Prosiding SNRT (Seminar Nasional Riset Terapan) **ISSN 2341-5662 (Cetak)**
Politeknik Negeri Banjarmasin, 7 November 2018 **ISSN 2341-5670 (Online)**

PERBANDINGAN METODE PERHITUNGAN KEMIRIPAN KATA

Reza Fauzan[1], Joni Riadi[2], Fuad Sholihin[2]
Politeknik Negeri Banjarmasin[123]
reza.fauzan@poliban.ac.id[1]
joni_akademika@yahoo.com[2]
fuad.sholihin@gmail.com[3]

ABSTRACT

Measuring similarities between words, sentences, paragraphs, and documents is an important component in various tasks such as information retrieval, document grouping, word-sense disambiguation, automatic essay giving, short assessment, machine translation and text summarization. This research discusses good methods used in determining the similarity of words. The method used is levenshtein and jaro winkler. Testing is done by comparing the results of the system to the expert so that it gets the accuracy value from the system. The test results show that levenshtein is better than jaro winkler.

Keywords: *jaro winkler, levenshtein, similarity*

ABSTRAK

Mengukur kemiripan antara kata, kalimat, paragraf, dan dokumen adalah komponen penting dalam berbagai tugas seperti pengambilan informasi, pengelompokan dokumen, disambiguasi kata-akal, pemberian esai otomatis, penilaian singkat, terjemahan mesin dan peringkasan teks. Penelitian ini membahas metode yang baik digunakan dalam menentukan kemiripan kata. Metode yang digunakan adalah levenshtein dan jaro winkler. Pengujian dilakukan dengan membandingkan hasil sistem terhadap pakar sehingga mendapatkan nilai akurasi dari sistem. Hasil pengujian menunjukkan levenshtein lebih baik daripada jaro winkler.

Kata Kunci: jaro winkler, kemiripan, levenshtein

Dalam dunia akademik, karya tulis seseorang dinilai berdasarkan keorisinalannya. Dengan adanya perkembangan pengarsipan berkas, terutama melalui internet yang terus berkembang saat ini, akses untuk mendapatkan karya-karya tulis tersebut menjadi semakin mudah. Contoh karya tulis akademis yang tersebar di internet adalah jurnal ilmiah dan tugas akademis mahasiswa. Karenanya, kemungkinan untuk menyalin karya-karya tulis tersebut menjadi lebih tinggi, dan keorisinalan karya tulis akademis pun semakin dipertanyakan. Dengan demikian, perlu adanya pendeteksian kemiripan teks untuk karya-karya tulis akademis sehingga keorisinalannya dapat diketahui dengan cepat (Imbar *et al.*, 2014).

Pengukuran kesamaan teks memainkan peran yang semakin penting dalam penelitian dan aplikasi terkait teks dalam tugas-tugas seperti pengambilan informasi, klasifikasi teks, pengelompokan dokumen, pendeteksian topik, pelacakan topik, pembuatan pertanyaan, penjawab pertanyaan, penilaian esai, pemberian skor jawaban singkat, terjemahan mesin, teks summarization dan lainnya. Menemukan kesamaan antara kata-kata adalah bagian mendasar dari

kesamaan teks yang digunakan sebagai tahap utama untuk kalimat, paragraf, dan kesamaan dokumen.

Namun, pendeteksian kemiripan teks menjadi tugas yang sulit dilakukan oleh manusia karena banyak dan besarnya teks untuk dibandingkan serta strukturnya yang tidak konsisten dan kompleks. Pendeteksian kemiripan teks dapat dilakukan untuk berbagai tujuan, salah satunya adalah untuk mencegah plagiarisme (tindakan menyatakan karya orang lain sebagai karya sendiri tanpa mengacu pada karya asli).

Text Mining adalah sebuah penerapan yang berasal dari information retrieval (IR) dan natural language processing (NLP). Definisi text mining secara sempit hanya berupa metode yang dapat menemukan informasi baru yang tidak jelas atau mudah diketahui dari sebuah kumpulan dokumen. Sedangkan secara lebih luas, text mining mencakup teknik text-processing yang lebih umum, seperti pencarian, pengambilan intisari, dan pengkategorian (Konchady, 2006).

Permasalahan yang dihadapi pada text mining adalah jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah, serta data noise. Sehingga sumber data yang digunakan pada text mining adalah kumpulan teks yang memiliki bentuk yang tidak terstruktur atau setidaknya semi terstruktur (Triawati, 2009).

Text mining memiliki beragam metode dalam menentukan kemiripan kata. Dua dari beragam metode tersebut adalah levenshtein dan jaro winkler. Pada penelitian ini penulis mencoba membandingkan keakuratan kedua metode tersebut dalam menentukan kemiripan kata.

METODE PENELITIAN

Tahapan penelitian yang akan dilakukan dalam beberapa tahapan (Fauzan, Indrasary and Muthia, 2018). Kegiatan penelitian mencakup studi pendahuluan, pengumpulan data, analisa kebutuhan, implementasi metode, dan pengujian sistem.

Pengumpulan Data

Masukkan yang digunakan adalah kata berbahasa Inggris. Kata pertama adalah kata yang selalu tetap di setiap pengujian. Kata kedua lebih bervariasi, yaitu kata-kata yang memiliki huruf dan makna yang mirip terhadap kata pertama. Kata dengan tulisan yang mirip dimaksudkan agar apakah sistem dapat mendeteksi kesalahan penulisan. Dan kata dengan makna yang mirip dimaksudkan agar apakah sistem dapat mengetahui kata yang mirip berdasarkan makna.

Analisa Kebutuhan

Analisa kebutuhan merupakan fase penting dalam pengembangan perangkat lunak (Fauzan and Pramono, 2013). Kesalahan pendefinisian kebutuhan pengguna akan berakibat pada kegagalan perangkat lunak meskipun secara teknis, perangkat lunak bisa digunakan dengan baik. Sebelum dilakukan analisa kebutuhan, keberadaan data multak diperlukan (SBN *et al.*, 2017).

Metode Levenshtein

Pada teori informasi dan ilmu komputer, Levenshtein Distance merupakan matriks untuk mengukur nilai jumlah perbedaan antara 2 string yaitu string sumber (s) dan string target (t). Nilai Levenshtein Distance antara dua kata merupakan nilai minimum dari pengeditan single-character (yaitu insertion, deletion maupun substitution) membutuhkan perubahan pada salah satu kata (Fahma, Cholissodin and Perdana, 2018).

Levenshtein Distance antara dua string ditentukan berdasarkan jumlah minimum pengeditan yang diperlukan untuk melakukan transformasi dari satu bentuk string ke bentuk string yang lain. Notasi yang digunakan untuk Levenshtein Distance adalah $LD(s, t)$ dengan s yaitu sumber dan t adalah target. Misalnya, jika source string (s) adalah "tihun" dan target string (t) dalah "tahun" maka nilai Levenshtein Distance adalah 1, dalam hal ini berarti dibutuhkan sebuah operasi yaitu substitution untuk mengubah source string (s) menjadi sama dengan target string (t).

Operasi dilakukan dengan cara menukar posisi karakter yang berdekatan dan menemukan kata yang sama dalam dictionary (Naradhipa *et al.*, 2011). Secara matematis, Levenshtein Distance antara dua string, misal string sumber a dan string target b (panjang |a| dan |b|) dengan $lev_{a,b}(|a|,|b|)$ pada indeks i dan j dimana telah dijelaskan pada Persamaan 1.

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & \min(i,j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & lainnya \end{cases} \quad \text{Persamaan 1}$$

Metode Jaro Winkler

Jaro-Winkler merupakan varian dari Jaro Distance metrik yaitu sebuah algoritma untuk mengukur kesamaan antara dua buah string, biasanya algoritma ini digunakan di dalam pendeteksian duplikat. Semakin tinggi nilai Jaro-Winkler untuk dua string, maka semakin tinggi presentase kemiripan kedua buah string tersebut (Kurniawati, 2014).

Langkah dasar untuk menghitung algoritma Jaro antara dua string s1 dan s2 adalah sebagai berikut (Suryaningrum and T, 2016).

a. Menghitung panjang string s1 dan s2.
b. Menemukan jumlah karakter yang "sama persis"(m) dalam dua string yang dibandingkan.
c. Menghitung jumlah transposisi kedua buah string.

Rumus dari jaro winkler ditampilkan pada Persamaan 2.

$$d_j = \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right) \quad \text{Persamaan 2}$$

Pengujian Sistem

Pengujian dilakukan menggunakan instrumen akurasi hasil (Fauzan and Basuki, 2016; Fauzan, Fitri and Fadliansyah, 2017). Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual (Taylor, 1999). Akurasi dapat dihitung dengan Persamaan 3.

$$Akurasi = \frac{Jumlah\ Sesuai}{Jumlah\ Masukkan} \times 100\% \qquad \text{Persamaan 3}$$

HASIL DAN PEMBAHASAN

Hasil pengujian menggunakan metode levenshtein ditampilkan pada Tabel 1. Sistem memberikan hasil mirip dengan nilai ambang batas sebesar 70.

Tabel 1. Pengujian Metode Levenshtein

No	Kata 1	Kata 2	Nilai Kemiripan	Hasil Sistem	Hasil Pakar
1	Home	Home	100	Mirip	Mirip
2	Home	Homes	80	Mirip	Mirip
3	Home	House	60	Tidak	Mirip
4	Home	Hom	75	Mirip	Mirip
5	Home	H	25	Tidak	Tidak
6	Home	me	50	Tidak	Tidak
7	Home	ho	50	Tidak	Tidak
8	Home	Place	20	Tidak	Mirip
9	Home	Recidence	11	Tidak	Mirip
10	Home	Apartment	22	Tidak	Mirip
11	Home	flat	0	Tidak	Mirip
12	Home	bungalow	0	Tidak	Mirip
13	Home	cottage	29	Tidak	Mirip

Dari Tabel 1, akurasi yang didapatkan adalah sebagai berikut.

$$akurasi = \frac{Jumlah\ Sesuai}{Jumlah\ Masukkan} \times 100\%$$

$$akurasi = \frac{6}{13} \times 100\%$$

$$akurasi = 46{,}15384615384615$$

Hasil pengujian menggunakan metode Jaro Winkler ditampilkan pada Tabel 2. Sistem memberikan hasil mirip dengan nilai ambang batas sebesar 70.

Tabel 2. Pengujian Metode Jaro Winkler

No	Kata 1	Kata 2	Nilai Kemiripan	Hasil Sistem	Hasil Pakar
1	Home	Home	100	Mirip	Mirip
2	Home	Homes	96	Mirip	Mirip
3	Home	House	83	Mirip	Mirip
4	Home	Hom	94	Mirip	Mirip
5	Home	H	78	Mirip	Tidak
6	Home	me	0	Tidak	Tidak

7	Home	ho	87	Mirip	Tidak
8	Home	Place	48	Tidak	Mirip
9	Home	Recidence	45	Tidak	Mirip
10	Home	Apartment	0	Tidak	Mirip
11	Home	flat	0	Tidak	Mirip
12	Home	bungalow	0	Tidak	Mirip
13	Home	cottage	0	Tidak	Mirip

Dari Tabel 2, akurasi yang didapatkan adalah sebagai berikut.

$$akurasi = \frac{Jumlah\ Sesuai}{Jumlah\ Masukkan} \times 100\%$$

$$akurasi = \frac{5}{13} \times 100\%$$

$$akurasi = 38{,}46153846153846$$

Kata pertama sampai dengan yang keempat dikatakan mirip oleh pakar karena terjadi sedikit kesalahan penulisan. Kata kelima sampai ketujuh dikatakan tidak mirip oleh pakar karena terdapat banyak kesalahan penulisan. Kata kedelapan sampai ketigabelas dikatakan mirip oleh pakar karena memiliki kemiripan makna.

Hasil pengujian menunjukkan metode levenshtein lebih baik daripada metode jaro winkler. Tetapi dari hasil akurasi kedua metode tersebut, masih belum dapat dinyatakan baik dengan nilai yang kurang dari 50 %. Hal ini disebabkan kedua metode tersebut hanya melihat dari huruf yang ada di dalam kata tersebut sehingga kata-kata yang dianggap mirip adalah kata yang memiliki tulisan hampir sama. Sedangkan kata yang memiliki tulisan berbeda tetapi makna yang sama tidak bisa diselesaikan menggunakan kedua metode ini.

KESIMPULAN

Pengujian menunjukkan nilai akurasi metode levenshtein adalah 46,15 %. Nilai akurasi metode jaro winkler adalah 38,46 %. Hal ini menunjukkan metode lenenshtein lebih baik dari metode jaro winkler. Kedua metode ini hanya dapat menemukan kemiripan dari kata-kata mirip yang memiliki sedikit kesalahan penulisan. Kedua metode ini tidak dapat menangani kata-kata yang mirip dari sisi makna kata.

Saran untuk penelitian selanjutnya adalah agar dimunculkan metode yang dapat mengetahui kata-kata yang mirip berdasarkan makna kata.

UCAPAN TERIMA KASIH

Terima kasih kepada Politeknik Negeri Banjarmasin yang telah membantu dalam hal pendanaan pada penelitian ini.

DAFTAR PUSTAKA

Fahma, A. I., Cholissodin, I. and Perdana, R. S. (2018) ‘Identifikasi Kesalahan Penulisan Kata (Typographical Error) pada Dokumen Berbahasa

Indonesia Menggunakan Metode N-gram dan Levenshtein Distance', *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(1), pp. 53–62.

Fauzan, R. and Basuki, S. (2016) 'PURWARUPA WEBSITE MANAJEMEN BENCANA', in *Seminar Nasional Riset Terapan*, pp. 9–10.

Fauzan, R., Fitri, R. and Fadliansyah, M. (2017) 'Sistem informasi penjurusan dan penerimaan peserta didik baru menggunakan metode weighted product', 1(1), pp. 11–22.

Fauzan, R., Indrasary, Y. and Muthia, N. (2018) 'Sistem Pendukung Keputusan Penerimaan Beasiswa Bidik Misi di POLIBAN dengan Metode SAW Berbasis Web', Jurnal Online Informatika, 2(2), p. 79. doi: 10.15575/join.v2i2.101.

Fauzan, R. and Pramono, D. (2013) 'Pemeriksaan Kemiripan Diagram Kasus Penggunaan Terhadap Skenario', JUTI:Jurnal Teknologi Informasi, 11, pp. 49–55. doi: http://dx.doi.org/10.12962/j24068535.v11i2.a11.

Imbar, R. V., Ayub, M., Rehatta, A., Jurusan, S., Informasi, S., Jurusan, S. and Informatika, T. (2014) 'Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks', Jurnal Informatika, pp. 31–42.

Konchady, M. (2006) Text Mining Application Programming. Boston: Charles River Media.

Kurniawati, A. (2014) 'Implementasi Algoritma Jaro-Winkler Distance Untuk Membandingkan Kesamaan Dokumen Berbahasa Indonesia', in. Indonesia.

Naradhipa, A. R., Kamayani, M., Reinanda, R., Simbolon, S., Soleh, M. Y. and Purwarianti, A. (2011) 'Application of Document Spelling Checker for Bahasa Indonesia', in ICACSIS.

SBN, A., Fauzan, R., Badriansyah, Halimah and Nawisworo, P. B. (2017) 'A45', in Seminar Nasional Riset Terapan. P3M Politeknik Negeri Banjarmasin, pp. 44–52.

Suryaningrum, K. M. and T, A. (2016) 'Pengkoreksian dan Suggestion Word pada Keyword Menggunakan Algoritma Jaro-Winkler', Jurnal Teknologi Informasi-AITI, 13(2), pp. 169–181.

Taylor, J. R. (1999) An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books.

Triawati, C. (2009) Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia. Bandung, Indonesia. Available at: http://digilib.ittelkom.ac.id/index.php?option=com_content&view=article&id=590:text-mining&catid=20:informatika&Itemid= 14.