SEMESTA TEKNIKA
Vol. 22, No.1, 41-50, Mei 2019
DOI: 10.18196/st.221235

Penerapan Algoritma Cosine Similarity pada Text Mining Terjemah Al-Qur'an Berdasarkan Keterkaitan Topik

(Implementation of the Cosine Similarity Algorithm on Text Mining of Al-Qur'an Translations Based on the Relationship of Topics)

M. DIDIK R. WAHYUDI

ABSTRAK

Al-Qur'an merupakan sumber hukum dan panduan dalam pemecahan berbagai masalah umat Islam dalam menjalani kehidupan beragama, bermasyarakat, dan bernegara. Pemecaham masalah di dalam Al-Qur'an tidak hanya mengacu pada satu atau dua ayat. Jumlah ayat dan surat Al-Qur'an yang sangat banyak menyebabkan pencarian suatu ayat Al-Qur'an menggunakan cara konvensional akan memerlukan waktu lama. Oleh karena itu, dibutuhkan sebuah sistem untuk mengenali, mencari topik, dan mengelompokkan suatu permasalahan. Pencarian topik dalam terjemahan Al-Qur'an merupakan salah satu penerapan dari metode klasifikasi pengelompokan teks yang melakukan proses secara otomatis menempatkan dokumen teks ke dalam suatu kategori berdasarkan isi teks tersebut. Pengelompokan terjemah ayat Al-Qur'an berbahasa Indonesia dapat dilakukan berdasarkan tingkat kemiripan antar ayat. Algoritma yang bisa dipergunakan dalam permasalahan ini adalah Cosine Similarity. Algoritma ini akan menghitung tingkat kemiripan antar ayat yang akan menghasilkan beberapa kelompok ayat yang diambil untuk dibandingkan dengan index Al-Qur'an. Hasil penelitian menunjukkan bahwa tingkat kemiripan antar ayat sebesar 20% memberikan hasil terbaik pada pengelompokan index Al-Qur'an rata-rata sebesar 46,42%. Tingkat kemiripan antar terjemah ayat Al-Qur'an sebesar 40% memberikan rata-rata sebesar 15, 39% pada pengelompokan index Al-Qur'an. Untuk tingkat kemiripan antar ayat diatas 40%, ada kelompok similaritas ayat yang tidak masuk dalam index Al-Qur'an.

Kata kunci: text mining, cosine similarity, indeks Al-Qur'an, pengelompokan

ABSTRACT

Al-Qur'an as a Muslim holy book, contains life guidance and instructions on how to solve various problems faced by humans on earth. The Qur'an also discusses how life after death fetches every human being. The Qur'an has specific methods in grouping specific themes or problems. There are groupings based on the themes of the stories of previous people, groupings based on verses, juz, and groupings based on the place where the verses or letters of the Qur'an are revealed. In-text mining, grouping a text object can be done in various ways. One of them is based on the level of similarity. This text grouping method, of course, can be implemented in the Qur'an to find out specific patterns. The grouping of Indonesian verses in the Qur'an can be based on the level of similarity between verses using the Cosine Similarity algorithm. This algorithm will calculate the level of similarity between verses. This process will produce several groups of verses that will be taken to compare with the index of the Qur'an. The results showed that the similarity between verses was 20%, giving the best results with an average of 46.42%. The highest level of similarity where each group can still be included in the Al-Qur'an index is 40% with an average of 15, 39%. For the level of similarity between verses above 40%, there is a similarity group of verses that are not included in the Qur'anic index.

***Keywords**: text mining, cosine similarity, index Al-Qur'an, clustering*

PENDAHULUAN

Al-Qur'an merupakan suatu sumber hukum yang menjadi panduan umat Islam dalam menjalani kehidupan beragama, bermasyarakat, dan bernegara. Di dalam kehidupan ini banyak permasalahan tentang agama yang sulit untuk terpecahkan. Untuk menyelesaikan permasalahan tersebut diperlukan pemahaman dalam mengkaji Al-Qur'an. Bila penyelesaian dari masalah tersebut tanpa didasari oleh pemahaman yang mendalam mengenai Al-Qur'an, maka dapat menimbulkan perpecahan ataupun perselisihan. Dalam Al-Qur'an pemecahan dari suatu permasalahan tidak hanya mengacu pada satu atau dua ayat. Sedangkan jumlah ayat dan surat dalam Al-Qur'an sangatlah banyak, sehingga pencarian suatu ayat Al-Qur'an menggunakan cara konvensional untuk keperluan tertentu dirasakan sebagian orang akan memerlukan waktu yang agak lama. Mungkin sebagian orang ingin mencari suatu ayat dengan hanya menggunakan suatu kata kunci tertentu dikarenakan tidak mengetahui bunyi atau arti dari sebuah ayat sepenuhnya.

Proses pencarian ayat - ayat Al-Qur'an dengan cara konvensional ataupun dengan Al-Qur'an digital yang selama ini ada di internet tidak cukup membantu, jika hasil yang kita inginkan adalah ayat - ayat tertentu yang sesuai dengan masalah yang kita hadapi. Dengan begitu dibutuhkan sebuah sistem untuk mengenali, mencari dan mengelompokkan masalah yang diinputkan oleh user. Sehingga sistem tersebut dapat menampilkan ayat-ayat Al-Qur'an sebagai referensi dan solusi.

Penelitian dengan topik terjemah Al-Qur'an di antaranya adalah :

1. Penggalian data untuk mengetahui adanya keterkaitan antar topik dalam terjemahan ayat-ayat Al-Qur'an menggunakan metode Weighted K-Nearest Neighbour (WKNN) dengan Multiple Direct Hashing and Prunning (M-DHP) (Muflikhah & L, 2013).
2. Sistem Qur'an Retrieval Terjemahan Bahasa Indonesia Berbasis Web dengan Reorganisasi Korpus yang dibangun dengan menggunakan model ruang vektor (Surya Agustian, 2014).
3. Implementasi Cosine Similarity Dalam Aplikasi Pencarian Ayat Al-Qur'an Berbasis Android. Penelitian ini membahas tentang pencarian ayat berdasarkan tema tertentu dengan meode cosine similarity dengan membuat aplikasi berbasis android (Chaerul Hadi, 2017)
4. Penerapan Algoritma K-Means Clustering Untuk Pengelompokan ayat Al-Qur'an Pada Terjemahan Bahasa Indonesia dengan mempergunakan kata kunci tertentu (Robani & Widodo, 2017)
5. Klasifikasi Text Mining untuk terjemah ayat Al-Qur'an menggunakan metode Klasifikasi Naive Bayes. Dalam penelitian ini, ayat Al-Qur'an di klasifikasi kedalam 15 kelompok, dan masih ada ayat Al-Qur'an yang tidak masuk kedalam 15 kelompok tersebut (Hilwah, Kudus, & Sunendiari, 2017).

Text Mining

Text mining merupakan suatu teknologi untuk menemukan suatu pengetahuan yang berguna dalam suatu koleksi dokumen teks sehingga diperoleh tren, pola, atau kemiripan teks bahasa alamiah yang berguna untuk tujuan tertentu. Text mining adalah proses penggalian, valid, dan dapat ditindaklanjuti pengetahuan yang tersebar di seluruh dokumen dan memanfaatkan pengetahuan ini untuk lebih mengorganisir informasi untuk referensi di masa mendatang. Text mining, biasa dikenal dengan Text Data Mining (TDM), adalah penemuan oleh komputer era baru, informasi yang sebelumnya tidak diketahui, secara otomatis dengan mengekstrasi informasi dari sumber daya yang datanya tidak terstruktur (Ojo & Adeyemo, 2012).

Text mining memiliki tugas yang lebih kompleks karena melibatkan data teks yang sifatnya tidak terstruktur dan kabur (fuzzy). Text mining merupakan bidang multidisiplin yang melibatkan analisis teks, ekstraksi informasi, clustering, kategorisasi, visualisasi, teknologi basis data, machine learning, dan data mining (Konchady, 2006). Perbedaan mendasar antara text mining dan data mining terletak pada sumber data yang digunakan. Pada data mining, pola-pola diekstrak dari basis data yang terstruktur, sedangkan di text mining, pola-pola diekstrak dari data tekstual (natural language) (Marti Hearst, 2003). Secara umum, basis data didesain untuk program dengan tujuan melakukan pemrosesan secara otomatis, sedangkan teks ditulis untuk dibaca langsung oleh manusia.

Pemrosesan text mining dimulai dari text preprocessing yaitu melakukan analisis text,

setelah itu tahap features generation adalah mengumpulkan kata-kata yang sudah dibersihkan pada saat text preprocessing (Even-Zohar, 2002) . Proses selanjutnya adalah feature selection yang merupakan proses penghitungan sesuai yang dibutuhkan. Setelah itu dilakukan text/data mining yang dalam penelitian ini proses yang dilakukan adalah pencarian similaritas antar ayar terjemah Al-Qur'an. Tahap terakhir adalah visualisasi dan intepretasi hasil text/data mining.

Dalam melakukan text mining, teks dokumen yang digunakan harus dipersiapkan terlebih dahulu, setelah itu baru dapat digunakan untuk proses utama. Proses mempersiapkan teks dokumen atau dataset mentah disebut juga dengan proses text preprocessing. Text preprocessing berfungsi untuk mengubah data teks yang tidak terstruktur menjadi data yang terstruktur dengan cara memilih setiap kata dari dokumen dan merubahnya menjadi kata dasar yang memiliki arti sempit dan proses teks mining akan memberikan hasil yang lebih memuaskan (Septiawan, Suprayogi, Mukhtar, & Hatiyanto, 2010).

Cosine Similarity

Cosine similarity adalah metode similaritas yang digunakan untuk menghitung similartias dua buah dokumen. Metode yang dipergunakan adalah melakukan perhitungan ukuran kesamaan antara dua buah vektor dalam sebuah ruang dimensi yang didapat dari nilai cosinus sudut dari perkalian dua buah vektor yang dibandingkan karena cosinus dari 0 adalah 1 dan kurang dari 1 untuk nilai sudut yang lain (Cios, Swiniarski, Pedrycz, & Kurgan, 2007). Nilai similarity dari dua buah vektor dikatakan mirip ketika nilai cosine similarity adalah 1.

Cosine similarity digunakan dalam ruang positif, dimana hasilnya dibatasi antara nilai 0 dan 1. Kalau nilainya 0 maka dokumen tersebut dikatakan mirip jika hasilnya 1 maka nilai tersebut dikatakan tidak mirip. Batas ini berlaku untuk sejumlah dimensi dan Cosine similarity ini paling sering digunakan dalam ruang positif dimensi tinggi. Misalnya, dalam Information Retrieval, masing-masing kata/istilah (term) diasumsikan sebagai dimensi yang berbeda dan dokumen ditandai dengan vector dimana nilai masing-masing dimensi sesuai dengan berapa istilah muncul dalam dokumen.

Index Al-Qur'an

Dalam upaya untuk membumikan ajaran Islam di tengah-tegah masyarakat Indonesia ada dasarnya telah dimulai sejak pertama kali Islam masuk ke Nusantara yaitu pada awal abad III. Hal ini terbukti diantaranya dengan adanya proses islamisasi masyarakat Indonesia yang dilakukan oleh para juru dakwah Islam dengan menggunakan sistem akulturasi dengan budaya setempat yang pada waktu itu masih kental dengan nuansa Hindu Budha. Ajaran Islam yang bersumber dari dua pilar utama yakni Al-Qur'an dan sunnah Rasulullah saw telah mendapat perhatian besar dari para sarjana muslim Indonesia. Hal ini terlihat dengan banyaknya kajian-kajian yang dilakukan terhadap Al-Qur'an khususnya terutama pada aspek penafsiran Al-Qur'an (Atabik, 2014).

Pada tabel 1 ditunjukkan beberapa pengkajian tentang tafsir Al-Qur'an yang telah dirintis oleh beberapa tokoh (Azra, 2013).

TABEL 1 : Aktivitas pengkajian tentang penafsiran Al-Quran

No.	Tahun	Buku	Penulis
1	1928	Al-Furqan	A.Hassan Bandung
2	1935	Tafsir Quran Indonesia	Mahmud Yunus
3	1955	Tafsir Al-Quran Al-Karim	Halim Hassan
4	1959	Tafsir Al-Quran	Zainuddin Hamidi
5	1960	Tafsir Al-Quran al-Hakim	Iskandar Idris dan Kasim Bakry
6	2001	Tafsir Al-Mishbah	Quraish Shihab
7	1973	Tafsir Al-Azhar	Hamka

Tafsir Al-Qur'an juga dilakukan oleh suatu lembaga, organisasi atau penerbitan. Misalnya oleh Muhammadiyah, Persis Bandung, beberapa penerbitan terjemah di Medan, Minangkabau, dan kawasan lainnya. Upaya ini bahkan ditindak lanjuti secara resmi oleh pemerintah Indonesia melalui proyek Penerjemahan dan penafsiran Al-Qur'an di bawah naungan Departemen Agama RI.

Tafsir-tafsir yang ditulis dalam bahasa Indonesia pada hakikatnya sudah memenuhi kebutuhan umat Islam yang meyakini betapa pentingnya memahami dan mengamalkan ajaran yang terkandung dalam Al-Qur'an. Kebutuhan masyarakat akan indeks Al-Qur'an bisa dilihat dari maraknya penggunanaan indeks-indeks Al-Qur'an oleh sebagaian masyarakat Indonesia. Usaha untuk membuat indeks Al-Qur'an sebagai salah satu alat untuk mempermudah memahami dan membumikan ajaran Al-Qur'an telah dilakukan oleh para mufassir Indonesia pada generasi kedua dalam kajian tafsir di Indonesia.

METODOLOGI PENELITIAN

Metode penelitian yang dipakai dalam penelitian ini adalah metode eksperimen atau penelitian terapan, yaitu dengan menerapkan metode *Cosine Similarity* untuk mengklasfikasi atau mengelompokkan terjemah Al-Qur'an kedalam beberapa kelompok sesuai dengan tingkat kemiripan antar ayat. Terjemah Al-Qur'an yang dipergunakan adalah terjemah Al-Qur'an yang dikeluarkan oleh Kementerian Agama Republik Indonesia. Tahapan penelitian terdapat pada Gambar 1.

1. Konversi Data

Terjemah Al-Qur'an dan Index Al-Qur'an yang menjadi kedua sumber data tersebut akan dikonversi kedalam dataset untuk pemrosesan lebih lanjut. Konversi data dilakukan pada terjemah Al-Qur'an dan index Al Quran.

2. Preprocessing

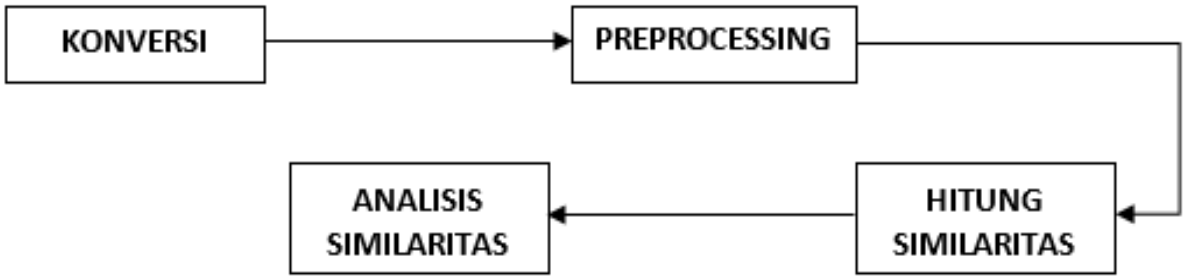
Preprocessing merupakan tahap untuk mempersiapkan teks menjadi data yang akan diolah. Input berupa dokumen atau string. Secara umum, proses ini memiliki beberapa tahapan yaitu lemmatizing, case folding, tokenizing, stop word removal, stemming, dan lain-lain. Preprocessing pada penelitian ini meliputi lemmatizing dan stop word tidak dihapus. Proses lemmatizing adalah suatu proses untuk mengembalikan suatu kata ke kata dasar. Pada penelitian ini, stop word dibiarkan saja dan tidak dihapus. Hal ini dilakukan agar makna sebenarnya dari terjemah Al-Qur'an bermakna tetap. Penghapusan stop word akan berakibat pada berubahnya makna yang sebenarnya dai terjemah Al-Qur'an menjadi makna yang sebaliknya. Hal ini bertentangan justru menghilangkan nilai kebenaran dari suatu kalimat.

3. Hitung Similaritas

Setelah selesai melakukan preprocessing terhadap ayat Al-Qur'an maka proses berikutnya adalah mencari tingkat similaritas antar ayat terjemah Al-Qur'an. Setiap ayat akan dibandingkan dan dicari seberapa besar tingkat kemiripan antara ayat satu dengan yang lainnya. Proses ini mempergunakan algoritma Cosine Similarity.

4. Analisis Similaritas

Setelah nilai similaritas antar ayat terjemah Al-Qur'an diperoleh, maka akan dilakukan pengujian hasil similaritas ini terhadap index Al-Qur'an tematik yang sudah disiapkan. Proses ini dilakukan untuk melihat sejauh mana kinerja Algoritma Cosine Similarity terhadap index Al-Qur'an tematik.



GAMBAR 1. Alur metode penelitian

HASIL DAN PEMBAHASAN

Penelitian ini mempergunakan data digital terjemah Al-Qur'an dalam bahasa Indonesia yang diterbitkan oleh Kementerian Agama Republik Indonesia. Setiap ayat terjemah Al-Qur'an berbahasa Indonesia ini akan diproses dengan algoritma Cosine Similarity untuk mencari nilai similaritas antar ayat. Nilai similaritas antar ayat yang dihasilkan kemudian akan di uji dengan Index Al-Qur'an.

Konversi data dilakukan pada terjemah Al-Qur'an dan index Al Quran. Terjamah Al-Qur'an berbahasa Indonesia yang dipergunakan berformat file Excell dengan bentuk tabel Excell yang belum terstruktur untuk pengolahan data dengan bahasa Python.

Dari format Excell, proses konversi dilakukan kedalam bentuk dataset dengan format CSV (Comma Separated Values) agar bisa di proses lebih lanjut oleh bahasa pemrograman python. Proses konversi dilakukan dengan pengeditan file Excell tersebut secara manual dengan menghilangkan kolom dan baris yang tidak dibutuhkan. Proses selanjutnya adalah dengan menyimpan hasil editan tersebut kedalam format CSV.

Proses konversi index Al-Qur'an ini relatif lebih rumit jika dibandingkan dengan konversi terjemah Al-Qur'an diatas. Proses konversi yang dilakukan meliputi :

1. Ambil per tema index.
2. Untuk setiap index yang dipilih, identifikasi ayat-ayat yang termasuk dalam anggota index tersebut
3. Hapus kolom dan baris yang tidak dibutuhkan.
4. Simpan dalam bentuk file excell setiap tema yang sudah diidentifikasi ayat-ayat yang masuk kategorinya.
5. Setelah semua tema selesai dikelompokkan, maka akan terbentuk file excell yang berisi sekumpulan terjemah ayat-ayat Al-Qur'an untuk setiap tema.
6. Buat tabel daftar tema dan notasikan dalam bentuk angka terlihat pada Gambar 2. Konversi setiap tema yang sudah berisi ayat tersebut ke dalam file csv untuk membentuk dataset yang dibutuhkan. Sehingga nanti akan tampil dataset.
7. Hilangkan duplikasi ayat yang ada dalam satu tema, kemudian gabungkan seluruh dataset setiap tema kedalam 1 file dengan format CSV. Proses penghilangan duplikasi dan penggabungan ini dengan mempergunakan MS-ACCESS.
8. Setelah digabungkan, maka diperoleh dataset index Al-Qur'an dengan jumlah ayat setiap tema.

Hasil preprocessing disimpan kedalam kolom tjclean dalam file CSV. Isi kolom tjclean ini yang nanti akan dipergunakan untuk pengolahan text mining terjemah Al-Qur'an. Hasil preprocessing dapat dilihat pada Gambar 3.

NOIDX	NAMA INDEX
1	IMAN
2	ILMU
3	BANGSA-BANGSA TERDAHULU
4	SEJARAH
5	AL QUR'AN
6	AKHLAQ & ADAB
7	IBADAH
8	MAKANAN & MINUMAN
9	PAKAIAN & PERHIASAN
10	HUKUM PRIVAT
11	MUAMALAT
12	PERADILAN & HAKIM
13	HUKUM PIDANA, JINAYAH
14	JIHAD

GAMBAR 2. Daftar Tema

```
id,terjemah,tjclean
[1.1],Dengan menyebut nama Allah Yang Maha Pemurah lagi Maha
Penyayang,menyebut nama allah maha pemurah maha penyayang
[1.2],Segala puji bagi Allah Tuhan semesta alam,puji allah tuhan semesta alam
[1.3],Maha Pemurah lagi Maha Penyayang,maha pemurah maha penyayang
[1.4],Yang menguasai hari pembalasan.,menguasai pembalasan
[1.5],Hanya kepada Engkaulah kami menyembah dan hanya kepada Engkaulah kami
mohon pertolongan,engkaulah menyembah engkaulah mohon pertolongan
[1.6],Tunjukilah kami jalan yang lurus,tunjukilah jalan lurus
[1.7],(yaitu) jalan orang-orang yang telah Engkau anugerahkan nikmat kepada
mereka bukan (jalan) mereka yang dimurkai dan bukan (pula jalan) mereka yang
sesat.,jalan orangorang engkau anugerahkan nikmat jalan dimurkai jalan sesat
[2.1],Alif Laam Miim.,alif laam miim
[2.2],Kitab (Al Qur'an) ini tidak ada keraguan padanya petunjuk bagi mereka
yang bertakwa,kitab al quran keraguan petunjuk bertakwa
```

GAMBAR 3. Hasil Preprocessing

Cosine Similarity berfungsi membandingkan kemiripan antar dokumen atau teks yang akan dianggap sebagai vektor. Sehingga nilai Cosine Similarity dari Vektor A dan B dapat dihitung seperti persamaan berikut :

$$Cosine\ Similarity = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

Untuk pencocokan text, nilai dari vektor A dan B adalah vektor term-frequency dari dokumen. Nilai Cosine Similarity berada pada range 0-1. Semakin mirip teks A dan B, maka nilai Cosine Similarity makin mendekati angka 1. Sebaliknya, makin tidak mirip antara teks A dan B, maka nilai Cosine Similarity makin kecil atau mendekati angka 0.

Proses perhitungan nilai Cosine Similarity pada penelitian ini, dilakukan dengan cara membandingkan antar ayat setiap terjemah Al Quran berbahasa Indonesia dengan memanggil fungsi Cosine Similarity. Keluaran dari proses ini adalah nilai kemiripan antar ayat terjemah Al-Qur'an.

Hasil perhitungan nilai similaritas antar ayat disimpan dalam file scoresim-ALL-NS.CSV. File ini berisi tiga atribut yaitu ayat kedua ayat yang dibandingkan serta nilai similarity diantara keduanya. Dari 6.234 jumlah ayat terjemah Al-Qur'an, diperoleh sebanyak 1.647.138 data perbandingan kemiripan antar ayat. Tampilan isi file scoresim-ALL-NS.CSV yang dihasilkan ditunjukkan oleh Gambar 4.

Setelah nilai similaritas antar ayat terjemah Al-Qur'an diperoleh, maka akan dilakukan pengujian hasil similaritas ini terhadap index Al-Qur'an tematik yang sudah disiapkan. Proses ini dilakukan untuk melihat sejauh mana kinerja Algoritma Cosine Similarity terhadap index Al-Qur'an tematik. Langkah analisis yang dilakukan adalah :

1. Dari data similaritas antar ayat yang telah dihitung, maka akan diperoleh kelompok-kelompok ayat berdasarkan pada tingkat kemiripan. Urutkan dari jumlah kemiripan paling banyak menuju paling sedikit. Ambil 14 besar kelompok ayat yang memiliki member paling banyak. Kelompok 14 besar ini yang akan kita jadikan acuan kelompok ayat yang akan kita proses dengan kelompok index Al-Qur'an. Mulai proses ini dilakukan dengan bantuan MS-ACCESS. Gambar 5 menunjukkan 14 besar kelompok ayat yang dimaksud.

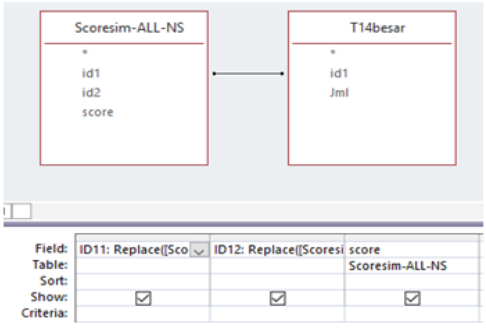
```
1         id1,id2,score
2         [1.1], [1.3], 0.816496580927726
3         [1.1], [2.28], 0.20100756305184242
4         [1.1], [2.29], 0.3162277660168379
5         [1.1], [2.32], 0.45883146774112343
6         [1.1], [2.37], 0.5217491947499509
7         [1.1], [2.54], 0.30151134457776363
8         [1.1], [2.80], 0.2357022603955158
9         [1.1], [2.95], 0.3162277660168379
10        [1.1], [2.96], 0.20851441405707477
...
...
1647130   [114.1], [114.3], 0.3162277660168379
1647131   [114.1], [114.5], 0.22360679774997896
1647132   [114.1], [114.6], 0.2581988897471611
1647133   [114.2], [114.3], 0.4999999999999999
1647134   [114.2], [114.5], 0.35355339059327373
1647135   [114.2], [114.6], 0.408248290463863
1647136   [114.3], [114.5], 0.35355339059327373
1647137   [114.3], [114.6], 0.408248290463863
1647138   [114.4], [114.5], 0.25
1647139   [114.5], [114.6], 0.28867513459481287
```

GAMBAR 4. Nilai similaritas antar ayat

id1	JML
[2.218]	2163
[2.165]	2092
[2.153]	2018
[2.8]	1989
[4.140]	1942
[3.32]	1941
[5.72]	1914
[2.116]	1907
[3.19]	1870
[2.62]	1850
[3.63]	1831
[3.57]	1818
[3.20]	1806
[3.155]	1788
[3.152]	1760
[2.190]	1757
[3.173]	1749
[3.176]	1743

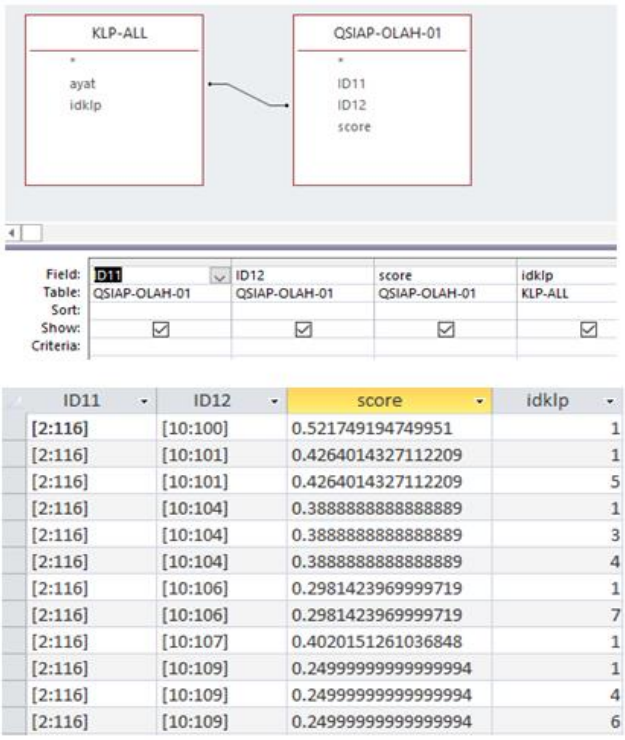
GAMBAR 5. Daftar 14 besar kelompok ayat

2. Buat relasi antara 14 kelompok ayat yang diproses pada langkah 1 diatas dengan hasil penilaian similaritas antar ayat. Query proses tersebut dan hasil eksekusinya query ditunjukkan oleh Gambar 6.
3. Tentukan kelompok index yang menjadi member kelompok similaritas antar ayat yang dihasilkan. Proses ini akan menghasilkan similaritas antar ayat dan masuk kategori apa ayat tersebut dalam index tematik terjemah Al-Qur'an. Gambar 7 menunjukkan proses pencocokan yang dilakukan.
4. Hitung jumlah ayat berdasarkan index Al-Quran. Ada kemungkinan satu ayat akan menjadi lebih dari satu bagian kelompk index. Filter kelompok index yang paling banyak memiliki member. Hal ini menunjukkan kelompok index tersebut yang paling dominan.



ID11	ID12	score
[2:218]	[2:220]	0.47087095579
[2:218]	[2:221]	0.27174648819
[2:218]	[2:222]	0.34995661963
[2:218]	[2:223]	0.25087260300
[2:218]	[2:224]	0.44992127066
[2:218]	[2:225]	0.47795211939
[2:218]	[2:226]	0.65271395186
[2:218]	[2:227]	0.47304991679
[2:218]	[2:228]	0.42365927286
[2:218]	[2:229]	0.35154137266
[2:218]	[2:231]	0.42295493443
[2:218]	[2:232]	0.40492914359

GAMBAR 6. Relasi kelompok ayat dan similaritas antar ayat



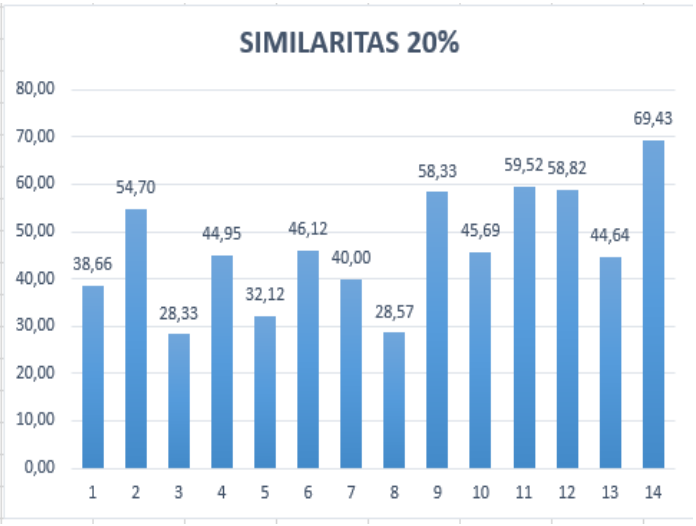
ID11	ID12	score	idklp
[2:116]	[10:100]	0.521749194749951	1
[2:116]	[10:101]	0.4264014327112209	1
[2:116]	[10:101]	0.4264014327112209	5
[2:116]	[10:104]	0.3888888888888889	1
[2:116]	[10:104]	0.3888888888888889	3
[2:116]	[10:104]	0.3888888888888889	4
[2:116]	[10:106]	0.2981423969999719	1
[2:116]	[10:106]	0.2981423969999719	7
[2:116]	[10:107]	0.4020151261036848	1
[2:116]	[10:109]	0.24999999999999994	1
[2:116]	[10:109]	0.24999999999999994	4
[2:116]	[10:109]	0.24999999999999994	6

GAMBAR 7. Proses pencocokan similaritas dengan index temati

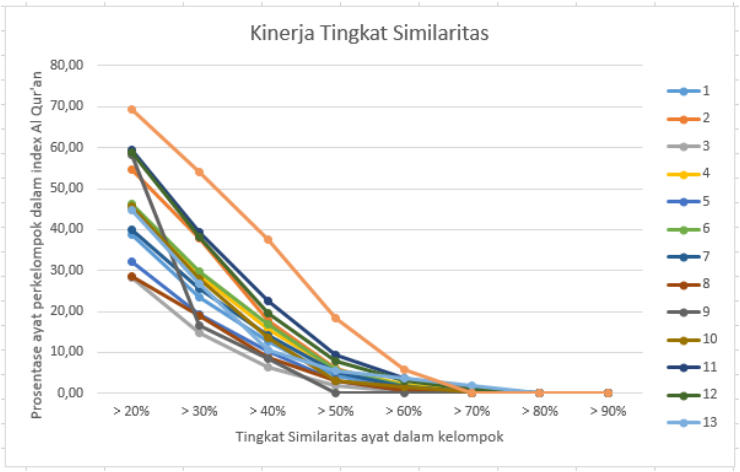
5. Hitung prosentase kecocokan Index Al-Qur'an terhadap kelompok similaritas antar ayat secara silang antara nilai similaritas terhadap index tematik Al-Qur'an dilakukan secara bertahap dari similaritas 20% hingga 90%. Hasil pencocokkan ditunjukkan oleh Gambar 8.

Hasil perhitungan similaritas terjemah ayat Al-Qur'an berbahasa Indonesia terhadap index tematik diatas menunjukkan bahwa dengan tingkat kemiripan antar ayat sebesar 20%, algoritma Cosine Similarity memberikan kesesuaian paling optimal yaitu sebesar rata-rata 46,42% terhadap index tematik Al-Qur'an. Sedangkan tingkat kemiripan sebesar 90%, memberikan kesesuaian paling rendah yaitu sebesar rata-rata 0,01% ditunjukan pada Gambar 9.

IdKlp	JML	JML-IDX	PROSEN
1	4969	1921	38,66
2	298	163	54,70
3	2090	592	28,33
4	832	374	44,95
5	2030	652	32,12
6	1483	684	46,12
7	775	310	40,00
8	126	36	28,57
9	12	7	58,33
10	197	90	45,69
11	84	50	59,52
12	102	60	58,82
13	56	25	44,64
14	157	109	69,43
RATA-RATA			46,42



GAMBAR 8. Similaritas 20% terhadap index Al-Qur'an



GAMBAR 9. Grafik rekap pengujian similaritas berdasarkan prosentase kemiripan

KESIMPULAN

Berdasarkan penelitian yang sudah dilakukan, didapatkan kesimpulan bahwa *text mining* terjemah ayat Al-Qur'an berdasarkan keterkaitan topik dengan mempergunakan algoritma *Cosine Similarity* dalam pengelompokan ayat berdasarkan tingkat kemiripan atau similaritas berhasil dilakukan. Hasil pengelompokan ayat berdasarkan tingkat kemiripan ini kemudian dicocokkan dengan index Al-Qur'an untuk mencari kelompok ayat mana yang paling sesuai dengan tema dalam index Al-Qur'an. Tingkat kemiripan antar ayat sebesar 20%, memberikan hasil terbaik dalam pengujian pencocokan kelompok ayat dengan rata-rata masuk dalam index Al-Qur'an sebesar 46,42%.

Tingkat kemiripan tertinggi dimana setiap kelompok masih dapat masuk dalam index Al-Qur'an adalah sebesar 40% dengan rata-rata masuk dalam index Al-Qur'an sebesar 15, 39%. Untuk tingkat kemiripan antar ayat diatas 40%, sudah mulai ada kelompok similaritas ayat yang tidak masuk dalam index Al-Qur'an. Hal ini menunjukkan makin penurunan prosentase kelompok ayat dalam index Al-Qur'an. Berikut hasil kinerja algoritma *Cosine Similarity* pada *text mining* terjemah ayat Al-Qur'an berdasarkan keterkaitan topik dengan kemiripan 20% hingga 90%.

UCAPAN TERIMA KASIH

Terimakasih kepada Lembaga Penelitian dan Pengabdian Masyarakat UIN Sunan Kalijaga yang telah mendanai penelitian ini.

DAFTAR PUSTAKA

Atabik, A. (2014). Perkembangan Tafsir Modern di Indonesia. *Perkembangan Tafsir Modern Di Indonesia*, 318–322. Retrieved from http://journal.stainkudus.ac.id/index.php/Hermeneutik/article/viewFile/895/831

Azra, A. (2013). *Edisi Perenial Jaringan Ulama Timur Tengah dan Kepulauan Nusantara Abad XVII & XVIII Akar Pembaharuan Islam Nusantara*.

Chaerul Hadi, M. R. M. (2017). *Implementasi Cosine Similarity Dalam Aplikasi Pencarian Ayat Al-Qur'an Berbasis Android*. *6*(2), 71–79.

Cios, K. J., Swiniarski, R. W., Pedrycz, W., & Kurgan, L. A. (2007). *Data mining. [electronic resource] : a knowledge discovery approach Data Mining*. Retrieved from http://sfx.fcla.edu/usf?genre=bookitem

Even-Zohar, Y. (2002). Introduction to text mining. *Automated Learning Group National Center for Supercomputing Applications University of Illinois*.

Hilwah, N., Kudus, A., & Sunendiari, S. (2017). Klasifikasi Text Mining untuk Terjemahan Ayat-Ayat Al- Qur'an menggunakan Metode Klasifikasi Naive Bayes. *Prosiding Statistika*, *2*(2), 179–185.

Konchady, M. (2006). *Text Mining Application Programming (Charles River Media Programming)*. Retrieved from http://www.amazon.com/Mining-Application-Programming-Charles-River/dp/1584504609

Marti Hearst. (2003). What Is Text Mining? Retrieved from http://people.ischool.berkeley.edu/~hearst/text-mining.html

Muflikhah, L., & L, D. Y. (2013). *Penggalian Data dalam Penentuan Keterkaitan Topik pada Terjemahan Ayat- ayat Al-Qur ' an*. *1*(1).

Ojo, A., & Adeyemo, A. (2012). Framework for Knowledge Discovery from Journal Articles Using Text Mining Techniques. *African Journal of Computing & ICT*. Retrieved from http://www.ajocict.net/uploads/Pre-print_-_O__Ojo___A_B__Adeyemo__2012___Framework_for_Knowledge_Discovery_from_Journal_Articles_Using_Text_Mining_Techniques.pdf

Robani, M., & Widodo, A. (2017). Algoritma K-Means Clustering Untuk Pengelompokan Ayat Al Quran Pada Terjemahan Bahasa Indonesia. *Jurnal Sistem Informasi Bisnis*, *6*(2), 164. https://doi.org/10.21456/vol6iss2pp164-176

Septiawan, D., Suprayogi, D. A., Mukhtar, A. M., & Hatiyanto, W. (2010). *Klasifikasi Iklan pada Online Shop dengan Metode Naive Bayes*. (8).

Surya Agustian, I. S. W. (2014). *Sistem Qur'an Retrieval Terjemahan Bahasa Indonesia Berbasis Web dengan Sistem Qur'an Retrieval Terjemah Bahasa Indonesia*. (November). https://doi.org/10.13140/2.1.5168.6083

PENULIS:

M. Didik R. Wahyudi

Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Kalijaga, Yogyakarta.

Email: m.didik@uin-suka.ac.id