ABSTRAK

Menemukan identitas dengan menggunakan nama adalah sesuatu yang sering kita lakukan setiap hari. Sebuah kecocokan mudah ditemukan bila nama yang dicari tersebut sama persis seperti yang tercatat di database. Namun, sebuah nama seringkali ambigu, tidak unik dan salah penulisan bisa dengan mudah terjadi. Secara khusus, bila ada banyak variasi nama, untuk mendeteksi semua variasi tersebut sekaligus mengkonsolidasikannya menjadi sebuah entiti merupakan masalah yang signifikan. Masalah ini lebih dikenal sebagai masalah pencocokan nama, keterkaitan data atau masalah resolusi entitas. Nama yang cocok memainkan peran penting dan krusial dalam banyak aplikasi, mulai dari pencarian, penghilangan duplikasi, keuangan, penegakan hukum, bibliografi, dll.

Sebagian besar metode yang telah ada, dikembangkan untuk bahasa Inggris, dan dengan demikian mencakup karakteristik bahasa ini. Sampai saat ini belum ada metode yang spesifik dirancang dan diimplementasikan untuk nama orang Indonesia. Tujuan dari tesis ini adalah untuk mengembangkan dataset nama orang Indonesia sebagai sebuah kontribusi untuk riset akademis dan mengusulkan fitur set yang tepat, yaitu yang memanfaatkan kombinasi konteks string nama orang Indonesia dan nilai permute-winkler nya. Algoritma klasifikasi mesin pembelajaran digunakan sebagai metode untuk pencocokan nama. Berdasarkan percobaan, dengan menggunakan algoritma Random Forest yg sudah di-tuning serta menggunakan fitur set yg diusulkan, terjadi peningkatan kinerja pencocokan sekitar 1.7% dan mampu mengurangi kesalahan klasifikasi sebesar 70% dari metode terbaik. Peningkatan kinerja ini membuat sistem pencocokan lebih efektif dan menurunkan risiko kesalahan klasifikasi.