Algoritma Clustering untuk Seleksi Ciri pada Kategorisasi Terjemah Hadits

Arief Fatchul Huda, Qonita Ummi Safitri, Firda Ayu Setiawati
Mathematics Department
UIN Sunan Gunung Djati
Bandung, Indonesia
afhuda@uinsgd.ac.id, qonita.safitri@gmail.com, firdaayu_setiawati@yahoo.com

Abstrak - Permasalahan dalam kategorisasi teks adalah besarnya jumlah ciri. Jumlah ciri tergantung pada jumlah kata yang digunakan dalam seluruh dokumen yang diolah. Masalah lain adalah banyaknya elemen nol dalam ciri untuk tiap dokumen. Dalam penelitian ini penulis mengajukan metode untuk memilih (menyeleksi) ciri dengan menggunakan proses klastering. Algoritma klastering yang digunakan adalah Partition Around Medoid (PAM). Kategorisasi dilakukan dengan menggunakan algoritma k-NN dan Nearest Centroid. Penggunakan Seleksi Ciri dapat meningkatkan akurasi kategorisasi, untuk k-NN sebesar 3% dan untuk NC sebesar 6%. Selain itu juga pemilihan ciri dapat mengurangi waktu komputasi.

Keywords—Seleksi CIri; Kategorisasi ; k-Neirest Neighbor (k-NN); Clustering; k-Medoid; Neirest Centroid (NC).

I. INTRODUCTION

Text mining memiliki peranan seperti data mining, sehingga metode dalam teks mining pun mengadopsi dari metode-metode data mining, seperti clustering dan klasifikasi (kategorisasi). Text clustering adalah proses mengumpulkan dokumen-dokumen yang serupa pada cluster yang sama berdasarkan suatu tingkat keserupaan yang disebut proximity. Dalam clustering, label dokumen sebelumnya tidak diketahui. Clustering sering disebut unsupervised learning, karena cluster hanya diperoleh berdasarkan karakteristik internal data[1]. Sedangkan klasifikasi text atau kategorisasi teks adalah menempatkan suatu dokumen pada kelas yang telah terdefinisi sebelumnya, berdasarkan isi dokumen tersebut [2]. Konsep dasar dari kategorisasi teks adalah memasukkan teks baru yang belum diketahui kategorinya dengan melakukan pelatihan pada data yang telah diketahuii kategorinya.

Metode kategorisasi teks ini dilakukan melalui algoritma machine learning, seperti k-NN (k Neirest Neighbor) dan Nearest Centroid. k-NN melakukan pelatihan dengan mengukur kedekatan data uji pada seluruh data latih, kemudian dipilih k tetangga terdekat untuk mengetahui mayoritas kelas data latih terhadap data uji. Sedangkan Nearest Centroid hanya membandingkan data uji dengan center dari setiap kelas.

Namun, untuk melakukan kategorisasi teks, membutuhkan komputasi yang mahal, karena sangat bergantung pada banyaknya ciri yang ditemukan pada dokumen. ciri dari dokumen adalah banyaknya kosa kata (term) unik yang ditemukan dari kumpulan dokumen. Akibatnya, semakin banyak dokumen yang diolah, maka semakin banyak pula kosa kata (term) yang ditemukan. Untuk mengurangi biaya komputasi, maka diperlukan suatu metode untuk mengurangi jumlah ciri dokumen.

Penelitian ini mencoba untuk mengurangi jumlah ciri dengan melakukan clustering pada kumpulan term yang ditemukan. Teknik clustering yang dipilih adalah k-Medoid, agar setiap medoid dapat diambil sebagai term baru yang akan digunakan sebagai ciri terpilih untuk melakukan kategorisasi. k-Medoid merupakan salah satu teknik clustering partisi yang membagi *n* buah dokumen, menjadi *k* buah cluster, berdasarkan kedekatannya dengan *k* buah medoid terpilih yang menghasilkan total jarak minimal [1,3]. Kategorisasi dokumen akan dilakukan untuk hadits dalam Kitab Shahih al-Bukhary dengan k-Nearest Neighbor (k-NN) dan Nearest centroid (NC). Hasil penelitian ini diharapkan mendapatkan hasil akurasi kategorisasi yang baik berdasarkan ciri yang lebih sedikit.

II. Metode Klastering k-MEDOID dan Kategorisasi

A. Teknik Klastering k-Medoid

Klustering memiliki beberapa pendekatan, diantaranya adalah clustering partisi berbasis *center* (*centre-based*). Klustering partisi berbasis center adalah pembagian objek ke dalam kelompok yang saling lepas dan direpresentasikan oleh suatu "*center*". "*Center*" dalam cluster dapat berupa nilai rata-rata objek atau data aktual cluster. Jika *center* adalah rata-rata dari objek dalam cluster, maka disebut k-means, sedangkan jika center adalah data aktual, maka disebut k-medoid. Umumnya letak medoid berada di tengah cluster.

Adapun algoritma *k-medoid* yang paling umum adalah algoritma *Partitioning Around Medoids* (PAM). PAM melakukan pencarian medoid pada semua objek, sehingga semua objek memiliki kemungkinan dipilih menjadi medoid. Akibatnya, PAM akan menemukan cluster dengan global minimun secara tepat. Namun, hal itu membuat PAM kurang efektif untuk data berukuran besar, sehingga [4] mengajukan metode k-medoid yang lebih cepat. Metode tersebut mengadosi konsep k-Means dan PAM.

Algoritma PAM

1. Pilih *k medoid* awal
2. Hitung cost (jumlah total similarity terbesar semua dokumen terhadap semua medoid)
3. Ulangi langkah berikut ketika nilai cost bertambah:
   a. Untuk setiap medoid m:
      1) Untuk setiap dokumen non-medoid o:
         - ✓ Tukar m dengan o, hitung kembali cost
         - ✓ Jika nilai cost berkurang dari iterasi sebelumnya, batalkan langkah menukar m dengan o.

Algoritma K-MEDOID yang dimodifikasi

1. Pilih *k medoid* awal dan tetapkan setiap objek berdasarkan medoid terdekat.
2. Hitung total jarak semua objek terhadap medoid clusternya.
3. Ulangi hingga total jarak tidak berubah:
   a. Pada setiap cluster, pilih objek yang meminimalkan jarak setiap objek dalam cluster dengan medoidnya. Dan tetapkan sebagai medoid baru.
   b. Tetapkan semua objek pada medoid terdekat.
   c. Hitung total jarak tiap objek ke medoid baru.

Hasil klastering bersifat subjektive, sehingga memerlukan metode validasi untuk menentukan kualitas cluster yang dibentuk. Beberapa metode validasi hasil clustering adalah *Average within Cluster Distances*, dan *Davis Bouldin Index* [7,8,9,10,11]. *Average within Cluster Distances* didefinisikan sebagai jarak rata-rata *centroid* terhadap seluruh elemen pada *cluster*nya, dapat dihitung menggunakan (1).

$$Avg = \frac{\sum_{i=1}^{k}\left(\sum_{j=1, x_j \in C_i}^{n} \left\| x_j - c_i \right\|\right)}{n \cdot k} \quad (1)$$

Sedangkan *Davis Bouldin Index* didefinisikan sebagai metrik yang mengukur rata-rata kemiripan setiap *cluster* dengan *cluster* yang paling mirip dengannya, dapat dihitung dengan (2).

$$DB = \frac{1}{k}\sum_{i=1, i\neq j}^{k} \max\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) DB = \frac{1}{k}\sum_{i=1, i\neq j}^{k} \max\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \quad (2)$$

B. Metode Kategorisasi

Categorisasi atau klasifikasi merupakan proses 2 tahap dalam analisis data. Tahap pertama adalah learning step, yakni membangun model klasifier dengan menganalisa data training yang bersesuaian dengan label kategori. Tahap kedua, adalah menguji model terhadap data test, kemudian keakuratan model diperiksa dengan membandingkan kelas data test hasil kategorisasi dengan label aslinya. Oleh karena itu, kategorisasi sering disebut supervised learning [6]. Metode categorisasi telah banyak dikembangkan, diantaranya adalah k-NN dan Nearest centroid.

1) k-Nearest Neighbor (k-NN)

k-NN merupakan metode kategorisasi yang paling sederhana. k-NN menghitung jarak data uji pada setiap data latih, kemudian dipilih k buah tetangga terdekat. Lalu, data uji dilabeli berdasarkan mayoritas kelas tetangga terdekat itu. Parameter k diambil ganjil dan biasanya lebih dari satu.

Algoritma k-NN

1. Tentukan parameter k.
2. Hitung jarak data uji dengan semua data latih.
3. Pilih k jarak terdekat data uji terhadap data latih
4. Tentukan kelas dari data uji berdasarkan mayoritas kelas pada k kelas terdekat pada langkah 3.

2) Nearest Centroid

Nearert centroid (NC) adalah metode kategorisasi dengan membandingkan data uji dengan centroid data latih. Centroid tersebut diperoleh dari nilai rata-rata setiap kelas. Data uji akan dilabeli pada kelas yang memiliki jarak terdekat dengannya.

Algoritma NEAREST CENTROID

1. Hitung rata-rata setiap kelas pada data latih sebagai centroid setiap kelas
2. Hitung jarak data uji kepada setiap centroid.
3. Labeli data uji dengan label kelas centroid terdekat.

Untuk mengetahui keakuratan hasil kategorisasi dipilih beberapa metode validasi, yakni purity, f-measure dan akurasi manual. Nilai purity menyatakan berapa besar kecocokan hasil kategorisasi dengan kelas aslinya. Nilai purity dapat dihitung dengan (3).

$$\text{purity} = \sum_j \frac{n_j}{n} pur(j),\ pur(j) = \frac{\max_i (n_{ij})}{n_j},\ j = 1,2,...,k \qquad (3)$$

dimana n_{ij} adalah jumlah objek di kelas j yang berlabel i, n_j adalah jumlah objek di kelas j, dan n adalah jumlah seluruh objek.

F-Measure merupakan kombinasi harmonik dari nilai presisi dan recall. Presisi (p_{ij}) adalah ketepatan label hasil kategorisasi model dengan label model semula, sedangkan recall (r_{ij}) adalah tingkat keberhasilan model melabeli objek dengan benar. F-Measure dapat dihitung dengan (4).

$$F = \sum_i \frac{n_i}{n} \max_j \left(\frac{2 p_{ij} r_{ij}}{p_{ij} + r_{ij}} \right),\ p_{ij} = \frac{n_{ij}}{n_j},\ r_{ij} = \frac{n_{ij}}{n_i} \qquad (4)$$

Sedangkan akurasi manual hanya membandingkan objek yang berhasil dilabeli dengan benar dengan seluruh objek yang dilabeli. Nilai akurasi ini serupa dengan purity, dan dapat diperoleh dengan (5).

$$accurasy = \frac{total\_benar}{total\_objek} \times 100\% \qquad (5)$$

III. EXPERIMENT

Percobaan dilakukan untuk mengkategorisasi hadits dalam kitab Shahih al-Bukhary dengan k-NN dan Nearest Centroid (NC) pada matriks bobot dokumen. Hasil kategorisasi dengan ciri selection dan tanpa ciri selection akan dibandingkan untuk menentukan metode terbaik berdasarkan validasi nilai purity, f-measure dan accurasy. Metode ciri selection yang dipilih adalah clustering k-medoid pada ciri yang ditemukan dalam dokumen. Gambar 1 menunjukkan diagram alur percobaan yang dilakukan.

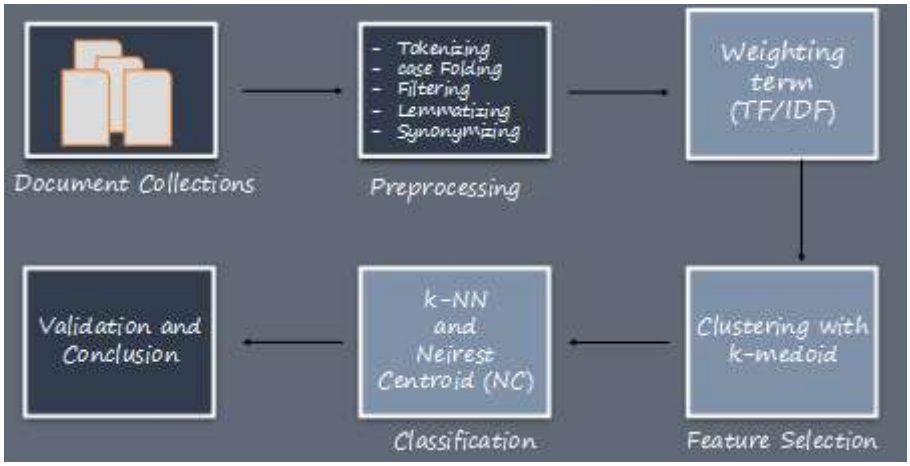

Fig. 1. Diagram Alur Percobaan

A. Dataset

Dataset yang digunakan adalah terjemahan kitab hadits Shahih al-Bukhary. Shahih al-Bukhary merupakan salah satu kitab hadits terbaik yang memuat sekitar 77 bab dengan 7563 hadits [12]. Namun, untuk penelitian ini hanya

menggunakan 5 bab dengan rincian seperti pada Tabel VI. Berdasarkan tabel VI, maka total data yang akan diolah adalah 539 dokumen yang terbagi menjadi 5 kelas.

B. Preprocessing

Preprocessing data dilakukan dengan bantuan Python versi 3.5, dengan mengalami beberapa tahap, yakni tokenizing, POS Tagging, Case folding dan Penghilangan Stop-word.

TABLE V
Proses Membaca Dokumen

Dok	Dataset Terbaca
D1	Bilal was ordered to repeat the wording of the Adhan for prayers twice, and to pronounce the wording of the Iqamas once except "Qad-qamat-is-Salat".
D2	Abu Barza said, "(O people!) Allah makes you self-sufficient or has raised you high with Islam and with Muhammad

D. Seleksi Ciri dengan Algoritma k-Medoid dan Kategorisasi

Tujuan dari ciri selection adalah mengurangi jumlah ciri, sehingga ciri minimal, tetapi mendapatkan akurasi maksimal. Oleh karena itu, pengurangan ciri dilakukan dari 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85% dan 90%. Dengan kata lain, term yang digunakan adalah 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45% dan 50%. Akibatnya, jumlah cluster k yang diuji disesuaikan dengan jumlah term yang digunakan. Hasil clustering divalidasi menggunakan Average Within Cluster Distance dan Davies Bouldin Index.

Untuk prosess kategorisasi, data train dan data test diambil secara merata dari data secara berurutan, yakni 50:50. Rincian pembagian data dapat dilihat pada Tabel VI.

TABLE VI
PEMBAGIAN DATASET EXPERIMENT

No Bab	Nama Bab	Hadist	Data Train	Data Test
1	Adzan	106	53	53
2	Holding into al-Quran and Sunnah	86	43	43
3	Knowledge	77	38	39
4	Tawheed	182	91	91
5	Wudhu	88	44	44
Total Document		***539***	***269***	***270***

Kategorisasi dilakukan menggunakan k-NN dan NC. Parameter yang dipilih untuk k-NN adalah k=5. Hasil Kategorisasi divalidasi menggunakan f-measure, purity dan akurasi manual.

IV. HASIL DAN DISKUSI

Hasil clustering menunjukkan bahwa, semakin banyak term yang digunakan (reduksi diperkecil), maka hasil cluster yang terbentuk semakin baik. Hal ini tampak dari nilai davies bouldin index dan average within cluster distance yang semakin mengecil untuk jumlah term yang meningkat. Namun, semakin banyak term yang digunakan (kluster semakin banyak) maka waktu komputasi semakin lama. Hasil clustering disajikan pada Tabel VII.

TABLE VII
VALIDASI HASIL CLUSTERING UNTUK ciri SELECTION

Term Used	Jumlah Ciri	Avrg. Distance	Davies Bouldin	Waktu(detik)
10%	262	0,0789	0,3202	2,3868
15%	393	0,0624	0,3201	4,0596
20%	524	0,0486	0,2542	4,9096
25%	655	0,0390	0,2429	5,7378
30%	786	0,0323	0,2306	5,6800
35%	917	0,0266	0,2252	6,4691
40%	1048	0,0225	0,2078	9,5109
45%	1179	0,0188	0,2047	10,5552
50%	1311	0,0156	0,1813	11,6203

Kategorisasi dilakukan pada dokumen dengan semua term dan term hasil reduksi untuk metode k-NN dan NC. Untuk kategorisasi k-NN dan NC yang menggunakan seluruh term, nilai validasi tidak berubah, sedangkan untuk kedua metode yang disertai ciri selection memiliki nilai yang beragam. k-NN yang disertai ciri selection menghasilkan nilai optimal jika menggunakan 30% term, sedangkan NC yang disertai ciri selection menghasilkan akurasi optimal dengan menggunakan 20% term saja. Selain itu, jika memandang rata-rata hasil akurasi dari semua metode yang digunakan pada percobaan ini, maka NC yang disertai ciri selection menghasilkan akurasi yang paling optimal. Selain itu, NC yang disertai ciri selection memiliki hasil eksekusi yang relatif lebih cepat dibandingkan metode lainnya. Hasil ini dapat dilihat pada tabel VIII.

TABLE VIII
HASIL VALIDASI PROSES KATEGORISASI PADA 5 BAB HADIST KITAB SHAHIH AL-BUKHARY

Metode	Akurasi	FMeasure	Purity	Waktu Eksekusi
NC	46,6667	49,3180	60,0000	0,0413
KNN	25,5556	36,0860	37,0370	1,7747
NC+FS	49,7119	52,6632	65,0617	0,0360
KNN+FS	31,3992	36,2275	48,1070	1,5211

Namun, hasil ini tidak menunjukkan bahwa pola tetap, karena sifat clustering yang hanya menemukan solusi optimal lokal. Akibatnya, setiap kali clusterting dilakukan, mungkin saja akan menghasilkan cluster yang berbeda, bergantung pada inisial medoid yang dipilih. Sebagai tambahan, clustering untuk jumlah cluster yang banyak un masih menjadi masalah.

V. KESIMPULAN

Clustering term dapat menjadi pilihan dalam metode ciri selection untuk menghasilkan akurasi yang baik pada proses kategorisasi dokumen. Clustering ciri selection yang dikombinasikan dengan NC menghasilkan akurasi yang lebih baik dibandingkan metode k-NN dan NC tanpa ciri selection. Namun, hasil percobaan menunjukkan waktu komputasi yang tidak berbeda secara signifikan pada proses kategorisasi antara menggunakan semua term ataupun menggunakan seleksi ciri.

REFERENCES

[1] P. N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining,, Addison Wesley, 2018.
[2] Fabrizio Sebastian, Text Categorization. Universita di Padova: Italy.
[3] L. Kauffman and P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, New York: John Wiley & Sons, 1990.
[4] H.S. Park., C.H. Jun, "A Simple and Fast Algorithm for K-medoids Clustering", *Expert Systems with Application*, vol. 36. Maret, 2009
[5] C. Luo, Y. Li, and S. M. Chung, "Text Document Clustering Based on Neighbors", *Data & Knowledge Engineering*, no. 68(11):1271–1288, 2009.
[6] Jiawei Han and Michelinde Kamber, Data Mining: Concepts and Techniques 2nd ed, San Fransisco: Morgan Kaufmann Publisher, 2006.
[7] Ichwanul Muslim Karo Karo, Kiki Maulana Adhinugraha, Arief Fatchul Huda. 2017. A cluster validity for spatial clustering based on davies bouldin index and Poligon Dissimilarity function. *Second International Conference on Informatics and Computing (ICIC)*, pg. 1-6.
[8] Ichwanul Muslim Karo Karo and Arief Fatchul Huda. 2016. Spatial Clustering for determining Rescue Shelter od Flood Disaster in South BAndung Using CLarans Algorithm with Polygon Dissimilarity Function. International Conference on Mathematics, Statistics, and Their Applicationns (ICMSA) 12th , Banda Aceh, Indonesia.
[9] Bernard Desgraupes. 2013. Clustering Indices. University Paris Puest.
[10] S. Al-Anazi, H. Al-Mahmoud, and Isra Al-Turaiki, "Finding Similar Documents using Different Clustering Techniques", dipresentasikan pada *Symposium on Data Mining Applications* (SDMA2016), Riyadh, Saudi Arabia, 30 Maret 2016.
[11] D. L. Davies, and D. W. Bouldin, "A Cluster Separation Measure", dipresentasikan pada *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI), 1979.
[12] Al-Bukhary, Shahih al-Bukhary, Lebanon: Darl Fikr, 2006.
[13] R. Feldman and S. James, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2006.
[14] F. Sebastian, Machine learning in automated text categorization, ACM Computing Surveys, Vol. 34 (1), 1-47, 2002.
[15] Pangestu Widodo, dkk, Klasifikasi Kategori Dokumen Berita Berbahasa Indonesi dengan Metode Kategorissi Multi-Label berbasis Domain-Specific Ontology. Vol 6 (2), 59-138, 2017.
[16] Riu Xu and Donld C. Wunsch II, Clustering, A John Wiley &Sons, Inc., Publication, 2009.
[17] Bachler MArtin, ciri Selection in Machine Learning, 2005.
[18] Xiaofei Zhou, Text Categorization Based on Clustering ciri Selection, 2014.