Algoritma K-Means *Clustering* Untuk Pengelompokan Ayat Al Quran Pada Terjemahan Bahasa Indonesia

Miftachur Robani*[a], Achmad Widodo[b]

[a] Magister Sistem Informasi Universitas Diponegoro
[b]Fakultas Teknik Universitas Diponegoro

Naskah Diterima : 11 Juni 2016; Diterima Publikasi : 30 Juli 2016

DOI: 10.21456/vol6iss2pp164-176

Abstract

Clustering process can make the process of grouping data so that the data in the same cluster have high similarity with the data in the same cluster. One of the clustering algorithm that is widely used is the K-Means because it has advantages such as simple, efficient, easy to understand and easy to apply. Grouping paragraph dealing with similar themes will allow users to find a theme in the Qur'an. This study aims to produce an information system that can perform grouping Quran with K-Means method. This research was conducted with a pre-processing stage process for text data, weighting by TFIDF, grouping data with K-Means clustering, labeling data for keywords. The resulting system is able to display a verse in groups associated with the keyword. The test results by using the index on the silhouette of Surah Al Fatihah generate positive value of 0.336 which means that the data in the right group, while the frequency of keywords versus the amount of data to produce a percentage of 53%, which means the keyword represents half of the data in the cluster. Tests also showed that the test results silhouette will be directly proportional to the number of clusters and inversely proportional to the number of data dimensions. To increase the value of testing required centroid method for early elections, the reduction of data dimensions and methods of measurement of distance and similarity.

Keywords : Clustering, K-Means, Al Quran, Silhoutte *etection, Recognition, Haar-Like Feature, ANN Backpropagation*

Abstrak

Proses Clustering dapat melakukan proses pengelompokan data sehingga data pada klaster yang sama memiliki kesamaan yang tinggi dengan data di klaster yang sama. Salah satu algoritma clustering yang banyak digunakan adalah K-Means karena memiliki kelebihan antara lain sederhana, efisien, mudah dipahami dan mudah diterapkan. Pengelompokan ayat yang memiliki kemiripan tema akan memudahkan pengguna menemukan suatu tema dalam Al Quran. Penelitian ini bertujuan menghasilkan suatu sistem informasi yang dapat melakukan pengelompokan ayat Al Quran dengan metode K-Means. Penelitian ini dilakukan dengan tahapan proses pra pemrosesan untuk data teks, pembobotan dengan TFIDF, pengelompokan data dengan K-Means clustering, pelabelan data untuk kata kunci. Sistem yang dihasilkan mampu menampilkan ayat dalam kelompok yang sesuai dengan kata kunci. Hasil pengujian dengan menggunakan silhouette index pada Surat Al Fatihah menghasilkan nilai positif sebesar 0,336 yang artinya data pada kelompok yang tepat, sedangkan dari frekuensi kata kunci berbanding jumlah data menghasilkan presentase 53% yang artinya kata kunci merepresentasikan setengah dari data dalam klaster. Pengujian juga menunjukkan bahwa hasil pengujian silhouette akan berbanding lurus dengan jumlah klaster dan berbanding terbalik dengan jumlah dimensi data. Untuk meningkatkan nilai pengujian diperlukan metode untuk pemilihan centroid awal, reduksi dimensi data dan metode pengukuran jarak dan kemiripan.

Kata kunci : Clustering, K-Means, Al Quran, Silhoutte

1. Pendahuluan

Ayat dalam kitab suci Al Quran merupakan objek menarik bagi ilmuwan komputer untuk menunjukkan pengetahuan, kearifan dan hukum dari ayat Al Quran di dalam sistem komputer. Dengan membangun sistem cerdas yang dapat menjawab berbagai macam pertanyaan berdasarkan pengetahuan dari ayat dan dapat membantu masyarakat, baik muslim maupun non muslim untuk memahami dan mengerti ayat dalam Quran (Atwell *et al.*, 2011).

Memahami maksud ayat dengan membaca tafsir (penjelasan detil dari maksud ayat) akan sangat membantu tetapi belum cukup memberikan gambaran utuh dari pesan yang kitab ini coba untuk sampaikan kepada pembaca. Hal ini dikarenakan Al Quran

*) Penulis korespondensi: miftachurrobani@gmail.com

mencakup satu tema di banyak surat yang berbeda dan untuk mendapat gambaran utuh, pembaca harus merujuk semua bagian yang saling berhubungan (Abbas, 2009).

Teks klasik agama adalah salah satu sasaran utama penggunaan *text mining*. Secara komputasi, buku seperti Quran memiliki informasi semi terstruktur karena diatur dalam struktur nomor surat dan ayat. Ini memudahkan pemodelan, berbeda dengan teks tidak terstruktur seperti novel atau biografi (Ahmad, 2013).

Berbagai metode telah digunakan untuk melakukan pengelompokan pada suatu data tertentu salah satunya *clustering*. *Clustering* merupakan metode analisis data yang penting dan algoritma ini dapat diklasifikasikan sebagai pengelompokan hirarki dan pengelompokan partitional. Sebagai metode klasifikasi terawasi, *clustering* membagi satu set objek ke dalam kelompok individu yang sama. Hal ini banyak digunakan untuk pengenalan pola, komputasi biologi, ilmu atmosfer, segmentasi gambar, analisis dokumen teks, diagnosis medis dan lain sebagainya (Wu, 2015).

Clustering teks adalah bagian yang penting dalam metode text mining, dan juga merupakan bagian dari data mining. Clustering teks adalah klasifikasi dokumen tanpa pengawasan, yang membagi koleksi teks menjadi beberapa subset yang disebut klaster, teks masing-masing klaster memiliki kesamaan yang lebih besar daripada yang berada dalam klaster yang berbeda. Clustering secara khusus sangat berguna untuk mengorganisir dokumen untuk meningkatkan penemuan kembali informasi dan mendukung proses browsing (Aggarwal, 2012).

Algoritma K-Means merupakan algoritma *clustering* partisi terbaik yang paling dikenal. K-Means barangkali juga paling luas dipakai pada algoritma *clustering* karena sederhana dan efisien. Dengan memberikan kumpulan data point dan jumlah k klaster yang diinginkan, algoritma ini akan mengulangi partisi data ke k klaster berdasarkan fungsi jarak. Kelebihan utama dari algoritma k-means adalah sederhana, efisien, mudah dipahami dan mudah diterapkan. Kompleksitas waktunya adalah O(tkn) dengan n adalah jumlah data, k adalah jumlah klaster dan t adalah jumlah iterasi. Dengan k dan t yang lebih kecil daripada n, algoritma k-means adalah algoritma yang linier dengan jumlah data (Liu, 2007). Sedangkan pada *clustering* hirarki, kompleksitas waktu adalah kuadratik $O(n^2)$ karena mengukur jarak dari seluruh data ke data lain (Steinbach *et al*., 2000).

Kualitas dari sebuah metode data mining seperti klasifikasi dan clustering sangat tergantung dengan proses penghilangan gangguan dari pola yang digunakann dalam proses clustering. Maka diperlukan proses pra-pemrosesan seperti pemisahan kata dari dokumen (*tokenization*), penghilangan kata yang sering muncul namun tidak relevan (*stopword removal*) dan pengubahan kata menjadi kata dasar (*stemming*). Dan setiap kata akan dilakukan representasi dengan metode pembobotan berdasarkan frekuensi muncul kata yaitu TF-IDF (Aggarwal, 2012).

Dari latar belakang tersebut dirumuskan masalah yaitu penggunaan metode *Clustering* dengan Algoritma K-Means untuk pengelompokan ayat-ayat Al Quran pada terjemahan Bahasa Indonesia. Dokumen akan dilakukan tahapan pra-pemrosesan dan ditentukan bobot berdasarkan frekuensi muncul pada proses *clustering*. Sehingga diperoleh klaster yang berisi ayat-ayat yang memiliki kemiripan dengan tema tertentu (keimanan, ibadah atau lainnya). Dan judul yang diambil adalah Algoritma K-Means *Clustering* untuk Pengelompokan Ayat Al Quran pada Terjemahan Bahasa Indonesia.

2. Kerangka Teori

2.1. Penyajian Ayat

Penyajian ayat dalam kelompok berdasarkan tema yang sama diyakini lebih mudah dipahami bagi pengguna. Pendekatan berdasarkan ontologi memperlihatkan bahwa ayat dapat diklasifikasi dan ditampilkan ke pengguna secara sistematis. Ontologi digunakan untuk menyajikan ayat dalam bentuk sistematis dan terstruktur dengan pemetaan maksud tema pada ayat yang sesuai dan yang memiliki relasi dengan ayat tersebut, contohnya adalah pada tema iman memiliki sub tema yaitu iman kepada Allah yang ada di ayat 21 surat ke 2 (Ta'a *et al.*, 2013).

Ayat dapat diorganisir menggunakan ontologi yang dipakai untuk menampilkan struktur ayat secara sederhana. Dengan pengelompokan ini akan memungkinkan pengguna untuk menemukan informasi tentang ayat lebih cepat dan mengurangi kebingungan bagi pembaca non Arab (Ksasbeh, 2009).

Perbandingan dua pendekatan yaitu agglomerative hierarchical clustering dan K-Means. Clustering hirarkis sering digambarkan sebagai pendekatan clustering kualitas yang lebih baik, tetapi terbatas karena kompleksitas waktunya kuadrat. Sebaliknya, K-Means kompleksitas waktu yang linier dengan jumlah dokumen, tetapi diperkirakan menghasilkan kualitas cluster rendah. Namun, hasil penelitian ini menunjukkan bahwa teknik K-Means lebih baik dari pendekatan hirarki yang diuji untuk berbagai metrik evaluasi cluster (Steinbach *et al*, 2000).

Clustering k-means dapat juga digunakan pada text berbahasa China. Dapat dibuktikan bahwa algoritma ini adalah benar dan efektif. Meskipun hasil pengelompokan k-means telah baik, tetapi hasil keseluruhan tidak memuaskan, alasannya adalah karakteristik yang berbeda dari arti kata tersebut diasumsikan berbeda, dan ini justru merupakan faktor penting yang mengarah ke hasil tidak ideal (Yao *et al.*, 2009).

Terbatasnya penelitian *text mining* dalam bahasa asing menjadi suatu tantangan bagi peneliti untuk secara efektif mengelola data dan melakukan

klasifikasi informasi yang relevan bagi pengguna. Pendekatan ini mengintegrasikan *clustering* dokumen k-means dengan ekstraksi fitur semantik dan vektorisasi dokumen menjadi kelompok halaman web berbahasa Arab menurut kesamaan semantik. Vektorisasi dokumen membantu untuk mengubah dokumen teks ke dalam distribusi probabilitas kelas semantik atau kepadatan kelas semantik (Alghamdi, 2014).

Metode untuk meningkatkan interaksi waktu browsing adalah dengan pendekatan scatter-gather. Pendekatan ini menampilkan keyword yang berhubungan dengan keyword lain kepada pengguna. Pengguna bisa memilih satu keyword yang berhubungan dengan satu atau lebih klaster (Aggarwal, 2012).

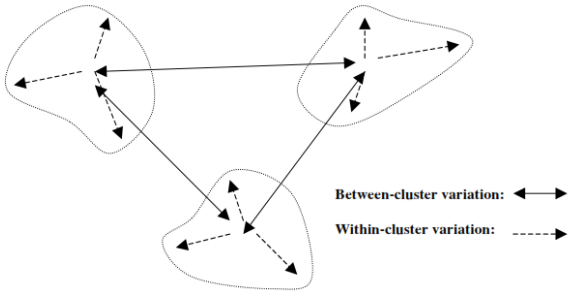
Pembangunan sistem managemen pengetahuan dengan pendekatan *clustering* dapat dilakukan untuk ekstraksi pengetahuan dari penulisan publikasi. Dengan menggunakan metode *clustering* k-means dapat membantu proses *organizing*, *filtering*, *browsing* dan *searching* pengetahuan. Dengan k-means rata-rata akurasi sebesar 89,13% dan kelengkapan dokumen kembali sebesar 85,73% (Pulukadang, 2014).

2.2. Clustering

Algoritma Clustering mengelompokan satu set dokumen ke dalam himpunan bagian atau klaster. Tujuan algoritma klaster adalah menciptakan klaster yang koheren secara internal, tetapi jelas berbeda satu sama lain. Dengan kata lain, dokumen dalam sebuah klaster harus semirip mungkin; dan dokumen dalam satu klaster harus sebeda mungkin dari dokumen dalam klaster lainnya (Manning, 2008).

Clustering merujuk pada pengelompokan dokumen, observasi atau kasus pada kelas yang objeknya mirip. Klaster adalah kumpulan dokumen yang mirip satu sama lain dan berbeda dengan dokumen pada klaster lain. Clustering berbeda dengan Clasification, pada clustering tidak ada target variabel untuk dikelompokkan. Algoritma clustering mencoba untuk membagi kumpulan data menjadi klaster yang anggotanya relatif sama, dimana kemiripan dokumen di klaster yang sama tinggi, dan kemiripan dokumen di klaster lain kecil. Dengan kata lain, seperti pada Gambar 1, Algoritma clustering mencoba membuat klaster dokumen yang between-clusters variation (BCV) lebih tinggi dibandingkan dengan within-cluster variation (WCV) (Larose,2005).

Pada penelitian ini, data yang akan digunakan adalah data teks berbahasa Indonesia yang diambil dari basis data terjemahan Al Quran berbahasa Indonesia versi Departemen Agama yang sudah dalam bentuk digital. Maka data yang akan diolah bersifat tidak terstruktur, oleh karena itu perlu adanya tahapan pra-pemrosesan sebelum dilakukan *clustering*. Tahapan pra-pemrosesan terdiri dari *tokenization*, *stopword removal*, *stemming* (Darawat, 2010) dan menggunakan pembobotan pada setiap kata di seluruh dokumen menggunakan skema TF.IDF (*term frequency-inverse document frequency*) (Ahlgren dan Colliander, 2009).



Gambar 1. Klaster memiliki BCV lebih tinggi daripada WCV (Larose,2005)

- *Tokenization*
  Proses tokenization berguna untuk memecah setiap kalimat dari seluruh dokumen pengetahuan ke dalam kata-kata (term) dengan menggunakan pembatas tab dan karakter spasi (Darawaty, 2010). Hal yang perlu dilakukan juga adalah menjadikan kata menjadi huruf kecil menghilangkan karakter tanda baca seperti tanda titik(.), koma (,), petik(“’), kurung(()), tanda tanya(?), tanda seru(!) dan tanda baca lainnya. *RFID Tag* : adalah device yang menyimpan informasi untuk identifikasi objek. *RFID tag* sering juga disebut sebagai *transponder*. *Tag* yang digunakan pada penelitian ini adalah *tag* bertipe kartu seukuran kartu nama.
- *Stopword removal*
  Kualitas metode data mining seperti clustering sangat berpengaruh pada proses penghilangan noise yang digunakan pada proses clustering. Sebagai contoh kata yang sering digunakan seperti "the", mungkin tidak berguna untuk meningkatkan kualitas clustering. Dengan demikian, penting untuk memilih feature secara efektif sehingga kata-kata noise dapat dihilangkan sebelum clustering. Cara paling sederhana untuk pemilihan kata pada clustering dokumen adalah penggunaan frekuensi dokumen untuk menyaring kata yang tidak relevan. Dengan kata lain, kata yang sering muncul di dokumen dapat dihilangkan karena merupakan kata biasa seperti "a", "an", "the" dan "of" yang tidak cukup beragam dari segi clustering (Aggarwal, 2012). *Stopword* adalah kata-kata yang sering muncul dalam suatu dokumen yang kurang berguna dalam proses penggalian text. Proses *Stopword removal* yang berguna menghilangkan stopword, merupakan proses yang sangat penting dalam text mining. Dalam penelitian yang berbahasa Indonesia, *stopword* yang digunakan misalnya adalah “yang”, “seperti”, “merupakan”, “adalah”, “sebuah” dan lain-lain (Pulukadang, 2014).

- *Stemming*
  Proses stemming berguna untuk merubah suatu kata menjadi kata dasarnya, misalnya kata 'mendapatkan' menjadi 'dapat'. Stemming akan meningkatkan klasifikasi teks dalam bahasa tertentu, pada Bahasa Indonesia, stemmer telah banyak dikembangkan(Arifin *el al.*, 2010).
- TF.IDF
  Tahap terakhir dalam pre-pemrosesan adalah pembobotan setiap kata menggunakan TF.IDF (term frequency-inverse document frequency) (Ahlgren dan Colliander, 2009) dengan menggunakan persamaan 1

$$w_{m,i} = freq_{m,i} \times log\left(\frac{N}{nm}\right) \quad (1)$$

Dengan $w_{m,i}$ adalah bobot setiap term (m) terhadap setiap dokumen(i), freq $_{m,i}$ adalah jumlah frekuensi kemunculan term m pada setiap i, N adalah jumlah seluruh dokumen i, dan nm adalah jumlah i yang terdapat kemunculan m.

Pada representasi TFIDF, Term Frequency (TF) dinormalisasikan dengan Inverse Document Frequency (IDF). Normalisasi IDF mengurangi bobot kata yang muncul pada koleksi data. Ini akan mereduksi kata penting yang muncul pada koleksi data, menjamin dokumen yang cocok memiliki pengaruh lebih daripada kata lain yang relatif rendah frekuensinya di koleksi data (Aggarwal, 2012).

Setelah tahapan pra-pemrosesan selesai, maka akan didapatkan bobot setiap term yang dapat dilakukan proses text mining. Setelah itu dilanjutkan dengan penggunaan *clustering* dengan algoritma K-Means.

2.3. K-Means Clustering

K-means adalah algoritma *clustering* untuk menemukan kelompok dari objek yang *non-overlapping* (Wu, 2012). K-Means juga dianggap sebagai algoritma yang efektif untuk mengelompokkan suatu data (Larose, 2005). K-means adalah algoritma clustering dalam bidang data mining. Hal ini digunakan untuk cluster analisis, dan memiliki efisiensi tinggi pada partisi data terutama dalam dataset besar. Sebagai algoritma pembelajaran tidak terawasi, kita tidak tahu hasil klaster sebelum mengeksekusi algoritma, tidak seperti klasifikasi. Karena jumlah klaster tidak diketahui, sehingga biasanya menggunakan jumlah kelompok yang diinginkan sebagai masukan, dan dalam aplikasi nyata, kitaumumnya memutuskan itu berdasarkan eksperimen (Yao *et al.*, 2013).

K-means merupakan algoritma yang sangat sederhana berdasarkan kemiripan. Ukuran kesamaan memainkan peran penting dalam proses clustering. Data yang mirip dijadikan ke cluster yang sama, dan yang berbeda dalam cluster yang berbeda. Biasanya digunakan Euclidean Distance untuk mengukur kesamaan antara dua titik data. Metode metrik yang berbeda untuk pengukuran kemiripan tidak akan mengubah hasilnya, tetapi hasil dari K-Means lebih sensitif terhadap centroid awal. Dua faktornya adalah: satu adalah nilai K, dan lain adalah pemilihan nilai awal centroid. K-Means menerapkan teknik berulang. Proses ini tidak akan berhenti sampai nilai rata-rata dari semua klaster tidak berubah. Dalam algoritma K-means, pemilihan pusat awal adalah kunci untuk mendapatkan hasil yang tepat. Jika memilih awal yang tepat centroid akan mendapatkan hasil yang baik, tetapi jika tidak, hasilnyaakan bertambah buruk, hal itu mungkin membuat kepadatan besar dan klaster dibagi menjadi potongan-potongan, atau menggabungkan dua cluster dekat menjadi satu kelompok. Jadi kita biasanya memilih awal centroid secara acak (Yao *et al.*, 2013).

Adapun tahapan yang dilakukan dalam algoritma K-Means adalah :

1. Penentuan nilai k

Proses pertama adalah menginisialisasi nilai awal k sebagai jumlah kluster yang akan dipartisi. Salah satu cara untuk menentukan k adalah dengan menggunakan *rule of thumb* (Mardia *et al.*, 1979) yaitu dengan persamaan 2.

$$k \approx \sqrt{n/2} \quad (2)$$

Nilai n adalah jumlah objek yang akan diklaster. Nilai k adalah jumlah klaster yang akan dipartisi. Persamaan lain untuk menentukan nilai k pada basis data teks (Can dan Ozkarahan, 1990 *dalam* Mardia et al., 1979) adalah :

$$k \approx \frac{m \times n}{t} \quad (3)$$

Dalam menentukan nilai k diperlukan jumlah objek/dokumen (n), jumlah term (m) dan jumlah record bobot yang mempunyai nilai lebih dari 0 (t) (Pulukadang, 2014).

2. Penentuan pusat klaster awal

Menentukan secara acak bobot yang akan menjadi pusat kluster sebanyak jumlah k yang sesuai dengan tahap pertama. Salah satu masalah pada algoritma K-Means adalah beberapa klaster mungkin menjadi kosong selama proses clustering karena tidak ada data yang menempatinya. Klaster tersebut disebut klaster kosong. Untuk mengatasi klaster kosong, dapat dipilih data point sebagai pengganti centroid, data point yang paling jauh dari centroid pada klaster yang besar (Liu, 2007).

3. Pengukuran jarak

Menentukan jarak bobot pada masing-masing dokumen yang bukan pusat klaster dengan bobot setiap pada masing-masing dokumen pusat kluster menggunakan jarak Euclidean(d).

$$d_m = \sqrt{\sum_{i=1}^{N} \left(x_{m,i} - y_{m,i}\right)^2} \quad (4)$$

Dengan $d_m$adalah jarak dari setiap bobot (m), i adalah setiap dokumen, N adalah jumlah dokumen, $x_{m,i}$

adalah record pada m terhadap setiap i yang bukan pusat klaster dan $y_{m,i}$ adalah record pada m terhadap setiap i yang termasuk pusat klaster.

4. Penentuan jarak terpendek

Setelah mendapatkan jarak antar record dengan pusat klaster, maka tentukan jarak (d) yang bernilai minimum pada setiap dokumen untuk menjadi anggota klaster.

5. Penentuan pusat klaster baru

Setelah menghasilkan klaster dan anggotanya pada iterasi pertama, dihitung kembali nilai baru pusat klaster atau centroid dengan membagi bobot pada klaster yang sama.

$$nilai\ centroid = \sum \frac{a_i}{c_k} \quad (5)$$

Dengan a_i adalah record i terhadap setiap dokumen yang terpilih menjadi anggota claster pada tahapan 4 dan c = jumlah anggota klaster yang terbentuk pada tahapan 4.

6. Penghentian iterasi

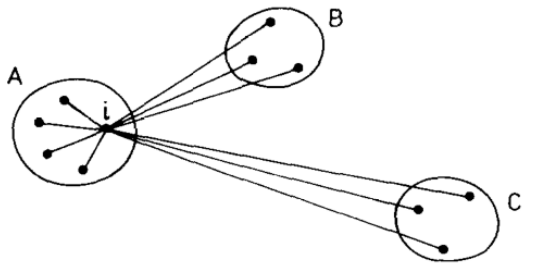
Mengulangi tahap 3-6 sampai nilai centroid atau anggota klaster sudah tidak berubah. Sehingga didapatkan klaster yang berisi dokumen yang mirip.

7. Penentuan Label

Klaster yang ada harus diberi label yang nantinya akan menjadi nama/tema klaster tersebut. Hal yang dapat dilakukan untuk memperoleh label adalah dengan menemukan kata yang paling sering diulang dalam masing-masing klaster. Kata yang paling sering disebut akan menjadi kata kunci dari klaster tersebut.

2.4. Pengujian Klaster

Pengujian kualitas klaster dapat menggunakan Silhouette index yang menggunakan lebar silhoutte pada masing-masing entitas. Untuk menghitung lebar silhoutte, digunakan rata-rata jarak terkecil ke entitas di klaster lain dan rata-rata jarak ke entitas lain di klaster yang sama digunakan. Perhitungan lebar silhouette menghasilkan nilai antara -1 dan 1. Nilai yang mendekati 1 menandakan entitas tersebut berada di klaster yang tepat. Namun jika nilai mendekati -1 menandakan entitas tersebut di klaster yang salah (Storlokken, 2007). Ilustrasi Silhoutte pada Gambar 2.



Gambar 2. Ilustrasi Silhoutte index (Rousseeuw, 1987)

Seperti pada Gambar 2.. Misalkan didefinisikan s (i) dalam kasus dissimilariti. Tentukan i dalam kumpulan data, dan dilambangkan dengan A cluster yang telah ditetapkan. Ketika klaster A berisi objek lain selain i, maka dapat dihitung (i) = perbedaan rata-rata i untuk semua data lain dari A. Pada Gambar 2.2, ini adalah panjang rata-rata semua data dalam klaster A. Untuk setiap klaster C yang berbeda dari A, dapat dihitung d (i, C) = perbedaan rata-rata i untuk semua objek dari C. Adapun rumus untuk menghitung silhoutte adalah (Rousseeuw, 1987):

$$s(i) = \frac{\mathrm{b\ (i)} - \mathrm{a(i)}}{\max\{\mathrm{a(i)}, \mathrm{b(i)}\}} \quad (2.6)$$

dengan a(i) adalah rata-rata jarak antara entitas i ke entitas lain dalam klaster, dan b(i) adalah rata-rata jarak minimum ke entitas di klaster lain.

2.5. Information Retrieval

Pencarian informasi di web merupakan dasar dalam information retrieval, sebuah cabang studi yang membantu pengguna untuk menemukan informasi yang dibutuhkan dari koleksi besar pada dokumen teks. Menemukan kembali (*retrieving*) informasi bermakna menemukan kumpulan dokumen yang relevan dengan query pengguna. Query pengguna yang paling sering digunakan adalah dalam format kata kunci (Liu, 2007). Metode yang dapat dipakai untuk *browsing* adalah pendekatan yang menampilkan kata kunci yang beragam untuk pengguna. Sehingga pengguna dapat memilih salah satu kata kunci yang berhubungan dengan satu atau lebih klaster (Aggarwal,2012).

Pengukuran keektifan sistem temu kembali dapat dilakukan dengan *precision* dan *recall*. *Precision* adalah jumlah dokumen ditemukan yang relevan dengan kata kunci. *Recall* adalah jumlah dokumen relevan yang ditemukan, adapun persamaanya adalah (Manning, 2008) :

$$Precision = \frac{dokumen\ relevan\ ditemukan}{dokumen\ ditemukan} \quad (7)$$

$$Recall = \frac{dokumen\ relevan\ ditemukan}{dokumen\ relevan} \quad (8)$$

3. Metodologi

3.1. Bahan dan Alat Penelitian

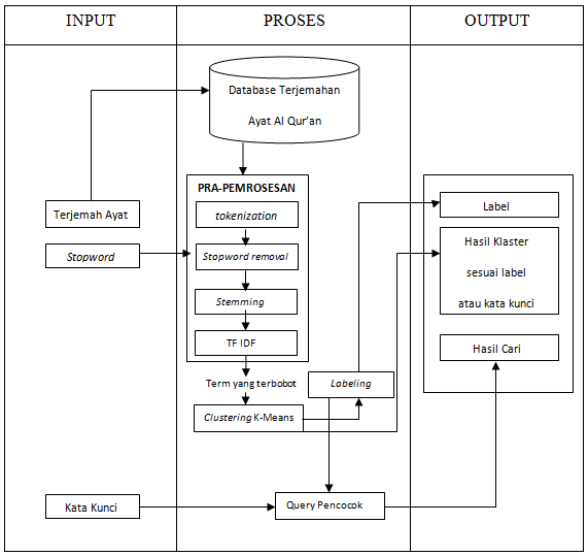
Penelitian ini menggunakan bahan yaitu data dari terjemahan ayat-ayat Al Quran dalam Bahasa Indonesia versi Kementrian Agama yang sudah didigitalisasi dalam bentuk basis data SQL. Basis data Al Quran didapatkan dengan mengunduh dari qurandatabase.org, sebuah web penyedia basis data Al Quran dengan 104 bahasa terjemahan. Basis data Al Quran berisi 114 surat dengan jumlah total ayat adalah 6236 ayat. Setiap surat memiliki jumlah ayat yang berbeda. Setiap ayat juga memiliki panjang yang berbeda. Alat yang digunakan untuk perancangan sistem adalah perangkat keras komputer dengan spesifikasi processor Intel Core i3 dan memori 2 GB. Sedangkan perangkat lunak yang digunakan adalah Sublime Text untuk editor PHP dan Navicat Premium untuk pengolahan basis data MYSQL.

3.2. Prosedur Penelitian

Penelitian ini akan dimulai dengan identifikasi masalah yaitu pengelompokan ayat, dilanjutkan dengan perancangan kerangka sistem informasi, implementasi dan pengujian. Penggunaan clustering K-Means akan menjadi solusi untuk masalah pengelompokan ayat. Pengelompokan ayat akan menjadikan proses penelusaran pada Al Quran menjadi lebih mudah dengan menghadirkan ayat-ayat yang memiliki kesamaan kata kepada pengguna. Maka dilakukan studi tentang penggunaan K-Means pada dokumen text, sehingga untuk penelitian ini didapatkan gambaran bagaimana cara kerja dan hasil dari K-Means untuk pengelompokan dokumen.

3.3. Kerangka Sistem Informasi

Sistem yang dibangun akan dimulai dengan memasukkan data terjemahan ayat ke dalam basis data, data berupa nomor ayat, nomor surat dan teks terjemahan ayat Al Quran dalam Bahasa Indonesia. Kemudian data ayat akan dilakukan proses prapemrosesan meliputi proses tokenisasi yaitu perubahan kalimat dalam ayat menjadi kata-kata terpisah, penghilangan stopword yakni kata-kata yang sering muncul namun tidak relevan, proses stemming yaitu perubahan menjadi kata dasar dan proses pembobotan menggunakan TFIDF yang berdasarkan kemunculan kata pad masing-masing ayat. Setelah setiap data memiliki bobot TFIDF maka dilakukan proses *clustering* menggunakan K-Means. Proses K-Means sendiri terdiri dari penentuan nilai k klaster, penentuan nilai awal centroid, penentuan jarak dengan Euclidean, penentuan jarak minimum dan penentuan klaster. Setelah data terkelompok maka dilakukan proses pelabelan untuk masing-masing klaster. Kerangka sistem informasi yang digunakan dapat dilihat pada Gambar 3.



Gambar 3. Kerangka Sistem Informasi

Kemudian sistem akan menampilkan hasil klaster kepada pengguna. Pengguna juga bias melakukan penelusuran ayat Al Quran berdasarkan label yang telah dihasilkan sistem. Label akan digunakan untuk proses pencarian kelompok ayat yang sesuai untuk pengguna. Pengguna juga bisa melakukan proses pencarian ayat dengan memasukkan keyword dan sistem akan mencocokan kata kunci dengan label yang ada. Label yang sesuai dengan keyword akan ditampilkan kepada pengguna. Kemudian pengguna daapat memilih label yang telah terhubung dengan klaster yang diwakilinya.

3.4. Implementasi

Ada dua sisi sistem yang akan dibangun yakni dari perancangan untuk pengelompokan oleh admin dan sisi penelusuran dan pencarian ayat untuk pengguna. Ada dua tahapan utama dalam sistem perancangan pengelompokan yaitu tahapan pra-pemrosesan dan tahapan clustering. Tahapan pra-pemrosesan terdiri dari *tokenization* yaitu proses perubahan kalimat pada ayat menjadi kata-kata terpisah sehingga dapat dilakukan proses selanjutnya, *stopword removal* yaitu penghilangan kata-kata yang sering muncul namun tidak relevan dengan dokumen, *stemming* yakni proses pengubahan semua kata menjadi kata dasar. dan pembobotan TF.IDF yang berdasarkan kemunculan kata pada tiap-tiap ayat. Sedangkan tahapan *clustering* terdiri dari penentuan jumlah k klaster yang didapat dengan menggunakan persamaan (2) dan (3), penentuan nilai awal centroid sejumlah k yang ditentukan dengan cara mengacak dan memilih dari bobot sejumlah k data yang dihasilkan dari pembobotan TFIDF, penghitungan jarak Euclidean antara semua data dengan nilai pusat centroid yang ditentukan, penghitungan jarak minimum antara jarak data ke masing-masing klaster, penentuan klaster yang berdasarkan nilai minimum pada jarak Euclidean dan penghitungan centroid baru untuk iterasi selanjutnya dengan membagi semua nilai bobot dari klaster yang sama. Setelah proses clustering selesai maka akan dilanjutkan dengan pelabelan untuk kata kunci masing-masing klaster.

Sedangkan sistem dari sisi pengguna akan disiapkan kata kunci untuk masing klaster yang telah dihasilkan dari proses pelabelan. Selain itu pengguna akan diberikan form untuk mengisi kata kunci yang ingin dicari yang kemudian sistem akan mengembalikan klaster yang memiliki kata kunci yang cocok. Kedua sistem ini, baik untuk pengelompokan dan pencarian oleh pengguna akan dibangun dengan menggunakan PHP dan MYSQL. Sedangkan keluaran dari sistem akan ditampilkan ke pengguna menggunakan web.

3.5. Pengujian

Aplikasi yang telah dibangun akan diuji kualitas klasternya menggunakan silhoutte index dengan Persamaan (6) dengan rentang hasil 1 dan -1, jika hasil mendekati 1 berarti klaster yang terbentuk sudah baik. Sedangkan pengujian penulusuran dokumen

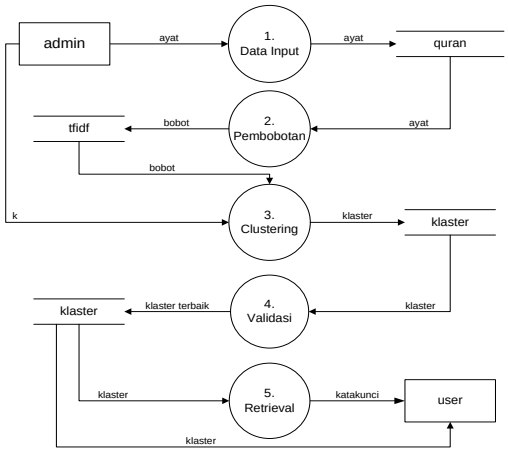
menggunakan *precision* dan *recall* dengan Persamaan (7) dan (8) dengan hasil berupa presentase, semakin tinggi nilai *precision* dan *recall* berarti hasil semakin baik.

3.6. Perancangan Sistem

Sistem yang dirancang sesuai dengan kebutuhan yang telah dijelaskan sebelumnya akan direpresentasikan dalam bentuk data flow diagram level 1 untuk menjelaskan proses secara umum yang terjadi. Pada Data Flow Diagram (DFD) Level 1 akan dijabarkan proses dan aliran data yang terjadi pada proses Pengelompokan ayat Al Quran. Proses yang terlibat meliputi : proses input data, proses pembobotan, proses clustering, proses validasi dan proses retrieval. Data Flow Diagram (DFD) level 1 terlihat di Gambar 4. Uraian masing-masing proses sebagai berikut :

a. Pada proses input data, admin akan memasukan data ayat yang berisi nomor surat , nomor ayat dan teks terjemahan masing-masing ayat dan selanjutnya disimpan dalam tabel quran.
b. Pada proses pembobotan TFIDF, data teks terjemahan ayat Al Quran dari tabel quran akan dilakukan pembobotan berdasarkan kemunculan kata yang menghasilkan bobot yang disimpan di table tfidf.
c. Pada proses clustering, data bobot setiap dokumen akan dilakukan pengelompokan menjadi k klaster yang telah ditentukan sebelumnya, proses ini akan menghasilkan klaster untuk masing dokumen yang disimpan di tabel klaster.
d. Pada proses validasi, hasil klaster akan dilakukan penghitungan validitas sehingga didapatkan klaster terbaik yang disimpan dalam tabel klaster.
e. Pada proses retrieval akan dimunculkan kata kunci dari masing-masing klaster yang terbentuk untuk pengguna akhir sistem.

Pengguna akhir akan mendapatkan hasil klaster dan kata kunci untuk mengakses kelompok ayat yang dihasilkan sistem.



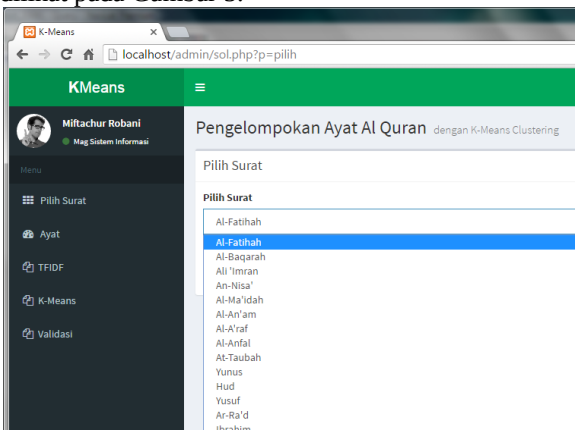
Gambar 4. *Data Flow Diagram level 1*

4. Hasil dan Pembahasan

4.1. Prapemrosesan

1. Tahapan pilih surat

Pada tahapan pilih surat, admin akan diberikan pilihan data surat mana yang akan diproses clustering. Sistem akan menampilkan 114 surat yang ada di dalam Al Quran yang dapat dipilih untuk tahapan selanjutnya. Tampilan halaman pilih surat dapat dilihat pada Gambar 5.



Gambar 5. Tampilan halaman pilih surat

Setelah memilih surat yang diinginkan, admin dapat melihat ayat pada menu Ayat. Misalkan surat yang dipilih adalah surat pertama, Surat Al Fatihah. Maka ayat yang akan dilakukan tahapan pra-pemrosesan adalah seperti pada Tabel 1.

Tabel 1. Ayat pada Surat Al Fatihah

No Ayat	Terjemah Ayat
1	Dengan menyebut nama Allah Yang Maha Pemurah lagi Maha Penyayang.
2	Segala puji bagi Allah, Tuhan semesta alam.
3	Maha Pemurah lagi Maha Penyayang.
4	Yang menguasai di Hari Pembalasan.
5	Hanya Engkaulah yang kami sembah, dan hanya kepada Engkaulah kami meminta pertolongan.
6	Tunjukilah kami jalan yang lurus,
7	(yaitu) Jalan orang-orang yang telah Engkau beri nikmat kepada mereka; bukan (jalan) mereka yang dimurkai dan bukan (pula jalan) mereka yang sesat.

2. Tahapan pra pemrosesan

Setelah dilakukan pemilihan surat yang akan dilakukan proses clustering, tahapan selanjutnya adalah dilakukan tokenisasi untuk menjadikan ayat menjadi kata per kata, kemudian dari masing-masing kata akan dihilangkan stopwordnya.

Stopword ditemukan dengan cara menghitung jumlah kemunculan kata. Kemudian data jumlah kemunculan kata diurutkan dari kata yang terbanyak

muncul, kata yang kemunculannya besar akan menjadi kandidat stopword. Pada data yang digunakan ditemukan sejumlah 878 kata yang dijadikan sebagai daftar kata yang akan dihilangkan pada tahapan pra pemrosesan. Adapun beberapa stopword yang ditemukan seperti ditunjukkan pada Tabel 2.

Tabel 2. Data Stopword

No	Kata	Jumlah Muncul
1	yang	9410
2	dan	8171
3	mereka	6008
4	orang	5904
5	kamu	3892
6	kami	2837
7	itu	2818
8	kepada	2629
9	tidak	2321
10	sesungguhnya	2182
11	di	1937

Setelah semua stopword yang ditemukan pada ayat dihilangkan maka proses selanjutnya adalah stemming yaitu perubahan kata menjadi kata dasar. Pada Surat Al Fatihah, hasil stemming seperti pada Tabel 3.

Tabel 3. Hasil stemming Surat Al Fatihah

No Ayat	Terjemah Ayat	Hasil Stemming
1	Dengan menyebut nama Allah Yang Maha Pemurah lagi Maha Penyayang.	allah murah sayang
2	Segala puji bagi Allah, Tuhan semesta alam.	puji allah tuhan semesta alam
3	Maha Pemurah lagi Maha Penyayang.	murah sayang
4	Yang menguasai di Hari Pembalasan.	kuasa balas
5	Hanya Engkaulah yang kami sembah, dan hanya kepada Engkaulah kami meminta pertolongan.	sembah tolong
6	Tunjukilah kami jalan yang lurus,	tunjuk jalan lurus
7	(yaitu) Jalan orang-orang yang telah Engkau beri nikmat kepada mereka; bukan (jalan) mereka yang dimurkai dan bukan (pula jalan) mereka yang sesat.	jalan nikmat jalan murka jalan sesat

3. Tahapan pembobotan TFIDF

Setelah dilakukan proses perubahan kata dasar kemudian dilakukan pembobotan. Proses pertama yang dilakukan adalah menghitung kemunculan kata atau term frequency (TF) untuk masing-masing dokumen berdasarkan kata yang telah dijadikan kata dasar pada proses stemming. Adapun hasil TF terlihat pada Tabel 4.4. Proses selanjutnya adalah menghitung inverse document frequency (IDF) yang merupakan normalisasi frekuensi kata. Untuk menghitung IDF kata "murah" yang kemunculannya(nm) di 2 ayat dan jumlah dokumen (n) adalah 7, maka dengan Persamaan (2.1) didapatkan :

$$IDF\ (murah) = log(\frac{nm}{n}) = log(\frac{2}{7}) = \ 0{,}544$$

Langkah selanjutnya adalah menghitung nilai TFIDF untuk masing-masing dokumen yang dihasilkan dari nilai TF dikalikan dengan nilai IDF.

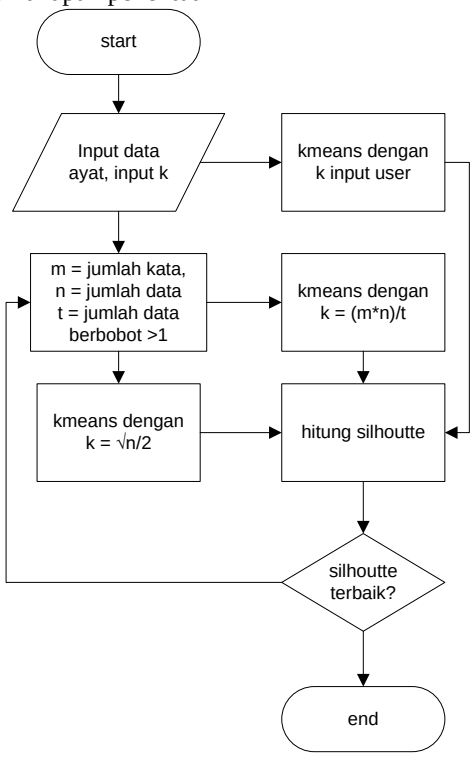
Pada perhitungan TFIDF ditemukan bahwa untuk setiap dokumen yang memiliki panjang berbeda akan mempengaruhi nilai bobot TFIDF. Untuk mengatasi hal ini, perlu dilakukan proses normalisasi. Adapun yang digunakan adalah cosine normalization dengan Persamaan 4.1 (Manning *et al.,* 2008).

$$\frac{1}{\sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}} \qquad (9)$$

dengan w adalah bobot setiap kata pada dokumen yang sama

4.2. Clustering

1. Tahapan penentuan k



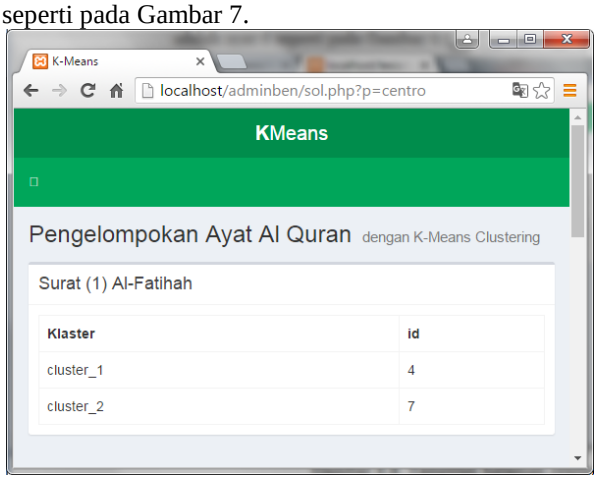
Gambar 6. Flowchart penentuan nilai k

Terlihat pada Gambar 4.3 bahwa untuk menentukan k dengan 3 macam cara yaitu dengan input dari user, dengan menggunakan rule of thumb (Persamaan 2.2) dan menggunakan jumlah bobot (Persamaan 2.3). Setelah satu tahap dengan salah satu k selesai, maka

akan dilakukan proses perhitungan silhouette. Kemudian berlanjut dengan k selanjutnya dimana k dengan nilai silhouette terbaik akan dipilih untuk output.

2. Tahapan penentuan centroid awal

Centroid atau pusat klaster awal dilakukan dengan cara acak. Dengan melakukan random centroid awal setiap sistem dijalankan memungkinkan output yang dihasilkan juga berbeda. Dengan demikian pada kasus terbaik akan ditemukan centroid yang tepat dengan hasil validasi yang tinggi, sebaliknya saat centroid awal yang ditemukan tidak tepat maka akan menghasilkan nilai validasi yang buruk juga. Pada tahapan penentuan centroid awal akan didapatkan hasil pengacakan nilai centroid awal sejumlah nilai k dari penentuan nilai k. Untuk surat Al Fatihah, pada sistem di halaman centroid awal akan menampilkan hasil acak dari sistem sejumlah k=2 yakni untuk klaster 1 adalah ayat 4 dan klaster 2 adalah ayat 7 seperti pada Gambar 7.



Gambar 7. Tampilan halaman centroid awal

3. Tahapan penentuan jarak

Setelah didapatkan pusat klaster yaitu pada ayat 3 dan ayat 6 kemudian dilakukan pengukuran jarak dengan menggunakan Euclidean dengan Persamaan (4). Sehingga didapatkan hasil seperti pada Tabel 4.

Tabel 4. Hasil pengukuran jarak

No ayat	cluster_1	cluster_2
1	0.6943419	1.2861672
2	0.8992827	1.2861672
3	0.7435548	1.2861672
4	0.9378019	1.2861672
5	0.9378019	1.2861672
6	1.1311376	0.5880254
7	1.1311376	0.5880254

Setelah diketahui jarak dari masing-masing ayat ke pusat klaster maka kemudian akan ditentukan nilai minimum jarak untuk masing-masing ayat sebagai dasar keanggotaan klaster. Misalkan pada ayat 1, jarak ke klaster 1 adalah 0,694 sedangkan ke klaster 2 adalah 1,286 dengan demikian ayat 1 akan masuk ke klaster 1.

4. Tahapan hasil klaster

Pada tahapan hasil klaster diperlihatkan letak masing-masing ayat di klaster mana, letak klaster ditentukan berdasarkan jarak minimum yang diperoleh pada penentuan jarak minimum. Hasil klaster terlihat pada Tabel 5.

Tabel 5. Hasil Klaster

No Ayat	Klaster
1	cluster_1
2	cluster_1
3	cluster_1
4	cluster_1
5	cluster_1
6	cluster_2
7	cluster_2

4.3. Retrieval

1. Tahapan pelabelan

Tahapan pelabelan akan memunculkan kata kunci untuk masing-masing klaster yang memiliki jumlah kemunculan tertinggi. Hasil untuk masing-masing klaster seperti pada Tabel 6.

Tabel 6. Hasil kata kunci

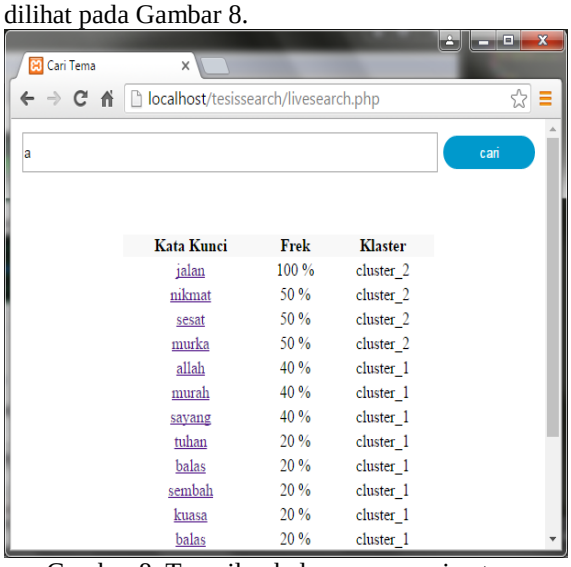
No	Kata kunci	frekuensi	klaster
1	murah	2	klaster 1
2	sayang	2	klaster 1
3	Allah	2	klaster 1
4	sembah	1	klaster 1
5	balas	1	klaster 1
6	jalan	2	klaster 2
7	tunju	1	klaster 2
8	murka	1	klaster 2
9	nikmat	1	klaster 2
10	lurus	1	klaster 2

Dengan didapatkan hasil pelabelan berupa kata kunci, akan menjadi rujukan saat pengguna memilih kata kunci “murah” misalnya, akan diberikan klaster 1. Sementara jika pengguna memilih kata kunci “jalan” akan diberikan klaster 2. Dengan demikian kata kunci ini akan memudahkan penulusuran ayat.

2. Tahapan pencarian tema

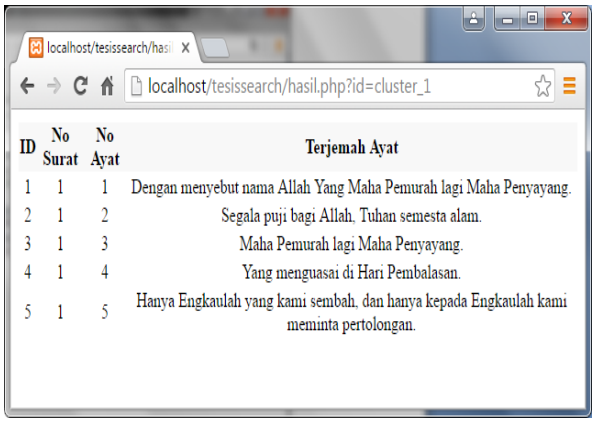
Pengguna dapat melakukan pencarian tema dengan menggunakan halaman pencarian tema yang telah disediakan sistem. Halaman ini akan menampilkan form pencarian kata yang dilengkapi dengan pencarian otomatis berdasarkan kata yang dimasukkan satu per satu. Kemudian sistem akan menampilkan kata kunci yang dihasilkan dari proses pelabelan. Setiap kata kunci akan terhubung dengan klaster yang diwakilinya. Dengan nilai presentase kata kunci dibandingkan dengan jumlah dokumen di

klaster yang sama. Tampilan halaman pencarian dapat dilihat pada Gambar 8.



Gambar 8. Tampilan halaman pencarian tema

Halaman hasil klaster untuk pencarian akan menampilkan klaster secara penuh dari masing-masing kata kunci yang telah dipilih. Tampilan halaman hasil klaster dari pencarian dapat dilihat pada Gambar 9.



Gambar 9. Tampilan halaman hasil klaster pencarian

4.4. Pengujian Hasil

Untuk mengetahui apakah sistem ini berhasil mengelompokkan data dengan benar maka akan dilakukan proses pengujian. Pengujian dilakukan secara internal dengan silhouette menggunakan Persamaan (6) dan dengan membandingkan ayat dengan kata kunci dengan jumlah ayat dalam klaster yang sama. Sedangkan secara eksternal menggunakan Persamaan (7) dan Persamaan (8). Adapun hasil perhitungan dengan silhouette pada Surat Al Fatihah dengan k=2 seperti pada Tabel 7. Dengan a adalah rata-rata jarak dokumen dengan anggota klaster yang sama, dan b adalah rata-rata jarak dokumen ke anggota klaster lain terdekat.

Tabel 7. Hasil silhouette Surat Al Fatihah

No Ayat	Klaster	Terjemah Ayat
		Klaster 1
1	Klaster 1	Dengan menyebut nama Allah Yang Maha Pemurah lagi Maha Penyayang.
2	Klaster 1	Segala puji bagi Allah, Tuhan semesta alam.
3	Klaster 1	Maha Pemurah lagi Maha Penyayang.
4	Klaster 1	Yang menguasai di Hari Pembalasan.
5	Klaster 1	Hanya Engkaulah yang kami sembah, dan hanya kepada Engkaulah kami meminta pertolongan.
		Klaster 2
6	Klaster 2	Tunjukilah kami jalan yang lurus,
7	Klaster 2	(yaitu) Jalan orang-orang yang telah Engkau beri nikmat kepada mereka; bukan (jalan) mereka yang dimurkai dan bukan (pula jalan) mereka yang sesat.

Dalam pengujian sistem didapatkan rata-rata nilai silhouette dokumen sebesar 0.336. Berdasarkan interpretasi nilai Silhoutte sebagaimana pada Tabel 8, maka struktur klaster masih termasuk kategori struktur yang lemah.

Tabel 8. Interpretasi nilai Silhoutte

Nilai Silhoutte	Interpretasi
0,71 – 1,00	Struktur kuat ditemukan
0,51 – 0,70	Struktur sedang ditemukan
0,26 – 0,50	Struktur lemah ditemukan
≤ 0,25	Struktur tidak substansial

Hasil proses clustering dan anggota masing-masing klaster dapat dilihat pada Tabel 9. Nilai frekuensi untuk masing-masing kata kunci dibandingkan dengan jumlah data pada klaster yang sama, misalnya kata “murah” muncul di dua ayat dari klaster 1 yang berisi lima ayat, maka diperoleh frekunsi per jumlah data adalah sebesar $^{2}/_{5}$ = 40%.

Tabel 9. Anggota Klaster

No ayat	Klaster	a	b	silhoutte
1	cluster_1	1.1262	2.0301	0.4453
2	cluster_1	1.2802	2.0301	0.3694
3	cluster_1	1.1952	2.0301	0.4113
4	cluster_1	1.2854	2.0301	0.3668
5	cluster_1	1.2854	2.0301	0.3668
6	cluster_2	1.2118	1.5581	0.2222
7	cluster_2	1.2880	1.5581	0.1733
		Rata-rata		0.3360

Maka pada surat Al Fatihah didapatkan hasil untuk masing-masing klaster dengan tiga kata kunci teratas seperti pada Tabel 10.

Tabel 10. Hasil perbandingan frekuensi kata kunci

No	Klaster	Kata kunci	Frekuensi/ Jumlah data
1	Klaster 1	murah	40%
2	Klaster 1	sayang	40%
3	Klaster 1	Allah	40%
4	Klaster 2	jalan	100%
5	Klaster 2	sesat	50%
6	Klaster 2	nikmat	50%
		Rata-rata	53%

Setelah dihitung didapatkan rata-rata sebesar 53% yang artinya setiap klaster memiliki kesamaan dalam rentang sedang. Dengan demikian, kata kunci akan merepresentasikan sebanyak setengah dari seluruh anggota klaster.

Sementara untuk perhitungan precision dan recall untuk tiga kata kunci teratas pada masing-masing klaster dengan menggunakan Persamaan (7) dan Persamaan (8). Precision didapatkan dengan membagi dokumen relevan yang ditemukan dengan jumlah seluruh dokumen yang ditemukan. Sedangkan recall didapat dari pembagian dokumen relevan yang ditemukan dengan jumlah dokumen relevan. Setelah dihitung didapatkan rata-rata precision adalah 53% dan recall 100%.

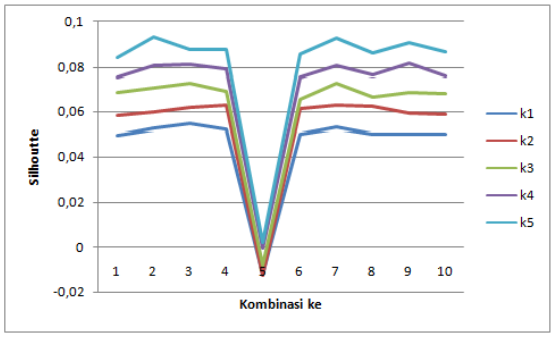
4.5. Pengujian Variabel

Untuk menentukan variabel apa sajakah yang mempengaruhi hasil pengujian silhouette, dilakukan percobaan dengan menentukan 10 kombinasi centroid awal seperti pada Tabel 11. Kombinasi didapatkan dari mengacak nomor id ayat pada Surat Al Baqarah yang dijadikan objek pengujian yaitu pada rentang 8 sampai 293. Masing-masing kombinasi ditentukan untuk 5 macam jumlah klaster (k) yaitu k1=10, k2=12, k3=14, k4=16 dan k5=18. Jadi dapat diketahui bagaimana pengaruh jumlah klaster dapat dilihat saat nilai k ditambah.

Tabel 11. Kombinasi centroid awal

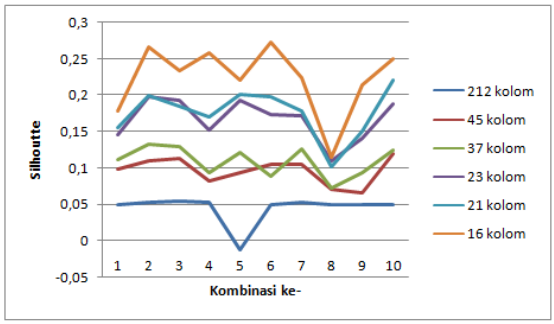
Jumlah Klaster (k)					Kombinasi (C) ke-									
k5	k4	k3	k2	k1	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
					58	64	37	51	8	12	30	22	15	21
					63	73	42	54	10	45	45	48	46	28
					86	108	74	78	41	58	62	90	55	42
					89	114	136	186	62	60	90	98	62	70
					154	140	156	202	73	73	98	131	72	133
					155	167	160	211	149	104	103	139	93	160
					181	220	181	222	158	176	135	229	142	216
					224	230	194	244	203	204	151	249	186	232
					270	239	229	245	234	232	207	273	199	251
					272	255	260	265	286	249	219	275	225	260
					178	228	66	221	208	122	217	193	164	240
					141	182	112	102	9	47	75	261	52	284
					50	118	161	40	127	81	129	83	264	36
					152	184	34	243	177	170	17	259	116	165
					195	67	33	223	187	258	172	71	271	210
					88	179	120	196	280	57	107	150	238	257
					82	109	145	100	124	183	235	132	84	53
					166	77	189	252	13	248	144	101	263	25

Adapun hasil sihoutte pada proses clustering K-Means dengan dimensi 212 kolom menggunakan 10 kombinasi untuk 5 buah jumlah klaster ini terlihat pada grafik dalam Gambar 10. Hasil menggambarkan semakin tinggi nilai k semakin baik hasil silhouette-nya.



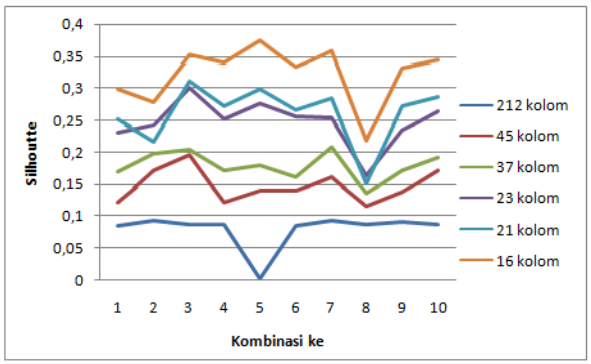
Gambar 10. Grafik silhouette dengan dimensi 212 kolom

Kemudian dilakukan percobaan dengan mengurangi jumlah dimensi kolom. Pada kombinasi C1 dengan k1=10 akan dicoba dengan 6 macam jumlah kolom yaitu 212 kolom, 45 kolom, 37 kolom, 23 kolom, 21 kolom dan 16 kolom seperti pada Gambar 11. Terlihat bahwa semakin sedikit jumlah kolom, hasil silhouette semakin tinggi.



Gambar 11. Grafik hasil silhouette dengan k1=10

Adapun dengan menaikan jumlah klaster dengan k5=18 maka hasil perhitungan silhouette seperti pada Gambar 12.



Gambar 12. Grafik hasil silhouette dengan k1=18

Seperti pada Gambar 12 terlihat bahwa hasil silhouette dapat dipengaruhi oleh jumlah klaster (k) dan jumlah dimensi data. Semakin jumlah klaster tinggi, hasil silhouette juga tinggi sedangkan jika jumlah kolom sedikit maka hasil silhouette semakin tinggi. Selain itu, penentuan centroid awal juga akan mempengaruhi hasil silhouette. Pada kasus terbaik, dengan kombinasi yang tepat maka akan dihasilkan silhouette yang tinggi, sebaliknya pada kasus terburuk akan didapatkan hasil yang buruk. Pada Gambar 4.9. terlihat pada kombinasi ke 5 awalnya memiliki nilai silhouette yang rendah, namun setelah jumlah kolom dikurangi maka menjadi kombinasi dengan silhouette tertinggi yaitu 0,3744. Sedangkan pada kombinasi ke 8 merupakan kasus terburuk karena hasil silhouette tetap rendah sebesar 0,218 walaupun jumlah kolom telah dikurangi.

5. Kesimpulan

Pengelompokan data ayat Al Quran dalam Bahasa Indonesia dengan menggunakan Algoritma K-Means akan menghasilkan kelompok ayat dengan kata kunci tertentu. Proses clustering dengan K-Means memiliki hasil pengujian silhouette pada Surat Al Fatihah bernilai positif sebesar 0,336 yang artinya data pada kelompok yang tepat. Frekuensi per jumlah data sebesar 53% yang artinya kelompok yang dihasilkan memiliki kemiripan yang sedang. Sedangkan untuk hasil perhitungan rata-rata precision sebesar 53% dan perhitungan rata-rata recall sebesar 100%. Pengujian juga menunjukkan bahwa hasil pengujian silhouette akan berbanding lurus dengan jumlah klaster dan berbanding terbalik dengan jumlah dimensi data. Pada kombinasi centoid awal yang tepat dengan jumlah klaster dan jumlah kolom yang tepat didapat silhouette 0,3744 pada Surat Al Baqarah.

Daftar Pustaka

Abbas, N.H, 2009. *Quran 'Search for a Concept' Tool and Website*, Thesis Master of Science, The University of Leeds.

Aggarwal C.C, Zhai C, 2012. *Mining Text Data*, Springer, New York.

Ahlgren, P. Colliander, C., 2009. Document-document similarity approaches and science mapping : Experimental comparison of five approaches. *Journal of Informetrics* 3. 49-63.

Ahmad, O., 2013. A Survey of Searching and Information Extraction on a Classical Text Using Ontology-based semantics modeling: A Case of Quran. *Life Science Journal*.

Alghamdi, H.M., 2014. Arabic Web Pages Clustering And Annotation Using Semantic Class Features, *Journal of King Saud University – Computer and Information Sciences* 26, 388–397.

Arifin, A.Z, Mahendra I., Ciptaningtyas H., 2010. Enhanced Confix Stripping Stemmer And Ants Algorithm For Classifying News Document In Indonesian Language, *The 5th International Conference on Information & Communication Technology and Systems*, pp 149-158.

Atwell, E., Dukes, K., Sharaf, A.-B., Louw, N. H. B., Shawar, B. A., McEnery, T., et al. 2010. Understanding the Quran: A new Grand Challenge for Computer Science and Artificial Intelligence. *Paper presented at the British Computer Society Workshop, Edinburgh*.

Darawaty, I, 2010. Intelegent Searching using Association Analysis for law Documents of Indonesian Government, *Second International Conference on Advances in Computing, Control and Telecomunication Technologies*, pp 122-124.

Ksasbeh M.Z., 2009. Using Ontology to Define the Structure of the Holy Quran, *4th International Conference on Information Technology*, Amman.

Larose, D.T., 2005. *Discovering Knowledge in Data : An Introduction to Data Mining*, Wiley-Interscience, New Jersey.

Liu B., 2007. *Web Data Mining*, Springer, New York.

Manning, C.D., 2008. *Introduction to Information Retrieval*, Cambridge University Press, New York.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. *Multivariate Analysis*. Academic Press, London.

Pulukadang D.R, 2014. *Pendekataan Clustering untuk Pengelolaan Pengetahuan pada Sistem Manajemen Pengetahuan*, Tesis Magister Sistem Informasi Undip.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* 20, pg 53-65.

Steinbach, M., Karypis, G., Kumar, V., 2000. *A Comparison of Document Clustering Techniques*, Technical Report of University of Minnesota, Minnesota.

Storløkken, R., 2007. *Labelling clusters in an anomaly based IDS by means of clustering quality indexes*, Thesis Master of Science in Information Security Gjøvik University College.

Ta'a, A., 2013. Al-Quran Themes Classification Using Ontology, *Proceedings of the 4th International Conference on Computing and Informatics, ICOCI 2013.*

Wu, X., Wu, B., Sun J., Qiu, S. and Li, X., 2015. A hybrid fuzzy K-harmonic means clustering algorithm, *Aplied mathematical Model*, 3398-3409.

Yao, Y., Liu, Y., Yu, Y., Xu, H., Lv, W., Li, Z. and Chen, X., 2013. K-SVM: An Effective SVM Algorithm Based on K-means Clustering, *Journal Of Computers*, 2632-2639