Normalisasi *String* untuk Optimasi *Phonetic String Matching* dalam Bahasa Indonesia

Arie Karhendana[1], Dicky Wizanajani R[2], Fajar Yuliawan[3]

Laboratorium Ilmu dan Rekayasa Komputasi
Departemen Teknik Informatika, Institut Teknologi Bandung
Jalan Ganesha 10, Bandung

E-mail : if13092@students.if.itb.ac.id[1], if13124@students.if.itb.ac.id[2], if13022@students.if.itb.ac.id[3]

Abstrak

Algoritma pencocokan *string* selalu menemui hambatan, terutama dalam mengidentifikasi nama (misal nama orang atau nama tempat). Hambatan ini merupakan faktor utama penyebab ketidakakuratan algoritma-algoritma pencocokan *string* berdasarkan kemiripan pengucapan, seperti Soundex, Metaphone, dan lainnya. Nama adalah suatu *string* khusus yang bisa sangat bervariasi. Penulisan dan cara pengucapan nama sangat tergantung kepada bahasa yang digunakan. Beberapa nama yang penulisannya berbeda, memiliki cara pengucapan yang sama. Oleh karena itu, perlu adanya suatu langkah normalisasi sebelum dilakukan langkah pencocokan *string*.

***Kata kunci**: phonetic string matching, optimasi pencocokan string, soundex, bahasa Indonesia, q-gram*

1 Pendahuluan

Saat ini sudah banyak ditemukan algoritma pencocokan *string* berdasarkan kemiripan ucapan (*phonetic string matching*). Masing-masing algoritma tersebut memiliki pendekatan yang hampir mirip. Namun, kebanyakan algoritma tersebut ditulis berdasarkan pengucapan dalam bahasa Inggris. Selain itu, algoritma-algoritma tersebut masih menghadapi masalah ketidakakuratan, seiring dengan begitu bervariasinya *string* yang akan dicocokan.

Oleh karena itu, makalah ini ditulis untuk mengatasi dua masalah tersebut, yaitu untuk mendukung algoritma pencocokan *string* berdasarkan bahasa Indonesia, sekaligus melakukan optimasi terhadap algoritma tersebut.

Aturan-aturan yang dipaparkan dalam makalah ini merupakan aturan yang ditemukan secara empirik dan belum dilandasi oleh alasan ilmiah. Namun, pengamatan empirik tersebut kami rasa masih cukup ideal.

2 Ketidakteraturan Nama dalam Bahasa Indonesia

Nama dalam bahasa Indonesia sangat dipengaruhi oleh bangsa-bangsa yang pernah berinteraksi dengan bangsa Indonesia. Oleh karena itu, banyak ditemukan nama yang menggunakan cara penulisan dan pengucapan asing, seperti Belanda, Arab, Cina, maupun Inggris.

Beberapa penulisan nama merujuk kepada satu cara pengucapan. Nama-nama tersebut merupakan variasi cara penulisan dari nama yang sudah ada.

Contoh berikut menunjukkan pernyataan tersebut.

Tabel 1: Contoh variasi nama dan pengucapannya dalam bahasa Indonesia

Nama	Cara Pengucapan
Rahmat	Rahmat
Rachmat	
Rakhmat	
Efendhi	Efendi
Efendi	
Efendy	
Effendhy	
Effendi	
Effendy	

* data diperoleh dari daftar mahasiswa ITB

Tabel 1 menunjukkan bahwa perbedaan nama tersebut bukanlah sebuah kesalahan entri data, namun merupakan variasi umum dari sebuah nama.

Damerau [1] menunjukkan setidaknya ada empat kejadian yang mengakibatkan variasi pada nama.

Tabel 2: Klasifikasi 'kesalahan' menurut Damerau

Jenis	Nama Dasar	Variasi
Insertion (penyisipan)	Fisher	Fischer
Omission (penghilangan)	Johnston	Johnson
Substitution (penggantian)	Catherine	Katherine
Transposition (pertukaran)	Hagler	Halger

Walaupun contoh yang diberikan berupa nama-nama asing, namun jenis-jenis variasi yang dipaparkan pun terjadi dalam nama Indonesia.

3 Bentuk Normal dalam Bahasa Indonesia

Cara pengucapan yang ditunjukkan pada tabel 1 merupakan bentuk paling sederhana yang dapat diucapkan dalam bahasa Indonesia. Dapat dikatakan, bentuk tersebut adalah bentuk normal dari nama-nama tersebut.

Untuk meningkatkan akurasi algoritma pencocokan *string*, maka sebelum 'dilewatkan' ke algoritma tersebut, harus dilakukan normalisasi terhadap *string* tersebut.

Proses normalisasi secara garis besar terbagi menjadi beberapa tahap, yaitu:

1. Normalisasi *q-gram*
2. Eliminasi duplikasi karakter

Tahap yang paling penting dalam proses normalisasi adalah normalisasi *q-gram*. *Q-gram* adalah susunan beberapa huruf yang berurutan. Normalisasi *q-gram* dilakukan dengan mengubah susunan huruf tersebut menjadi *q-gram* lain yang lebih sederhana.

Secara empirik, aturan-aturan tersebut ditunjukkan dalam tabel 3.

Tabel 3: Aturan translasi *q-gram*

q-gram		Contoh	
Awal	**Translasi**	**Awal**	**Translasi**
KH	HH	Rakhmat	Rahhmat
DJ	JJ	Endjang	Enjjang
TJ	CC	Itjang	Iccang
CQ, CK	KK	Erick	Erikk
PH	FF	Philip	Ffilip
DZ	ZZ	Dzikri	Zzikri
SJ	SY	Sjahrir	Syahrir
SY	SS	Syifa	Ssifa
BH,DH, GH,JH, SH,TH, ZH	BB,DD, GG,JJ, SS,TT, ZZ	Ardhi	Arddi
V	F	Saviena	Safiena
KS	XX	Wicaksono	Wicaxxono
OE	UU	Wahyoedi	Wahyuudi
IE	II	Arie	Arii
Y	I	Donny	Donni

Tahap eliminasi duplikasi karakter dilakukan dengan menghilangkan karakter-karakter berurutan yang sama. Kebanyakan duplikasi karakter ini muncul setelah langkah normalisasi *q-gram*.

4 Kendala dalam Proses Normalisasi

Walaupun aturan yang ada sudah cukup ideal dan memadai, namun masih terdapat beberapa *q-gram* yang didefinisikan aturannya. Hal ini disebabkan oleh ambiguitas pengucapan *q-gram* akibat serapan dari berbagai bahasa asing. Selain itu, ambiguitas juga disebabkan oleh perbedaan ejaan Indonesia lama dan baru.

Tabel 4 menunjukkan beberapa *q-gram* yang ambigu.

Tabel 4: Contoh ambiguitas *q-gram*

Q-gram	Variasi	Bentuk Normal	Translasi
CH	Nurcholis	Nurholis	HH
	Christian	Kristian	KK
	Chandra	Candra	CC
J	Joni	Joni	<tetap>
	Dajat	Dayat	Y

Ambiguitas ini menimbulkan masalah, karena *string* yang diperoleh tidak dapat dinormalisasi. Salah satu pemecahannya adalah dengan menggunakan pendekatan statistik untuk menghitung sebaran translasi mana yang lebih sering muncul. Namun, untuk melakukan hal ini dibutuhkan sampel nama yang cukup besar dan waktu yang relatif lama.

5 Analisis Efektivitas

Untuk menguji hasil normalisasi, kami mencoba menganalisis *string* dengan menggunakan algoritma Soundex standar. Algoritma Soundex yang digunakan adalah algoritma yang murni (untuk bahasa Inggris).

Hasil analisis tersebut ditunjukkan dalam tabel 5.

Tabel 5: Analisis efektivitas normalisasi

String	**Soundex** (tanpa normalisasi)	**Soundex** (dengan normalisasi)
Rahmat	R530	R530
Rachmat	R253	
Rakhmat	R253	
Dzikri	D226	Z260
Zikri	Z260	
Ginanjar	G552	G552
Ginandjar	G553	

Terlihat bahwa *string* yang memiliki kesamaan ucapan mendapat hasil Soundex yang sama setelah dinormalisasi.

Walaupun analisis yang dilakukan masih sederhana, namun sudah dapat menggambarkan efektivitas normalisasi yang dilakukan. Namun untuk pembuktian secara ilmiah, tentunya membutuhkan penelitian yang lebih lama dan mendalam.

6 Kesimpulan

Proses normalisasi menghasilkan *string* yang sederhana, dengan pengucapan yang sama dalam bahasa Indonesia.

Aturan-aturan normalisasi yang dipaparkan dalam makalah ini didapatkan berdasarkan pengamatan empirik dari daftar nama mahasiswa ITB. Oleh karena itu, aturan yang didapatkan belum dapat dibuktikan secara ilmiah.

Selain itu, aturan-aturan (*rules*) yang ada pun perlu diperbaiki agar dapat mengakomodasi *q-gram* yang lebih rumit maupun *q-gram* yang ambigu. Langkah yang dilakukan misalnya dengan menggunakan pendekatan statistik yang lebih rumit.

Bagaimanapun juga, proses normalisasi ini merupakan tahap awal untuk menemukan algoritma pencocokan *string* dalam bahasa Indonesia. Tentunya algoritma seperti ini akan sangat berguna untuk bidang-bidang lain seperti basis data, *data mining*, atau bahkan *natural language processing*.

Daftar Pustaka

[1] F. Damerau, *A Technique for Computer Detection and Correction of Spelling Errors*, Communications of the ACM 7, 171-176, 1964.