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

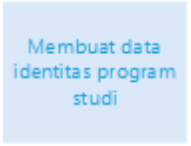
DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

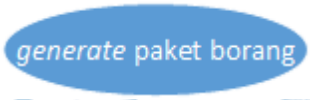
DAFTAR SIMBOL

1. *Flowchart*

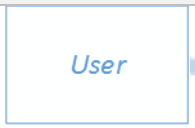
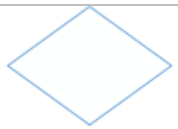
No.	Gambar	Nama Gambar	Deskripsi
1.		*Start/End state*	Menandai dimulai dan diakhirinya sebuah *flowchart*
2.		Kegiatan manual	Menunjukkan pekerjaan yang dilakukan dengan manual
3.		Dokumen	Menunjukkan dokumen
4.		*Decision*	Menyatakan kondisi dalam sebuah *flowchart*
5.		Simpanan *offline*	*File* non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

2. *Usecase*

No.	Gambar	Nama Gambar	Deskripsi
1.		*System Boundary*	Untuk menggambarkan jangkauan sistem dan memberikan alternatif pilihan sistem

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

3. ERD

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