

ANALISIS DATA TEKS MENGUNAKAN METODE HEURISTIC MATCHING STUDI KASUS : TEKS BAHASA BALI

Ricky Aurelius Nurtanto Diaz

Program Studi Sistem Komputer

STMIK STIKOM Bali

E-mail : ricky.aurelius@gmail.com

Abstrak

Teks mining merupakan salah satu bidang Data Mining dimana yang menjadi dasar dalam proses mining adalah data yang berupa data teks. Dengan memanfaatkan text mining maka pola tertentu dalam pencarian bisa ditemukan dengan waktu atau proses yang lebih singkat. Dalam penelitian ini, sebelum mengarah ke pencarian berbasis konten, maka perlu terlebih dahulu untuk dilakukan proses penelitian di sisi pencarian dengan menggunakan teknik Heuristic Matching. Teknik ini diterapkan dalam teks bahasa Bali, dimana dalam proses pencarian, teknik Heuristic Matching akan bekerja dengan konsep bigram. Hasil dari pengujian menunjukkan bahwa teknik tersebut dapat diterapkan dalam pencarian teks namun karena tingkatan bahasa Bali yang beragam menyebabkan, proses pencarian hanya menghasilkan teks yang memiliki kemiripan penulisan walaupun berbeda makna.

Kata kunci : Heuristic matching, teks, bahasa Bali

1. PENDAHULUAN

Dalam pengembangan sistem informasi, terdapat berbagai teknik atau algoritma yang saling berkaitan atau berhubungan dalam proses maupun fitur yang dimilikinya. Perkembangan berbagai algoritma atau teknik yang digunakan sangat mempengaruhi kinerja sistem tersebut. Hal ini juga berpengaruh terhadap beberapa bentuk penelitian yang dilakukan untuk algoritma atau teknik tertentu. Data mining adalah proses untuk penggalian pola-pola dari data. Alasan utama untuk menggunakan data mining adalah untuk membantu dalam analisis koleksi pengamatan perilaku. Data tersebut rentan terhadap *collinearity* karena diketahui keterkaitan. Fakta yang tidak terelakkan data mining adalah bahwa subset/set data yang dianalisis mungkin tidak mewakili seluruh domain, dan karenanya tidak boleh berisi contoh-contoh hubungan kritis tertentu dan perilaku yang ada di bagian lain dari domain.

Salah satu fitur yang tersedia dalam sistem informasi adalah fitur pencarian. Selama ini proses pencarian didasarkan pada bentuk pencarian dasar berdasarkan suku kata atau yang sering dikenal dengan *exact matching*. Disisi lain proses pencarian data tidak hanya berdasar pada suku kata dasar namun juga telah mengarah ke sisi pencarian berdasarkan isi atau konten atau makna dari dokumen tertentu. Hal ini tentu mengarah dengan salah satu bidang

dalam data mining yaitu text mining. Dengan memanfaatkan text mining maka pola tertentu dalam pencarian bisa ditemukan dengan waktu atau proses yang lebih singkat. Pada penelitian sebelumnya, Arrummaisha dkk, melakukan penelitian dengan judul Pemilahan Artikel Berita dengan Text Mining. Bambang Kurniawan dkk juga melakukan penelitian serupa dengan judul Klasifikasi Konten Berita dengan Metode Text Mining. Hasil dari penelitian ini menghasilkan sistem klasifikasi berita berbasis web dengan menggunakan bahasa pemrograman PHP dan database MySQL menunjukkan bahwa berita *testing* bisa terklasifikasi secara otomatis seluruhnya dengan bantuan Naïve Bayes Classifier. Mesran melakukan penelitian dengan judul penelitian Implementasi Algoritma Brute Force dalam Pencarian Data Katalog Buku Perpustakaan.

Dalam penelitian ini, sebelum mengarah ke pencarian berbasis konten, maka perlu terlebih dahulu untuk dilakukan proses penelitian di sisi pencarian dengan menggunakan teknik *Heuristic Matching*. Digunakannya bahasa Bali dalam penelitian ini disebabkan keunikan yang dimiliki oleh bahasa Bali dimana terdapat cukup banyak kata dengan bentuk yang sama namun berbeda makna atau sebaliknya. Hal ini tentu dapat menjadi pertimbangan khusus yang dapat diuji

coba dalam teknik *Heuristic Matching* untuk proses pencarian teks.

2. STRING MATCHING

String matching adalah proses pencarian semua kemunculan *query* yang selanjutnya disebut *pattern* ke dalam *string* yang lebih panjang atau teks. Dalam penerapannya string matching dibagi menjadi dua yaitu Exact Matching dan Heuristic Matching (Mesran, 2014). *Exact matching* digunakan untuk menemukan pattern yang berasal dari satu teks. Contoh pencarian exact mathing adalah pencarian kata “pelajar” dalam kalimat “saya seorang pelajar” atau “saya seorang siswa”. Sistem akan memberikan hasil bahwa kalimat pertama mengandung kata “pelajar” sedangkan kalimat kedua tidak. Beberapa algoritma yang tergolong dalam exact matching adalah Brute Force, Boyer Moore dan Colussi. *Heuristic Matching* adalah teknik yang digunakan untuk menghubungkan dua data terpisah ketika exact matching tidak mampu melakukan proses pencarian dengan sesuai.

3. HEURISTIC MATCHING

Dalam proses Heuristic Matching di penelitian ini, digunakan konsep *N-gram Similarity* yang bertujuan untuk membandingkan beberapa *string* (teks). Metode ini menghitung sejumlah *n-gram* bersama-sama seperti serangkaian *n* karakter diantara string. Sebagai contoh, bigram untuk string article adalah “ar”, “rt”, “ti”, “ic”, “cl” dan “le”. Untuk trigram pada string yang sama adalah “art”, “rti”, “tic”, “icl”, “cle”.

Selanjutnya akan dilakukan proses perhitungan kemiripan dua string dengan menggunakan *Jaccard Coefficient*, yaitu :

$$\text{Jaccard Coefficient (pattern, text)} = \frac{|pattern \cap text|}{|Pattern \cup text|} \quad (1)$$

Berikut adalah contoh perbandingan data string dari dua buah string menggunakan *n-gram similarity*. Terdapat dua buah string :

- Artikel : ar, rt, ti, ik, ke, el
- Arikel : ar, ri, ik, ke, el

dari dua string tersebut terdapat 4 model bigram yang sama yaitu : *ar*, *ik*, *ke* dan *el*. Maka perhitungan nilai similarity nya adalah :

$$\text{Sim (Artikel, Arikel)} = 4/7 = 0,57$$

Dari hasil perhitungan ini diperoleh hasil bahwa dua buah string diatas memiliki nilai kesamaan yang cukup tinggi karena mendekati 1.

4. BAHASA BALI

Bahasa Bali adalah sebuah bahasa Austronesia dari cabang Sundik dan lebih spesifik dari anak cabang Bali-Sasak. Bahasa ini terutama dipertuturkan di pulau Bali, pulau Lombok bagian barat, dan sedikit di ujung timur pulau Jawa. Di Bali sendiri Bahasa Bali memiliki tingkatan penggunaannya, misalnya ada yang disebut Bali Alus, Bali Madya dan Bali Kasar. Yang halus dipergunakan untuk bertutur formal misalnya dalam pertemuan di tingkat desa adat, meminang wanita, atau antara orang berkasta rendah dengan berkasta lebih tinggi. Yang madya dipergunakan di tingkat masyarakat menengah misalnya pejabat dengan bawahannya, sedangkan yang kasar dipergunakan bertutur oleh orang kelas rendah misalnya kaum sudra atau antara bangsawan dengan abdi dalemnya.

Dalam penggunaan bahasa Bali, didasarkan pada kata dasar. Kata dasar (kruna lingga) adalah kata yang belum mendapatkan atau mengalami proses pengimbuhan maupun pengulangan. Kata dasar juga bisa diartikan kata yang belum mendapatkan awalan (prefiks), sisipan (infiks), maupun akhiran (sufiks). Pengertian lain tentang kata dasar adalah kata yang belum mendapatkan apa-apa, misalnya belum mendapat imbuhan (afiks), persengauan (nasalisasi), atau dijadikan kata mjemuk sebagai hasil proses pembentukan kata menjadi kata jadian atau kata bersusun.

5. MODEL KONSEPTUAL PENELITIAN

Dalam pelaksanaan penelitian akan dilakukan beberapa tahap yaitu komponen pertama adalah input, dimana pada bagian ini sistem akan menerima inputan dari *user* berupa data teks dalam bahasa Bali. Teks tersebut kemudian akan disimpan dalam database. Selanjutnya, jika ada teks lain yang akan dibandingkan dengan teks yang telah ada sebelumnya, maka akan dilakukan proses perbandingan dengan konsep *N-gram Similarity* dan diukur menggunakan *Jaccard Coefficient*. Hasil akhirnya adalah nilai koefisien kemiripan data.

6. IMPLEMENTASI PROSES

PENCARIAN TEKS BAHASA BALI

Untuk melakukan proses pencarian teks, digunakan dua interface yang menjadi *interface* utama dalam penerapan teknik *Heuristic Matching* ini. Halaman pertama adalah halaman pencarian seperti terlihat pada Gambar 1, dimana halaman ini digunakan oleh pengguna untuk memasukkan teks dalam bahasa Bali yang akan dicari.

Gambar 1 Halaman Pencarian

Interface kedua adalah halaman yang berfungsi untuk menampilkan hasil pencarian berdasarkan *Heuristic Matching*, dan jika tersedia teks lain yang memiliki *Jaccard Coefficient* yang cukup dekat dengan teks yang diinputkan atau teks bermakna sama maka akan ditampilkan oleh sistem pada urutan kedua dan ketiga.

Gambar 2 Hasil Pencarian

Berikut adalah contoh pencarian kata yang dilakukan. Terdapat dua buah string :

- Ngajen : Ng, ga, aj, je, en
- Ngajeng : Ng, ga, aj, je, en, ng

dari dua string tersebut terdapat 5 model *bigram* yang sama yaitu : Ng, ga, aj, je, en. Selanjutnya akan dilakukan proses perhitungan kemiripan dua string dengan menggunakan *Jaccard Coefficient*, yaitu :

$$Sim (Ngajen, Ngajeng) = 5/6 = 0,83$$

Dari hasil perhitungan ini diperoleh hasil bahwa dua buah string diatas memiliki nilai kesamaan yang cukup tinggi karena mendekati 1. Dengan nilai kesamaan yang tinggi ini, maka sistem akan menampilkan kata Ngajeng pada urutan kata teratas, lalu diikuti dengan pilihan kata yang lain namun bermakna sama. Kata lain yang bermakna sama ini disediakan agar pengguna dapat mengetahui padanan kata yang lain meskipun tentu saja berbeda waktu pemakaiannya (Bahasa Halus, Madya dan Kasar). Pada kondisi berikutnya, bila tidak terdapat padanan kata yang sama, maka sistem akan menampilkan hasil pencarian kata dengan *Jaccard Coefficient* yang terdekat meskipun berbeda makna. Sebagai contoh adalah kata *Sirah* yang berarti Kepala dan *Sira* yang berarti Siapa.

7. ANALISIS PENGUJIAN

Dari proses pengujian data, diperoleh hasil bahwa *heuristic matching* dapat digunakan untuk proses pencarian teks dalam bahasa Bali, dengan kemungkinan teks yang dicari telah dipelajari atau disimpan sebelumnya. Proses pencarian dapat ditambahkan dengan alternatif kata lain yang dapat juga dijadikan sebagai pembanding dengan teks utama yang dicari, meskipun memiliki kemungkinan besar akan berbeda makna. Berikut adalah hasil pengujian data uji yang dilakukan dalam penelitian ini :

Tabel 1. Data Uji

Kata Uji	Heuristic Matching (Alternatif 1)	Heuristic Matching (Alternatif 2)
Batis	Batis	Cokor
Jalan	Mejalan	Melali
Basng	Basang	Masang
Hati	Ati	Pasang
Keneh	Keneh	Mekeneh
Getih	Getih	Lete
Sirah	Sirah	Sira
Maong	Baong	Meong
Bok	Bok	Mbok
Bibi	Bibi	Bibih
Untuk	Untu	Entu
Wayah	Wayah	Layah
Kedek	Kedek	Bedeg
Eling	Ngeling	Beling

Kata Uji	Heuristic Matching (Alternatif 1)	Heuristic Matching (Alternatif 2)
Ngajen	Ngajeng	Ngamah
Makpak	Makpak	
Lebeng	Lebeng	Melebengan
Nginem	Nginem	
Cegut	Nyegut	Makpak
Sirep	Sirep	Arep
Rebahan	Mebahan	
Ngipi	Ngipi	Lipi
Bapak	Bapak	Bapa
Jait	Nyait	Pait
Jarum	Jaum	Paum
Tebek	Tebek,	Nebek
Lipi	Lipi	Ngipi
Cacing	Cacing	Cicing
Dadang	Padang	Gadang
Pasih	Pasih	Sapih
Nyah	Nyah	Nyuh
Kebus	Kebus	Kepus
Bara	Barak	Arak
Kuning	Kuning	Nuning
Gadang	Gadang	Padang
Gelem	Gelem	Selem
Lek	Lek	Belek
Tua	Tua	Lua
Beler	Beler	Baler
Beneh	Beneh	
Peteng	Peteng	Siteng
Pidan	Pidan	Selidan
Dija	Dija	Bija
Konyang	Iragejakonyang	Konyang
Onyan	Onyang	Konyang
Ken	Ken	Yen
Yen	Yen	Ken
Sing	Sing	Ring
Sepit	Sepit	
Linggah	Linggah	

8. KESIMPULAN DAN SARAN

Kesimpulan dari penelitian yang telah dilakukan adalah :

1. Data teks bahasa bali dapat digunakan dalam analisis data teks untuk berbagai

keperluan baik untuk tujuan clustering maupun klasifikasi.

2. *Heuristic Matching* memiliki kelemahan dalam menemukan kata yang memiliki kemiripan makna tetapi berbeda tulisan.
3. Kekurangan tersebut dapat semakin terlihat jika diterapkan dalam teks bahasa Bali, dimana dalam penulisan dan pembicaraan bahasa Bali terdapat perbedaan tingkat bahasa yaitu : Bahasa Bali Halus, Bahasa Bali Madya dan Bahasa Bali Kasar.

Adapun yang menjadi saran dalam penelitian ini adalah, dalam penelitian selanjutnya, proses pencarian data dapat diatasi dengan menggunakan pencarian menggunakan metadata. Hal ini dapat didasarkan dari sumber kata asli yang kemudian dijabarkan ke dalam tingkatan-tingkatan bahasa Bali.

9. DAFTAR PUSTAKA

- Arrummaisha Adrifina, 2008. Pemilahan Artikel Berita Dengan Text Mining, Proceeding, Seminar Ilmiah Nasional Komputer dan Sistem Intelijen (KOMMIT 2008)
- Bambang Kurniawan, 2012. Klasifikasi Konten Berita Dengan Metode Text Mining, JURNAL DUNIA TEKNOLOGI INFORMASI Vol. 1, No. 1, 2012
- Daniel T. Larose, 2005. Discovering Knowledge in Data: An Introduction to Data Mining
- Ed. I. (2012). *JavaScript untuk Membangun Website Profesional*. Yogyakarta : ANDI
- Kusrini, 2009, Algoritma Data Mining, Andi, Yogyakarta.
- Mesran, 2014. Implementasi Algoritma Brute Force Dalam Pencarian Data Katalog Buku Perpustakaan, Majalah Ilmiah Informasi dan Teknologi Ilmiah (INTI) Volume III Mei 2014.
- Saputra, Agus. (2011). *Trik dan Solusi Jitu Pemrograman PHP*. Jakarta : Elex Media Komputindo
- Supriyanto, Aji. (2007). *WEB dengan HTML dan XML*. Yogyakarta : Graha Ilmu
- Turban, Efraim, et al. 2005. Decision Support Systems and Intelligent Systems 7th Ed. New Jersey : Pearson Education.