Seminar Nasional Sistem Informasi dan Teknologi Informasi 2018
SENSITEK 2018
STMIK Pontianak, 12 Juli 2018

Pengelompokan Artikel Bahasa Bali Menggunakan Algoritma K-Means Clustering

Ricky Aurelius Nurtanto Diaz[1)]

STMIK STIKOM BALI
Jl. Raya Puputan No.86,Denpasar, 0361-244445
e-mail: ricky@stikom-bali.ac.id

Abstrak

Teks mining merupakan salah satu bidang data mining yang memiliki cukup banyak hal untuk diteliti, terutama karena Indonesia memiliki cukup banyak ragam bahasa dan tulisan-tulisan dalam bahasa daerah yang mempunyai ciri khas masing-masing. Dalam penelitian ini, akan dilakukan penelitian mengenai proses pencarian teks artikel yang tertulis dalam bahasa Bali. Digunakannya bahasa Bali dalam penelitian ini karena keunikan yang dimiliki oleh bahasa Bali dimana terdapat banyak kata dengan bentuk yang sama namun bisa berbeda makna atau sebaliknya dengan makna yang sama namun berasal dari kata yang berbeda. Pemanfaatan teknik N-Gram Similarity merupakan proses awal pengenalan teks yang terdapat pada sebuah artikel. Hasil pengenalan teks kemudian disimpan dalam nilai variable key yang terus meningkat seiring dengan ditemukannya teks lain yang sesuai dan variabel noise untuk kumpulan teks yang tidak sesuai. Hasil pengenalan teks ini kemudian akan dikelompokkan dengan menggunakan algoritma K-Means dan menghasilkan akurasi hingga 93%. Proses ini dapat menjadi dasar dalam penelitian berikutnya untuk pencarian artikel bahasa daerah menggunakan teknik semantik search.

Kata kunci: *clustering, k-means, artikel, bahasa, bali*

1. Pendahuluan

Dalam bidang data mining, beragam permasalahan pada berbagai bidang dapat diselesaikan dengan menerapkan metode data mining dengan berdasarkan pada teknik tertentu. Salah satu bidang yang memiliki keterkaitan dengan proses mining adalah pengelompokan artikel. Tujuan pengelompokan artikel pada dasarnya digunakan sebagai acuan dalam proses pencarian informasi, baik secara otomatis dimana fitur ini disediakan oleh mesin pencari, maupun proses pencarian berdasarkan permintaan pengguna seperti mencari artikel yang berada pada kategori tertentu. Banyaknya jumlah artikel yang tersedia baik dalam bentuk hardcopy maupun softcopy menghasilkan kebutuhan baru bagi pengguna sistem, dimana kemampuan sistem dalam mengelompokan artikel akan sangat membantu pengguna dalam proses pencarian dan menemukan informasi yang diinginkan. Selama ini telah banyak penelitian dilakukan untuk melakukan proses pengelompokan artikel baik dalam Bahasa Indonesia maupun bahasa asing dimana penelitian tersebut terfokus pada beberapa kategori tertentu. Akhmad Zaini pada tahun 2017 melakukan penelitian dengan judul Pengelompokan Artikel Berbahasa Indonesia Berdasarkan Struktur Laten Menggunakan Pendekatan Self Organizing Map, dimana hasilnya menunjukkan bahwa penggunaan struktur laten dapat mengurangi dimensi ciri sebesar 32% dari dimensi ciri kemunculan kata, sehingga berdampak pada efisiensi waktu. Hasilnya juga menunjukkan bahwa struktur laten, ketika diterapkan pada algoritma SOM, mampu menghasilkan kualitas yang baik jika dibandingkan dengan menggunakan frekuensi kemunculan kata [1]. Penelitian lain yang dilakukan oleh Ambarwati dan Edi pada judul penelitian Pengelompokan Berita Indonesia Berdasarkan Histogram Kata Menggunakan Self-Organizing Map, menunjukkan bahwa teknik atau algoritma yang digunakan dalam sistem dapat menampilkan koleksi dokumen dari lima kategori berita yang ada pada tiap tahunnya, selain itu sistem yang dihasilkan mampu menampilkan banyaknya kata yang sering muncul pada tiap artikel berita (top 10 words) [2].

Dari beberapa pilihan algoritma data mining dan berdasarkan kondisi artikel yang akan diolah, maka dalam penelitian ini digunakan metode K-Means sebagai algoritma yang digunakan dalam proses pengelompokan. Algoritma ini dipilih juga berasarkan hasil penelitian selama ini yang menunjukkan bahwa algortima K-Means memiliki unjuk kerja yang sangat baik dalam proses klusterisasi, seperti hasil penelitian yang dipublikasikan oleh Dyah Herawatie pada penelitian yang berjudul Perbandingan Algoritma Pengelompokan Non-Hierarki untuk Dataset Dokumen [3]. Dalam penelitian ini, data yang dipakai untuk eksperimen adalah artikel media masa yang berbahasa Indonesia yang diambil dari website Kompas dan Detik, dimana pada penelitian ini diperoleh hasil bahwa algoritma pengelompokan yang memberikan hasil yang terbaik adalah K-Means.

Seiring dengan semakin banyaknya artikel dalam bahasa Bali yang dapat dengan mudah ditemukan di internet, maka diperlukan juga penelitian yang mengarah kepada

pengelompokan artikel bahasa Bali. Hal ini tentu akan berguna bagi masyarakat yang ingin mencari artikel dalam bahasa Bali dan sistem bisa menemukan artikel tersebut dengan mudah karena proses penemuan artikelnya telah sesuai dengan pencarian yang diinginkan oleh pengguna. Penelitian yang dilakukan ini difokuskan pada proses pengelompokan artikel menggunakan algortima K-Means dengan berdasarkan pada kata kunci yang dimiliki oleh setiap artikel. Hasil dari penelitian ini selanjutnya dapat digunakan sebagai dasar pencarian semantik berbasis web untuk menemukan artikel khususnya artikel dalam bahasa Bali.

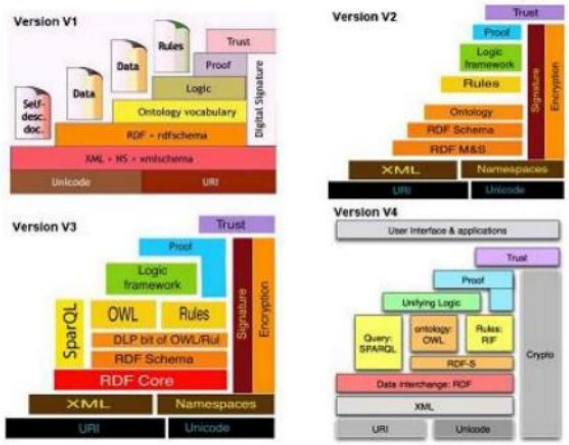
Bahasa Bali

Bahasa Bali merupakan bahasa austronesia yang menjadi bahasa sehari-hari dalam pergaulan masyarakat di daerah Bali. Dalam perjalanan waktu dan perkembangan teknologi, bahasa Bali tidak hanya digunakan sebagai alat komunikasi saja, tetapi juga sebagai sarana untuk melakukan abstraksi pemikiran dalam bidang yang sangat luas [4]. Hal ini dapat dilihat dari munculnya tulisan-tulisan dalam jaringan internet yang memuat berbagai bahasan dan tertuang dalam bahasa Bali.

Dalam penerapannya, bahasa Bali memiliki tingkatan saat digunakan dalam percakapan atau penulisan, dimana ada kelompok bahasa Bali halus, bahasa Bali madya dan bahasa Bali kasar, dimana dalam pemanfaatannya, pengguna perlu melihat konteks pembahasan dan audiens yang diajak berkomunikasi untuk memilih penggunaan bahasa Bali halus, madya atau kasar. Saat diimplementasikan, bahasa Bali didasarkan pada kata dasar dimana kata dasar (kruna lingga) adalah kata yang belum mendapatkan atau mengalami proses pengimbuhan maupun pengulangan. Konsep inilah yang menjadi dasar penentuan kata kunci pada sebuah artikel yang akan diangkat dalam penelitian ini.

Web Semantik

Web semantik merupakan generasi web terbaru yang memiliki tujuan untuk otomasi, integrasi, dan penggunaan kembali data pada aplikasi web yang berbeda. Web semantik adalah perluasan dari World Wide Web dengan teknik baru dan standar terhadap interoperation dan pemahaman oleh komputer. Semantik sendiri berarti ilmu yang mempelajari makna dan perubahan makna, sehingga makna dari suatu data tidak hanya bisa dipahami oleh manusia tetapi juga oleh mesin dimana dalam hal ini adalah aplikasi web. [5]. Berikut adalah komponen utama dalam web semantik :



Gambar 1. *Komponen Semantik Web* [5]

Algoritma K-Means

K-Means clustering merupakan algoritma *supervised learning* yang populer digunakan untuk mendapatkan dekripsi dari sekumpulan data dengan cara menemukan kecenderungan setiap data untuk berkelompok dengan individu data lainnya. Kecenderungan pengelompokan tersebut didasarkan pada kemiripan karakteristik individu-individu data yang ada. Ide dasar dari teknik ini adalah menemukan pusat dari setiap kelompok data yang mungkin ada untuk kemudian mengelompokkan setiap data individul kedalam salah satu dari kelompok-kelompok tersebut berdasarkan jaraknya [6]. Dalam penentuan kelompok berdasarkan jaraknya dapat menggunakan beberapa teknik penghitungan jarak antara lain :

1. Euclidean Distance

Euclidean Distance adalah metrika yang paling sering digunakan untuk menghitung kesamaan dua vektor. Rumus euclidean Distance adalah akar dari kuadrat perbedaan 2 vektor (root of square differences between 2 vectors) [7].

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad \ldots\ldots(1)$$

2. Manhattan Distance

Manhattan Distance/City Block Distance, merupakan salah satu teknik yang sering digunakan untuk menentukan kesamaan antara dua buah obyek. Pengukuran ini dihasilkan berdasarkan penjumlahan jarak selisih antara dua buah obyek dan hasil yang didapatkan dari Manhattan Distance bernilai mutlak. Manhattan Distance melakukan perhitungan jarak dengan cara tegak lurus [8]. Dalam koordinat cartesian,

jika p dan q merupakan dua titik dalam euclidean n-space yaitu :

$$d_1(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|, \quad \text{.....(2)}$$

Maka dalam ruang dimensi dua, Manhattan distance antar (p1,p2) dan (q1,q2) adalah :

$$| p_1 - q_1 | + | p_2 - q_2 | \quad \text{.....(3)}$$

3. Canberra Distance

Untuk setiap nilai 2 vektor yang akan dicocokkan, Canberra Distance membagi absolute selisih 2 nilai dengan jumlah dari absolute 2 nilai tersebut [7]. Hasil dari dua nilai yang dicocokkan lalu dijumlahkan untuk mendapatkan Canberra Distance. Jika koordinat nol-nol((0,0)) diberikan definisi dengan 0/0=0. Canberra Distance ini sangat peka terhadap sedikit perubahan dengan kedua koordinat mendekati nol.

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| - |x_{jk}|} \quad \text{.....(4)}$$

2. Pembahasan

Dalam makalah ini, jumlah data yang digunakan adalah 45 data artikel yang diperoleh dari berbagai sumber. 45 artikel ini terdiri dari 30 artikel murni bahasa Bali dan 15 artikel umum yang mengandung beberapa kata dalam bahasa Bali. Data artikel tersebut kemudian akan diproses terlebih dahulu dalam teknik pencarian teks menggunakan teknik *Heuristic Matching* [9]. Proses yang dilakukan adalah membandingkan teks artikel dengan teks yang telah ada sebelumnya pada database, dan dilakukan proses perbandingan dengan konsep *N-gram Similarity* dimana proses pengukuran menggunakan *Jaccard Coefficient*. Hasil akhirnya adalah nilai koefisien kemiripan data dalam bentuk angka yang akan menjadi nilai variabel kata kunci (key) dari keseluruhan data dalam artikel.

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

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

Ngajen : Ng, ga, aj, je, en

Metode dan Data

Dalam makalah ini untuk melakukan proses pengelompokan artikel, terlebih dahulu akan melewati tahap pencarian kata kunci dan *noise*. Proses penemuan kata kunci dilakukan dengan cara membandingkan teks artikel dengan teks yang telah ada sebelumnya pada database menggunakan konsep *N-gram Similarity* dan diukur menggunakan *Jaccard Coefficient*. Hasil akhirnya adalah nilai koefisien kemiripan data untuk setiap kata pada artikel yang sesuai yang selanjutnya akan ditotal untuk menjadi jumlah nilai kata kunci. Kata lain yang tidak sesuai dengan database bahasa Bali, akan dihitung dan menjadi total nilai noise. Kedua nilai untuk setiap artikel ini yang kemudian akan digunakan untuk proses pegelompokan dengan menggunakan metode K-Means Clustering. Berikut adalah diagram alur pengembangan sistem :

Gambar 2. *Alur Pengembangan Sistem*

Dari dua string tersebut terdapat 5 model bigram yang sama yaitu : Ng, ga, aj, je, en. Selanjutnya akan dilakuan 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 menyimpan kata Ngajeng dengan skor 1.

Jumlah data artikel yang digunakan dalam makalah ini berjumlah 45 artikel, yang akan dibagi menjadi lima buah variabel yaitu ID Artikel, Genre Artikel, Jumlah Halaman Artikel, Jumlah Key, dan Jumlah Noise. Berikut adalah contoh 10 data artikel yang digunakan dalam penelitian ini :

Tabel 1. Data Artikel

Artikel	Jumlah Kata
1	531
2	450
3	312
4	511

5	422
6	437
7	553
8	215
9	467
10	538

Selanjutnya, setelah data artikel dikumpulkan, tahap berikutnya adalah menemukan jumlah kata kunci dan noise dari sebuah artikel. Berikut adalah contoh artikel bahasa Bali yang digunakan :

Wénten satua "I Belog", sakéwanten durung wénten satua "I Ajum". Minab benjangan wénten satua "I Ajum". Menawi pacang wénten sané makarya satua "I Belog Ajum" utawi "I Ajum Belog".

Belog pangus adung mapasangan sareng ajum dados belog ajum. Belog ajum sampun dados klompok kruna sané ngwangun pangertian tunggal. Belog ajum sampun sakadi pasangan kruna sané tan dados kapasahang, tan dados kabalikang, ajum belog.

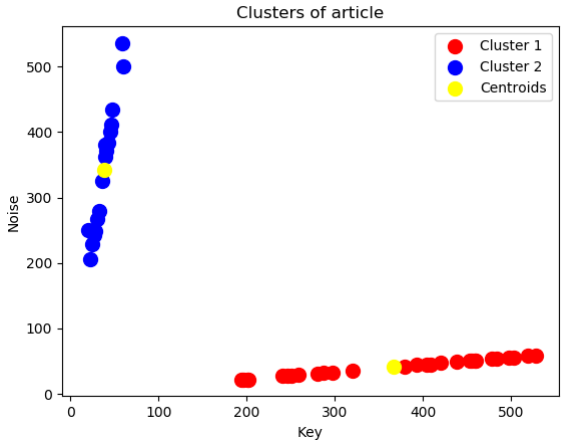
Wénten belog-belogan, wénten ajum-ajuman. Belog-belogan mawit saking kruna belog lan ajum-ajuman mawit saking kruna ajum. Belog-belogan maartos malaku sakadi anak belog. Ajum-ajuman maartos nyinahang parilaksana ajum, bonggan, sombong.

Dari data artikel awal yang digunakan, kemudian dicari jumlah kata kunci dan jumlah *noise* sehingga dihasilkan data artikel hasil pencarian sebagai berikut :

Tabel 2. Data Artikel Hasil Pencarian

Artikel	Jumlah Key	Jumlah Noise
1	478	53
2	405	45
3	281	31
4	460	51
5	380	42
6	393	44
7	498	55
8	194	22
9	420	47
10	480	54

Tahap selanjutnya adalah proses pengelompokan data artikel sesuai atribut yang tersedia. Proses pengelompokan menggunakan algoritma K-Means Clustering menggunakan *Euclidean Distance*. Hasil pengelompokan dapat dilihat pada grafik berikut ini :



Gambar 3. *Grafik Cluster Artikel*

Dari gambar terlihat bahwa artikel dibagi menjadi dua bagian, yaitu artikel yang memiliki kedekatan dengan kata kunci dan artikel yang lebih dekat dengan *noise*. Dari 45 data yang digunakan, dihasilkan 12 artikel yang tergabung dalam kelompok biru serta 33 artikel yang berada pada kelompok merah. Artikel yang memiliki kedekatan dengan kata kunci (merah) adalah kelompok artikel yang memiliki kesamaan data dengan data bahasa Bali yang dimiliki sistem. Artikel yang memiliki kedekatan dengan *noise* (biru) adalah kelompok artikel yang jumlah kata tidak dikenali atau tidak sama dengan sistem lebih banyak dari jumlah yang sesuai dengan sistem.

3. Kesimpulan

Proses yang dilakukan untuk pencarian teks artikel merupakan landasan awal untuk pengembangan pencarian artikel berbasis semantik. Penerapan teknik *N-gram Similarity* pada sistem pencarian teks bahasa Bali memiliki kemampuan yang cukup baik dimana pada penelitian ini, hasil pengelompokan artikel melalui algoritma *K-Means Clustering* menunjukkan presentase 73%. Nilai akurasi ini bisa semakin baik dengan memperbaiki data bahasa Bali yang dimiliki oleh sistem sehingga pada proses pencarian teks semakin tinggi peluang kata yang dikenali oleh sistem. Pada pengembangan selanjutnya, penelitian dapat dilakukan dengan penerapan metode *weighted tree similarity* untuk pencarian artikel bahasa Bali berbasis semantik.

Daftar Pustaka

[1]. A. Zaini, M. A. Muslim, and W. Wijono, "Pengelompokan Artikel Berbahasa Indonesia Berdasarkan Struktur Laten Menggunakan Pendekatan Self Organizing Map," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 6, no. 3, 2017.
[2]. A. Ambarwati and E. Winarko, "Pengelompokan Berita Indonesia Berdasarkan Histogram Kata Menggunakan Self-Organizing Map," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 8, no. 1, pp. 101–110, 2014.
[3]. D. Herawatie, "Perbandingan Algoritma Pengelompokan Non-Hierarki untuk Dataset Dokumen," *Semin. Nas. Apl. Teknol. Inf. Yogyakarta*, pp. 11–16, 2014.

[4]. I. K. Sudarsana, "Menumbuhkan Minat Belajar Bahasa Bali Pada Kalangan Remaja," *Pros. Sembada 2017*, no. 1, pp. 81–86, 2017.
[5]. K. Dwi, P. Novianti, R. Aurelius, N. Diaz, "Sistem Pencarian Program Studi Pada Perguruan Tinggi," *J. Sains dan Teknol.*, vol. 6, no. 1, pp. 93–104, 2017.
[6]. K. R. Prilianti and H. Wijaya, "Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering," *J. Cybermatika*, vol. 2, no. 1, pp. 1–6, 2014.
[7]. S. R. Wurdianarto, S. Novianto, and U. Rosyidah, "Perbandingan Euclidean Distance Dengan Canberra Distance Pada Face Recognition," *Techno.COM*, vol. 13, no. 1, pp. 31–37, 2014.
[8]. A.A.Ngr Wisnu Gautama, Yudha Purwanto, "Analisis Pengaruh Penggunaan Manhattan Distance Pada Algoritma Clustering Isodata (SelfOrganizing Data Analysis Technique) Untuk Sistem Deteksi Anomali Trafik," *e-Proceeding Eng.*, vol. 2, no. 3, pp. 7404–7411, 2015.
[9]. R. Aurelius, N. Diaz, "Analisis Data Teks Menggunakan Metode Heuristic Matching Studi Kasus : Teks Bahasa Bali," *Teknoif*, vol. 4, no. 2, pp. 7–10, 2016.