

## ABSTRAK

Kelengkapan basis data alumni adalah hal yang penting bagi universitas. Hubungan timbal balik alumni dengan instansi adalah salah satu butir penting akreditasi universitas maupun akreditasi program studi. Namun, basis data alumni cenderung tidak relevan atau cacat, karena kemungkinan perubahan nomor telepon, alamat rumah, alamat *email*, dan lain-lain. Padahal informasi tersebut sangat menunjang komunikasi dengan instansi. Dan lagi, tingkat keberhasilan pengajaran suatu universitas juga ditentukan oleh sejalannya bidang kerja alumni dengan bidang ilmu yang telah ditempuh. Oleh karena itu, perlu ditemukan cara untuk dapat melengkapi basis data alumni dan mengetahui pekerjaan alumni dengan tepat. Pertama-tama mengumpulkan informasi yang tersebar di internet. Hal itu dilakukan dengan *scraping*, yakni mengekstrak dari mesin temu balik, seperti: Google. Tidak hanya itu, perlu dilakukan *data filtering*. Namun permasalahan muncul karena perkembangan dunia (hingga saat ini populasi dunia 2015 mencapai 7.3 miliar) dan perkembangan teknologi, mengakibatkan banyak orang yang mengacu pada nama yang sama. Dengan demikian, perlu dipikirkan cara untuk membedakan individu yang dicari dari milyaran individu hasil pencarian. Cara yang dilakukan adalah melalui metode klusterisasi (penerapan algoritma *Unsupervised Person Name Disambiguator*). Algoritma ini mencoba untuk membedakan individu yang dicari dari individu-individu lainnya. Tidak cukup puas dengan performansi algoritma *UPND*, maka dikembangkan pula algoritma *Reduce-UPND*. Tidak hanya itu, berbagai tahap data preprocessing perlu dilakukan demi meningkatkan performansi aplikasi ini. Terlebih untuk melihat seberapa besar pengaruh kelas kata terhadap hasil penebakan bidang kerja, maka hasil kluster diklasifikasikan menjadi 2 kelompok; kata benda dan kata kerja (acuan: KBBI). Tidak hanya data mengenai individu yang diolah, tetapi juga perlu didefinisikan pekerjaan. Setelah itu, dilakukan penebakan profesi. Penebakan yang dilakukan berdasarkan informasi individu *non-social media*. Eksperimen yang dilakukan dalam penelitian ini adalah *cross-validation (5-fold)* dan *hold training-test* (3 kombinasi 80%:20%, 70%:30%, dan 60%:40% antara *training* dan *testing*). Akurasi dari penebakan bidang kerja dan asal fakultas sebesar 90.91% (pada komposisi 80%:20%). Tidak cukup puas dengan akurasi tinggi dari prediksi bidang kerja dan asal fakultas, aplikasi ini juga membandingkan hasil dari kluster sosial media. Sosial media yang dipilih adalah LinkedIn, mengingat LinkedIn adalah sosial media untuk para profesional dan informasi dalam LinkedIn diisi sendiri oleh individu yang bersangkutan. Aplikasi ini pada akhirnya mengkombinasikan prediksi bidang kerja dan asal fakultas alumni serta ekstraksi informasi dari kluster sosial media LinkedIn, yaitu: pekerjaan sekarang, informasi pendidikan, dan informasi pekerjaan yang lampau.

Kata kunci: *virtual alumni tracer (VILTER)*, disambiguasi nama, prediksi bidang kerja dan asal fakultas, ekstraksi sosial media, LinkedIn

## **ABSTRACT**

*Completeness of alumni database is important for the university. Alumnus reciprocal relationship with the agency is one of the important points university and the study program accreditation. However, the database of alumni tends to be flawed, because the possibility of changing phone numbers, home address, email address, and others. Though the information is very supportive communication with agencies. Moreover, the success of a university is also determined by field of alumni profession in the field of science that has been taken. Therefore, it is necessary to find a way to be able to complete the database of alumni and know alumni profession appropriately. First collect the scattered information on the Internet. This was done by scraping, which is extracted from the retrieval engine, such as Google. Not only that, there should be a data filtering. However, problems arise because the development of the world (to this day the world population reached 7.3 billion in 2015) and the development of technology, resulting in a lot of people who refer to the same name. Thus, it should be considered a way to distinguish individuals who sought billions of individual search results. How that is done is through the clustering (Unsupervised Person Name Disambiguator algorithms). These algorithms try to distinguish individuals who sought from other individuals. Not quite satisfied with the performance of the algorithm UPND, the algorithm also developed by the Reduce-UPND. Not only that, stages of data preprocessing needs to be done to improve the performance of this application. Especially to see how big class influence the outcome word guessing areas of work, then the cluster results are classified into two groups; noun and verb (reference: KBBI). Not only data about the individual that is processed, but also need to be defined job. After that, guessing professions. Guessing that based on the non-social media information. Experiments were performed in this study is cross-validation (5-fold) and hold training-test (3 combination, such as 80%:20%, 70%:30%, and 60%:40% between training and testing). Accuracy of guessing the field work and origin of the faculty reached 90.91% (on a composition of 80%:20%). Not quite satisfied with the high accuracy of prediction of the field work and the origin of the faculty, this application also compares the results of social media cluster. Social media chosen is LinkedIn, considering that LinkedIn is a social media and information for professionals in the LinkedIn filled solely by the individual concerned. These applications in turn combines the prediction field of work and the origin of alumni and faculty of extracting information from social media LinkedIn clusters, namely: highlight job, educational information, and information about past jobs.*

*Keywords: virtual alumni tracer (VILTER), name disambiguation, job prediction, social media extraction, LinkedIn*

# DAFTAR ISI

LEMBAR PENGESAHAN .....	i
PRAKATA.....	ii
PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH .....	iv
PERNYATAAN PERSETUJUAN ORISINALITAS KARYA .....	v
ABSTRAK .....	vi
<i>ABSTRACT</i> .....	vii
DAFTAR ISI.....	viii
DAFTAR GAMBAR .....	xiii
DAFTAR TABEL.....	xv
DAFTAR SIMBOL.....	xvi
BAB I PENDAHULUAN .....	1
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah.....	5
1.3. Tujuan .....	6
1.4. Ruang Lingkup .....	6
1.5. Sistematika Penulisan .....	7
BAB II LANDASAN TEORI.....	9
2.1. <i>Web Search Engine</i> .....	9
2.1.1. <i>Crawling</i> dan <i>Indexing</i> .....	9
2.1.2. Mesin Pencari sebagai Pemberi Jawaban.....	10
2.2. Relevansi Tautan .....	11
2.3. Media Sosial .....	11
2.4. Ekspansi Kueri.....	13
2.5. Teknik <i>Scraping</i> untuk Ekstraksi Informasi .....	15

2.6. Algoritma <i>Unsupervised Person Name Disambiguator (UPND)</i> .....	16
BAB III ANALISIS DAN DESAIN .....	19
3.1. Analisis .....	19
3.1.1. Pengumpulan Data .....	25
3.1.2. Pengambilan Data dari Mesin Temu Balik .....	25
3.1.3. Supervisi Kueri .....	27
3.1.4. <i>Data Preprocessing</i> .....	27
3.1.5. Pembobotan <i>Term</i> .....	29
3.1.6. <i>Unsupervised Person Name Disambiguator</i> .....	30
3.1.7. Reduksi Halaman Temu Balik .....	31
3.1.8. Pembentukan Model <i>Naïve Bayes</i> .....	33
3.1.9. Pembentukan Model <i>Naïve Bayes</i> dan Klasifikasi Berdasarkan Kelas Kata.....	38
3.1.10. Kata-Kata Profesi dalam Hasil Kluster <i>UPND</i> .....	39
3.1.11. Peta Penelitian .....	39
3.2. Gambaran Keseluruhan.....	46
3.2.1. Persyaratan Antarmuka Eksternal .....	46
3.2.2. Antarmuka Perangkat Keras .....	46
3.2.3. Antarmuka Perangkat Lunak.....	47
3.2.4. Fitur-Fitur Perangkat Lunak.....	47
3.3. Desain Perangkat Lunak .....	60
3.3.1. Pemodelan Perangkat Lunak.....	60
3.3.2. Desain Penyimpanan Data .....	65
3.3.3. Rancangan Antarmuka .....	65
BAB IV PENGEMBANGAN PERANGKAT LUNAK .....	68
4.1. Implementasi Algoritma <i>UPND</i> dan <i>Red-UPND</i> .....	68

4.2. Implementasi <i>Class</i> .....	68
4.2.1. Implementasi <i>Class ClassProbability</i> .....	68
4.2.2. Implementasi <i>Class Cluster</i> .....	68
4.2.3. Implementasi <i>Class CompleteClustering</i> .....	69
4.2.4. Implementasi <i>Class DetailPredictingTuple</i> .....	69
4.2.5. Implementasi <i>Class IndonesianStemmer</i> .....	70
4.2.6. Implementasi <i>Class JobPredictingTuple</i> .....	71
4.2.7. Implementasi <i>Class JobPredictionHoldTrainTest</i> .....	72
4.2.8. Implementasi <i>Class KebiReader</i> .....	72
4.2.9. Implementasi <i>Class ModelTuple</i> .....	73
4.2.10. Implementasi <i>Class NameEmbedder</i> .....	73
4.2.11. Implementasi <i>Class NBTuple</i> .....	73
4.2.12. Implementasi <i>Class NGramResult</i> .....	74
4.2.13. Implementasi <i>Class SearchResultTuple</i> .....	74
4.2.14. Implementasi <i>Class TermTuple</i> .....	75
4.2.15. Implementasi <i>Class TrainingTuple</i> .....	75
4.2.16. Implementasi <i>Class Alumni</i> .....	76
4.2.17. Implementasi <i>Class AlumniDAO</i> .....	76
4.2.18. Implementasi <i>Interface IAlumni</i> .....	77
4.2.19. Implementasi <i>Interface ILinkedInEntity</i> .....	77
4.2.20. Implementasi <i>Class InfoEntity</i> .....	77
4.2.21. Implementasi <i>Interface IPrediksi</i> .....	78
4.2.22. Implementasi <i>Class LinkedInEntity</i> .....	78
4.2.23. Implementasi <i>Class LinkedInEntityDAO</i> .....	79
4.2.24. Implementasi <i>Class Prediksi</i> .....	80
4.2.25. Implementasi <i>Class PrediksiDAO</i> .....	80

4.2.26. Implementasi <i>Class Utility</i> .....	81
4.3. Implementasi Simpanan Data .....	81
4.3.1. Implementasi Tabel Alumni .....	81
4.3.2. Implementasi Tabel Prediksi .....	82
4.3.3. Implementasi Tabel LinkedIn.....	82
4.4. Implementasi Antar Muka .....	83
4.4.1. Implementasi Antar Muka <i>Form Job Predicting</i> .....	83
4.4.2. Implementasi Antar Muka <i>Web Halaman Utama</i> .....	83
4.4.3. Implementasi Antar Muka <i>Web Informasi Alumni</i> .....	84
4.4.4. Implementasi Antar Muka <i>Web Informasi Alumni Retrieved</i> dari LinkedIn .....	85
4.4.5. Implementasi Antar Muka <i>Web Rekomendasi Alumni</i> .....	85
<b>BAB V TESTING DAN EVALUASI SISTEM</b> .....	<b>86</b>
5.1. Korelasi Jumlah Halaman Temu Balik dan Kluster .....	86
5.2. Pengaruh Kluster terhadap Basis Data Alumni .....	88
5.3. Identifikasi Halaman Media Sosial.....	89
5.4. Reduksi Halaman Temu Balik.....	90
5.5. Pembentukan Model <i>Naïve Bayes</i> .....	92
5.6. <i>First Step Job Prediction</i> .....	94
5.7. Eksperimen <i>Cross-Validation</i> .....	94
5.8. Eksperimen <i>Hold Training-Test</i> .....	96
5.9. Pengaruh Teknik Penebakan dan Asal Fakultas .....	101
5.10. Pengujian Penebakan Bidang Kerja dan Asal Fakultas .....	102
<b>BAB VI SIMPULAN DAN SARAN</b> .....	<b>104</b>
6.1. Simpulan .....	104
6.2. Saran .....	106



LAMPIRAN A HASIL <i>SCRAPING</i> .....	xviii
LAMPIRAN B HASIL <i>CRAWLING</i> .....	xix
LAMPIRAN C HASIL <i>CLUSTERING</i> .....	xix
LAMPIRAN D HASIL <i>CO-OCCURRENCE</i> .....	xx
DAFTAR PUSTAKA .....	xxiv
CURRICULUM VITAE .....	xxvi



## DAFTAR GAMBAR

Gambar 2.1 Algoritma <i>UPND</i> (Delgado et al., 2014).....	18
Gambar 3.1 Algoritma <i>Calculate The Most Suitable Cluster</i> .....	22
Gambar 3.2 Algoritma <i>Calculate The Co-Occurrence</i> .....	22
Gambar 3.3 Visualisasi Hasil Analisis Penelitian.....	23
Gambar 3.4 Visualisasi Hasil Analisis Penelitian (cont'd).....	24
Gambar 3.5 Contoh Data.....	27
Gambar 3.6 Contoh Data Setelah Konversi Tahap I.....	28
Gambar 3.7 Contoh Data Setelah Konversi Tahap II .....	28
Gambar 3.8 Contoh Data Setelah Konversi Tahap III .....	29
Gambar 3.9 Algoritma <i>Red-UPND</i> .....	32
Gambar 3.10 Contoh Reduksi Algoritma <i>Red-UPND</i> .....	33
Gambar 3.11 <i>Flowchart</i> Peta Penelitian .....	44
Gambar 3.12 <i>Flowchart</i> Peta Penelitian (cont'd) .....	45
Gambar 3.13 <i>Use Case Diagram</i> Prediksi Bidang Kerja dan Asal Fakultas untuk Alumni .....	62
Gambar 3.14 <i>Class Diagram</i> Prediksi Bidang Kerja dan Asal Fakultas untuk Alumni .....	63
Gambar 3.15 <i>Sequence Diagram</i> Keseluruhan Sistem Prediksi Bidang Kerja dan Asal Fakultas untuk Alumni .....	64
Gambar 3.16 Desain Penyimpanan Data Prediksi Bidang Kerja dan Asal Fakultas untuk Alumni .....	65
Gambar 3.17 Rancangan <i>Form</i> Utama.....	66
Gambar 3.18 Rancangan <i>Web</i> Halaman Utama.....	66
Gambar 3.19 Rancangan <i>Web</i> Informasi Alumni .....	67
Gambar 4.1 <i>Class ClassProbability</i> .....	68
Gambar 4.2 <i>Class Cluster</i> .....	69
Gambar 4.3 <i>Class CompleteClustering</i> .....	69
Gambar 4.4 <i>Class DetailPredictingTuple</i> .....	70
Gambar 4.5 <i>Class IndonesianStemmer</i> .....	71
Gambar 4.6 <i>Class JobPredictingTuple</i> .....	72
Gambar 4.7 <i>Class JobPredictionHoldTrainTest</i> .....	72





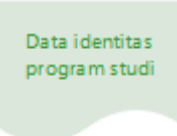



Gambar 4.8 Class <i>KebiReader</i> .....	72
Gambar 4.9 Class <i>ModelTuple</i> .....	73
Gambar 4.10 Class <i>NameEmbedder</i> .....	73
Gambar 4.11 Class <i>NBTuple</i> .....	74
Gambar 4.12 Class <i>NGramResult</i> .....	74
Gambar 4.13 Class <i>SearchResultTuple</i> .....	75
Gambar 4.14 Class <i>TermTuple</i> .....	75
Gambar 4.15 Class <i>TrainingTuple</i> .....	76
Gambar 4.16 Class <i>Alumni</i> .....	76
Gambar 4.17 Class <i>AlumniDAO</i> .....	77
Gambar 4.18 Interface <i>IAlumni</i> .....	77
Gambar 4.19 Interface <i>ILinkedInEntity</i> .....	77
Gambar 4.20 Class <i>InfoEntity</i> .....	78
Gambar 4.21 Interface <i>IPrediksi</i> .....	78
Gambar 4.22 Class <i>LinkedInEntity</i> .....	79
Gambar 4.23 Class <i>LinkedInEntityDAO</i> .....	80
Gambar 4.24 Class <i>Prediksi</i> .....	80
Gambar 4.25 Class <i>PrediksiDAO</i> .....	81
Gambar 4.26 Class <i>Utility</i> .....	81
Gambar 4.27 Implementasi Penyimpanan Data .....	81
Gambar 4.28 Antar Muka <i>Form Job Predicting</i> .....	83
Gambar 4.29 Antar Muka <i>Web Halaman Utama</i> .....	84
Gambar 4.30 Antar Muka <i>Web Informasi Alumni</i> .....	84
Gambar 4.31 Antar Muka <i>Web Informasi Alumni Retrieved</i> dari LinkedIn .....	85
Gambar 4.32 Antar Muka <i>Web Rekomendasi Alumni</i> .....	85
Gambar 5.1 Korelasi <i>Pearson</i> untuk Jumlah Rata-Rata dan Kluster .....	87
Gambar 5.2 Komposisi <i>5-Fold Cross-Validation</i> .....	95
Gambar 5.3 Performansi Eksperimen <i>Cross-Validation</i> .....	96
Gambar 5.4 Komposisi Keseluruhan Data Koleksi .....	97
Gambar 5.5 Komposisi Eksperimen <i>Hold Training-Test</i> (80%:20%) .....	98
Gambar 5.6 Komposisi Eksperimen <i>Hold Training-Test</i> (70%:30%) .....	99
Gambar 5.7 Performansi Eksperimen <i>Hold Training-Test</i> (60%:40%) .....	99

## DAFTAR TABEL


Tabel 3.1 Contoh Data Alumni .....	26
Tabel 3.2 Kueri yang Diujicobakan .....	26
Tabel 3.3 Metode I Pengelompokan Profesi berdasarkan Fakultas-Fakultas UKM .....	34
Tabel 3.4 Tabel Top-15 Sekolah Tinggi Indonesia.....	35
Tabel 3.5 Metode II Penentuan Jenis Profesi Umum (Utama) dari Top-15 Sekolah Tinggi .....	35
Tabel 3.6 Metode III Penentuan Jenis Profesi Umum dari Top-15 Sekolah Tinggi dan Pengetahuan Penulis.....	36
Tabel 3.7 Kelas Profesi Bidang Kerja.....	37
Tabel 4.1 Tabel Alumni .....	82
Tabel 4.2 Tabel Prediksi .....	82
Tabel 4.3 Tabel LinkedIn.....	82
Tabel 5.1 Korelasi Jumlah Halaman dan Kluster untuk Setiap Supervisi Kueri ..	86
Tabel 5.2 Pengaruh Kluster <i>UPND</i> terhadap Basis Data Alumni.....	88
Tabel 5.3 Sebaran Sosial Media Pada Tautan di Dalam Kluster .....	89
Tabel 5.4 Perbandingan Lama Waktu Eksekusi <i>UPND</i> dan <i>Red-UPND</i> .....	90
Tabel 5.5 Selisih Akurasi <i>Red-UPND</i> dan <i>UPND</i> .....	91
Tabel 5.6 Performansi Eksperimen <i>Cross-Validation</i> .....	95
Tabel 5.7 Akurasi Eksperimen <i>Hold Training-Test</i> .....	97
Tabel 5.8 Tabel Perbandingan Komposisi Data Testing 30% dan 40% .....	100
Tabel 5.9 Selisih <i>Data Testing</i> dan <i>Data Training</i> dari <i>Hold Train-Test</i> 60%:40% .....	101
Tabel 5.10 Pengaruh Metode Penebakan dan Asal Fakultas .....	101
Tabel 5.11 Tabel Pengujian Tebakan Bidang Kerja dan Asal Fakultas.....	102

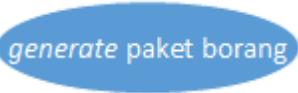
## DAFTAR SIMBOL

### 1. Flowchart

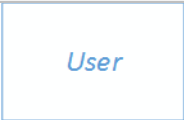



No.	Gambar	Nama Gambar	Deskripsi
1.		<i>Start/End state</i>	Menandai dimulai dan diakhirinya sebuah <i>flowchart</i>
2.		Kegiatan manual	Menunjukkan pekerjaan yang dilakukan dengan manual
3.		Dokumen	Menunjukkan dokumen
4.		<i>Decision</i>	Menyatakan kondisi dalam sebuah <i>flowchart</i>
5.		Simpanan <i>offline</i>	<i>File</i> non komputer yang diarsip baik terurut angka ( <i>numerical</i> ), huruf ( <i>alphabetical</i> ), dan tanggal ( <i>cronological</i> )
6.		Proses	Menunjukkan kegiatan proses dari operasi program komputer

### 2. Usecase

No.	Gambar	Nama Gambar	Deskripsi
1.		<i>System Boundary</i>	Untuk menggambarkan jangkauan sistem dan memberikan alternatif pilihan sistem

No.	Gambar	Nama Gambar	Deskripsi
2.		<i>Actor</i>	<i>Actor</i> mempresentasikan seseorang atau sesuatu yang berinteraksi dengan sistem
3.		<i>Communication</i>	Tujuan komunikasi adalah untuk memperlihatkan bahwa sebuah <i>actor</i> terlibat dalam <i>usecase</i>
4.		<i>Generalization</i>	Relasi antara dua <i>actor</i> atau dua <i>usecase</i> dimana salah satunya menurunkan, menambahkan atau <i>override</i> sifat dari yang lainnya
5.		<i>Usecase</i>	Gambaran fungsionalitas dari suatu sistem, sehingga pengguna dapat memahami guna dari sistem

### 3. ERD

No.	Gambar	Nama Gambar	Deskripsi
1.		<i>Entity</i>	Menyatakan sebuah obyek dalam sebuah ERD
2.		<i>Attribute</i>	Menyatakan elemen yang dimiliki obyek dalam sebuah ERD
3.		<i>Relationship connector</i>	Penghubung antar obyek, atribut, dan relasi dalam sebuah ERD
4.		<i>Relationship</i>	Menyatakan hubungan antar obyek dalam sebuah ERD