mempunyai dua kata, dan dapat dikatakan serupa jika minimal terdapat satu kata yang sama.
   Contoh : Prisa Marga dan Prisa Marga Kusumantara.
c. Kandidat dengan nama akun yang berbeda, atribut serupa dan merupakan orang yang sama. Contoh: Riri Novalia dan Ieiemimykayla

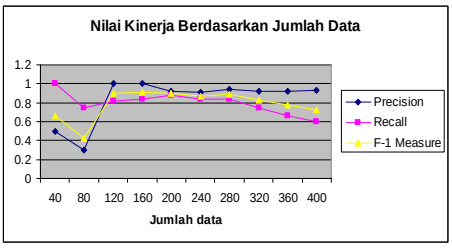
### Uji coba

Uji coba dilakukan dengan empat skenario yaitu:

a. uji coba terhadap sampel data.
b. uji coba terhadap sampel data dengan penambahan *Friend Recommender.*
c. akurasi terhadap sampel data.
d. uji coba terhadap kemampuan model untuk mengenali akun berdasarkan kategori kandidat.

Pengujian pertama dilakukan dengan mengambil sampel data secara acak dan sebanding antara kedua akun Facebook dan Google+. Data diambil dengan kelipatan 40 hingga jumlah sampel mencapai 400. Hasil pengujian diperlihatkan pada tabel 2.
Jumlah sampel adalah jumlah sampel yang diambil keseluruhan. Jumlah sampel 40 berarti sampel 20 diambil dari Google+ dan 20 diambil dari Facebook. Pasangan kandidat adalah berapa jumlah pasangan kandidat yang pada kenyataanya adalah pasangan akun milik satu orang. Sedangkan klasifikasi sesuai adalah jumlah pasangan akun yang dapat diprediksi dengan benar oleh model yang dibangun.

Dari tabel 2 dapat dilihat bahwa pada awal percobaan, ketika jumlah sampel 40 maka nilai *precision* ditemukan 0.5 dan *recall* 1, yang merupakan nilai tertinggi dari seluruh uji coba. Pada pengujian dengan sampel data 120, dan 160, nilai *recall* mengalami sedikit penurunan menjadi 0.81 dan 0.82, tetapi *precision* mengalami nilai tertinggi yaitu 1. Dan *F1-measure* juga mengalami nilai tertinggi yaitu sebesar 0.9. Kemudian seiring bertambahnya jumlah sampel, maka nilai nilai *precision* mengalami penurunan. Begitu juga dengan nilai *recall* dan *f-1 measure*. Nilai terendah *precision* terdapat pada jumlah sampel data 80. Nilai *recall* terendah terdapat pada uji coba dengan jumlah sampel data 400. Dan nilai *F-1 measure* terendah terdapat pada saat sampel berjumlah 80. Grafik *F-1 measure* dapat dilihat pada gambar 6.




*Gambar 6. Nilai kinerja berdasarkan jumlah data*

**Hasil Uji Coba terhadap jumlah sampel data dengan penambahan Friend Recommender**

Pada penelitian Veldman, (2009) ditemukan bahwa rata-rata jumlah jaringan pertemanan antara dua akun dari dua jejaring sosial berlainan, berbeda sangat jauh. Guna mensiasati hal tersebut, maka pada penelitian ini ditambahkan *Friend Recommender* untuk menambah jumlah jaringan yang timpang. Hasil Uji coba dengan penambahan Friend Recommender diperlihatkan oleh tabel 3. Pada tabel tersebut, terdapat 4 kolom yang menjelaskan hasil pemetaan akun sebelum dan sesudah penambahan *Friend Recommender*. SS (sama-sama) dimana hasil perhitungan menunjukkan bahwa klasifikasi kedua akun dengan jaringan asli adalah sama dan tetap sama setelah ditambahkan *Friend Recommender*. BB adalah beda-beda, artinya dengan jaringan asli kedua akun adalah akun yang berbeda dan tetap berbeda setelah ditambahkan dengan *Friend Recommender*. SB adalah sama beda artinya dengan perhitungan jaringan asli kedua akun adalah sama dan menjadi berbeda setelah ditambahkan *Friend Recommender*. BS adalah beda-sama artinya sebuah akun diklasifikasikan berbeda dengan perhitungan jaringan asli dan menjadi sama setelah ditambahkan *Friend Recommender.*

*Tabel 2. Hasil perhitungan model terhadap jumlah sampel*

| NO | Jumlah Sampel | Pasangan Kandidat | Klasifikasi sesuai | P | R | F1 |
|---|---|---|---|---|---|---|
| 1 | 40 | 4 | 2 | 0.5 | 1 | 0.666666667 |
| 2 | 80 | 8 | 3 | 0.3 | 0.75 | 0.428571429 |
| 3 | 120 | 19 | 11 | 1 | 0.81818 | 0.9 |
| 4 | 160 | 27 | 15 | 1 | 0.83333 | 0.909090909 |
| 5 | 200 | 34 | 21 | 0.91667 | 0.88 | 0.897959184 |
| 6 | 240 | 44 | 29 | 0.90909 | 0.83333 | 0.869565217 |
| 7 | 280 | 48 | 33 | 0.94595 | 0.83333 | 0.886075949 |
| 8 | 320 | 51 | 34 | 0.92308 | 0.75 | 0.827586207 |
| 9 | 360 | 52 | 36 | 0.92308 | 0.66667 | 0.774193548 |
| 10 | 400 | 60 | 43 | 0.92857 | 0.6 | 0.728971963 |

Tabel 3. Hasil pemetaan pada akun setelah ditambahkan *Friend Recommender*

| Jumlah Sampel | SS | BB | SB | BS |
|---|---|---|---|---|
| 40 | 0 | 0 | 0 | 0 |
| 80 | 7 | 7 | 0 | 0 |
| 120 | 9 | 14 | 0 | 2 |
| 160 | 15 | 22 | 0 | 3 |
| 200 | 24 | 24 | 0 | 4 |
| 240 | 32 | 33 | 0 | 4 |
| 280 | 37 | 27 | 0 | 11 |
| 360 | 43 | 44 | 0 | 11 |
| 400 | 45 | 49 | 0 | 11 |

**Mengukur Akurasi Terhadap Jumlah Data**

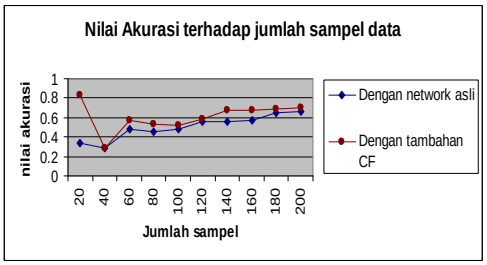
Akurasi menentukan seberapa banyak jumlah data benar yang dideteksi oleh model. Tabel 4 menunjukkan hasil perhitungan akurasi terhadap jumlah sampel data pada saat uji coba, dengan *Friend Recommender* dan tanpa *Friend Recommender*.

Tabel 4. Nilai Akurasi

| Sampel | Akurasi tanpa FR | Akurasi FR |
|---|---|---|
| 40 | 0.333333333 | 0.833333333 |
| 80 | 0.285714286 | 0.285714286 |
| 120 | 0.47826087 | 0.565217391 |
| 160 | 0.45 | 0.538461538 |
| 200 | 0.480769231 | 0.519230769 |
| 240 | 0.553846154 | 0.584615385 |
| 280 | 0.56 | 0.675675676 |
| 320 | 0.567901235 | 0.680555556 |
| 360 | 0.642857143 | 0.693181818 |
| 400 | 0.663265306 | 0.705128205 |

Dari tabel dapat dilihat bahwa nilai akurasi cenderung mengalami kenaikan baik dengan Friend Recommender atau tanpa Friend Recommender. Akan tetapi nilai akurasi dengan Friend Recommender sedikit lebih tinggi dibandingkan tanpa Friend Recommender. Nilai Akurasi tertinggi didapatkan pada saat jumlah sampel 400, dengan nilai 0.66 untuk profil tanpa Friend Recommender dan 0.83 untuk profil yang ditambahkan dengan Friend Recommender

Pada gambar 7 dapat dilihat bahwa profil dengan penambahan *Friend Recommender* nilai akurasinya sebikit lebih tinggi dibandingkan dengan profil tanpa *Friend Recommender (FR)* atau *Collaborative Filtering (CF).*




*Gambar 7. Nilai akurasi*

**Kemampuan Model Mendeteksi Kandidat berdasarkan kategori**

Pada bagian awal uji coba dijelaskan bahwa data dibagi berdasarkan tiga kategori. Berdasarkan kategori tersebut dilakukan uji coba, apakah model dapat mendeteksi akun yang merupakan milik orang yang sama. Tabel 5 menunjukkan hasil uji coba berdasarkan tiga kategori tersebut.

Tabel 5(a) Uji coba terhadap akun dengan nama identik

| Jumlah sampel | Nama Identik | Klasifikasi benar |
|---|---|---|
| 40 | 4 | 2 |
| 80 | 7 | 3 |
| 120 | 15 | 7 |
| 160 | 21 | 10 |
| 200 | 25 | 14 |
| 240 | 34 | 21 |
| 280 | 37 | 24 |
| 320 | 39 | 25 |
| 360 | 40 | 26 |
| 400 | 45 | 31 |

Tabel 5. (b) Uji coba terhadap akun dengan nama serupa

| Jumlah sampel | Nama Serupa | Klasifikasi benar |
|---|---|---|
| 40 | 0 | 0 |
| 80 | 1 | 0 |
| 120 | 4 | 3 |
| 160 | 6 | 5 |
| 200 | 9 | 7 |
| 240 | 10 | 8 |
| 280 | 11 | 9 |
| 320 | 11 | 9 |
| 360 | 12 | 10 |
| 400 | 13 | 11 |

Tabel 5(c) Uji coba terhadap akun dengan nama berbeda

| Jumlah sampel | Nama beda | Klasifikasi benar |
|---|---|---|
| 40 | 0 | 0 |
| 80 | 0 | 0 |
| 120 | 0 | 0 |
| 160 | 0 | 0 |
| 200 | 0 | 0 |
| 240 | 0 | 0 |
| 280 | 0 | 0 |
| 320 | 0 | 0 |
| 360 | 0 | 0 |
| 400 | 1 | 1 |

Dari perhitungan tabel 5 a,b,c didapatkan bahwa model yang dibangun pada penelitian ini dapat mendeteksi kandidat akun dengan nama identik, nama serupa dan nama berbeda. Nama identik dan nama serupa didapat dari membandingkan nama kedua akun. Sedangkan nama berbeda didapat dari membandingkan atribut nama lain dengan nama akun, sehingga didapatkan akun yang sama. Karena yang mengisi atribut nama lain tidak banyak, maka data untuk akun dengan nama berbeda pun tidak banyak.

**Analisis Hasil Uji Coba**

Setelah dilakukan uji coba terhadap sampel data, maka dapat disimpulkan bahwa semakin banyak sampel data, maka *precision recall* akan semakin menurun karena semakin banyaknya jumlah kemungkinan pasangan data. Akibatnya nilai F-1 measure juga semakin menurun.

Terdapat beberapa kesalahan pada saat klasifikasi data. Untuk nama yang sama, diprediksi sebagai orang berbeda. Tetapi pada kenyataanya akun tersebut adalah milik orang yang sama. Kesalahan ini terjadi karena akun tersebut tidak diisi dengan lengkap, dan jaringan pertemanannya pun kurang. Dari hasil perhitungan berdasarkan tabel 5, dari sejumlah pasangan akun milik orang yang sama pada sampel, model dapat memprediksi sekitar 60 % akun dengan benar.

Penggunaan fungsi string Jaro pada waktu seleksi kandidat tidak terlalu ketat menyaring calon kandidat. Dengan fungsi string Jaro, nama yang berbeda satu dua huruf, nilai kemiripannya tinggi. Untuk nama yang berbeda urutan, nilai kemiripannya rendah. Sehingga akan lebih baik jika digunakan fungsi Soft TF-IDF, guna menyaring calon kandidat dengan baik. Kelemahannya, fungsi simialritas Soft TF-IDF mempunyai waktu proses yang lama, terutama untuk pencocokan nama akun satu persatu.
Hasil uji coba dengan penambahan *Friend Recommender* dapat sedikit meningkatkan *precision recall* dibandingkan jaringan asli. Sehingga nilai F-1 measure dapat lebih tinggi dibandingkan dengan perhitungan jaringan pertemanan asli. Terdapat kesalahan di satu akun, dimana pada kenyataanya akun milik orang yang berbeda diklasifikasikan sama setelah *Friend Recommender*. Hal ini kemungkinan terjadi karena dua akun milik dua orang yang berasal dari lingkungan yang sama. Misal, sama-sama bersekolah di ITS dengan jurusan yang sama. Sehingga kemungkinan teman antara kedua akun kebanyakan sama, dan ketika ditambahkan *Friend Recommender* nilai kemiripan kedua akun menjadi sama. Namun kejadian seperti ini tidak banyak. Dari awal percobaan, hingga menggunakan 400 sampel, hanya terdapat satu akun dengan kesalahan seperti ini.

Hasil uji coba terhadap akurasi berada di atas 0.5. Hasil pengujian terhadap sampel tanpa *Friend Recommender*, nilai akhirnya hingga 0.66. sedangkan dengan penambahan *Friend Recommender*, nilai akhirnya sekitar 0.7. Dapat disimpulkan bahwa penambahan *Friend Recommender* dapat menambah nilai keserupaan untuk profil dengan jumlah jaringan pertemanan antara Facebook dan Google + tidak begitu berbeda. Sehingga *Friend Recommender* tidak akan berpengaruh banyak terhadap jaringan yang jumlah teman antar Facebook dan Google + berbeda jauh. Hal ini mungkin disebabkan karena batasan yang digunakan untuk memilih teman yang diusulkan terlalu ketat. Teman yang diusulkan harus mempunyai teman bersama sekitar lima orang dengan kandidat dan akun padanannya juga merupakan teman dari kandidat di jejaring sosial lain. Karena jumlah akun yang mempunyai padanan tidak banyak, maka jumlah teman usulan yang dapat ditambahkan pun tidak banyak. Penambahan *Friend Recommender*, dapat menaikkan nilai kemiripan jejaring pertamenan rata-rata sekitar 2.92 %.

Pada awal percobaan, nilai akurasi meroket tajam, dari 0.8333 menjadi 0.285714. Hal ini dimungkinkan karena pada awal percobaan tidak terdapat akun dengan nilai FP (*false positive*), sehingga nilai *Precision Recall* menjadi bagus. Tapi pada percobaan berikutnya, terdapat 4 data dengan nilai *False Positive*, artinya hasil prediksi menyatakan bahwa keempat data tersebut adalah orang yang sama, sedangkan pada kenyataanya, kedua akun adalah milik orang yang berbeda. Hal inilah yang mengakibatkan nilai akurasi menjadi turun.

Dari hasil pengujian terhadap sampel data, dapat disimpulkan bahwa model dapat mendeteksi akun dengan nama identik, serupa dan berbeda.

## *SIMPULAN dan SARAN*

Pada penelitian ini, untuk mengenali kandidat potensial dari dua akun yang berbeda adalah dengan membandingkan nama setiap akun Facebook dan Google+ dan menghitung keimiripannya.

Untuk memastikan bahwa kedua akun dari kandidat potensial adalah orang yang sama dihitung keserupaan profilnya berdasarkan atribut profil dan jaringan pertemanannya. Atribut yang dibandingkan adalah gender, tanggal lahir, pendidikan, pekerjaan dan lokasi. Untuk setiap atribut dari akun Facebook dibandingkan dengan setiap atribut dari akun Google + dan dihitung keserupaanya.
Guna meningkatkan nilai keserupaan jaringan, maka ditambahkan algoritma *Friend Recommender*. Dari hasil penelitian, penambahan algoritma *Friend Recommender* dapat menambah keserupaan jaringan sekitar 2.92%. Model dapat mendeteksi akun dengan nama identik, serupa, berbeda.

Pada penelitian selanjutnya, guna meningkatkan nilai keserupaan, pada atribut lokasi dapat dilakukan dengan membandingkan jarak kedua lokasi berdasarkan peta. Untuk itu dapat digunakan aplikasi Google Maps. Dengan tujuan yang sama, untuk atribut pendidikan dan pekerjaan perlu ditambahkan mekanisme mengenali singkatan dan akronim. Selain itu dapat ditambahkan juga pengetahuan atau semantik untuk nilai atribut yang berbahasa Inggris.

## *DAFTAR RUJUKAN*


[1] Talburt, Jhon R., 2011. *Entity Resolution and Information Quality*. Elsevier, USA

[2] Bartunov, Sergey, Korshunov, A., Park, S. T., Ryu, W., & Lee, H., 2012. *Joint link-attribute user identity resolution in online social networks.* Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis ACM.

[3] Niu, Lingfeng, Jianmin Wu, Yong Shi, 2011. *Entity resolution with attribute and connection graph.* In Data Mining Workshops (ICDMW), 2011 IEEEb/ 11th International Conference IEEE. (pp. 267-271).

[4] Motoyama, Marti, George Varghese., 2009. *I seek you: searching and matching individuals in social networks*. Proceedings of the eleventh international workshop on Web information and data management ACM (pp. 67-75).

[5] Jain, Paridhi, Ponnurangam Kamaguru, Anupam Joshi., 2013. *@I seek 'fb.me': Identifying Users across Multiple Online Social Networks*. Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee.

[6] Veldman, Irma., 2009. *Matching profiles from social network sites*. Master. University of Twente, Enschede.

[7] Robert Walters Company., _. *Using Scoial Media in the Recruitment Process*. Whitepaper, Insight Series, UK.

[8] Shresta, Amenda., 2013. *Visualization and Detection of Multiple Aliases in Social Media*. Master Thesis, Uppsala Universitet, Uppsala.

[9] Pabico, J.P., 2014. *An Analysis of Named Entity Disambiguation in Social Networks*. in Asia Pasific Journal of Multidisciplinary Research, Vol 2, No.4, hal 31-34.

[10] Goga, Oana., 2014. *Matching User Accounts Across Online Social Networks: Methods and Applications*. Dissertation. LIP6-Laboratoire d'Informatique de Paris 6, Université Pierre et Marie Curie, Paris, Perancis.

[11] Vosecky, Jan, Dan Hong, Vincent Y. Shen., 2009. *User identification across multiple social networks*. Networked Digital Technologies, 2009. NDT'09. First International Conference (pp. 360-365).

[12] Raad, Elie, Richard Chbeir, Albert Dipanda., 2010.*User profil matching in social networks.* In Network-Based Information Systems (NBiS), 2010 13th International Conference IEEE. (pp. 297-304).

[13] Peled, Olga, Michael Fire, Lior Rokach, Yuval Elovici., 2013. *Entity Matching in Online Social Networks*. In Social Computing (SocialCom), 2013 International Conference IEEE. (pp. 339-344).

[14] Brin, Sergey, Lawrence Page., 1998. *The anatomy of a large-scale hypertextual Web search engine*. Computer networks and ISDN systems 30.1 107-117.

[15] Soltani, Reza, 2013. *Identity Matchin in Social Media Platforms*. Thesis. York University, Toronto, Ontario, Canada.

[16] Zafarani, Reza, Mohammad Ali Abbasi, Huan Liu., 2014. *Social media mining: an introduction*. Cambridge University Press.

[17] Perito, Daniele., Claude Castellucia, Mohammed Ali Kaffar, Pere Manils.,

2011. *How unique and traceable are usernames?*. Privacy Enhancing Technologies. Springer Berlin Heidelberg.

[18] Josang, Audun, Stephen Marsh, Simon Pope., 2006. *Exploring Different Type of Trust Propagation*. Trust Management, Springer Berlin Heidelberg, 2006. 179-192.