ANALISIS DAN IMPLEMENTASI PENCARIAN KATA BERBASIS KONKORDANSI DAN N-GRAM PADA TERJEMAHAN AL-QURAN BERBAHASA INDONESIA

ANALYSIS AND IMPLEMENTATION CONCORDANCE SEARCH AND N-GRAM FOR WORDS IN AL-QURAN ENGLISH TRANSLATION

Devye Bellika Nugraheni[1], Moch. Arif Bijaksana[2], Eko Darmawiyanto[3]
[1]Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom
[2]Fakultas Informatika, Universitas Telkom
[3]Fakultas Informatika, Universitas Telkom
[1]devyebellika@gmail.com,[2]arifbijaksana@telkomuniversity.ac.id,[3]ekodarmawiyanto@telkomuniversity.ac.id

Abstrak

Pengkajian ayat-ayat Al-Quran sangatlah penting bagi umat muslim khususnya, namun sayang dalam pelaksanaannya masih menggunakan cara manual termasuk dalam pencarian ayat-ayat Al-Quran yang mengandung kata tertentu untuk dipergunakan sebagai contoh. Dengan menggunakan metode pencarian berbasis konkordansi, akan mempermudah dalam membantu pencarian seluruh ayat yang mengandung kata tertentu dengan kesamaan lema. Tidak hanya itu, dengan konkordansi dapat dilakukan pengolahan kata sehingga dapat menghasilkan informasi tambahan yang lebih lengkap. Salah satu pengolahan yang dapat digunakan untuk membantu adalah teori marcov assumption terkait n-gram. N-Gram akan membantu dalam mencari informasi terkait kombinasi kata tertentu dengan kata-kata disekitarnya.

Kata Kunci : Al-Quran, konkordansi, n-gram

Abstract

Quran Research is very important to the Moeslims, but unfortunately the implementation of this research is still done manually, for example, searching the verses of the Qur'an that contain certain words that will be used as an example. By doing concordance-based search, this method will help us to search the entire paragraph containing certain words with similar entries. Not only that, concordance can do a word-processing that will produce a more complete additional information. One function that can be used to help us is marcov theoretical assumption about n-grams. N-Gram will assist in searching the related information about combination of words and also the words around it

Keyword : Al-Quran, concordance, n-gram.

1. Pendahuluan

Al-Quran memiliki kandungan yang dibutuhkan kaum muslim dalam menjalani kehidupan. Oleh karena itu pengkajian ayat-ayat Al-Quran sangatlah penting bagi kaum muslim. Namun sayangnya kegiatan ini masih dilakukan secara manual, termasuk dalam pencarian ayat yang mengandung suatu kata tertentu untuk dijadikan contoh dan referensi tafsir Al-Quran. Pada saat ini sebenarnya sudah sangat memungkinkan pencarian suatu kata dalam kumpulan kalimat di dokumen tertentu dilakukan dengan bantuan aplikasi berbasis konkordansi.

Pencarian berbasis konkordansi sangat berbeda dengan pencarian menggunakan query search seperti kebanyakan. Pada query, pencarian harus terpaku pada aturan yang telah ditetapkan sebelumnya, sedangkan konkordansi mempunyai sifat yang lebih dinamis dan fleksibel. Konkordansi sendiri merupakan salah satu metode yang digunakan dalam pemadanan unsur leksikal yang konsisten [1]. Unsur leksikal yang biasa digunakan adalah kata, lema, ataupun akar kata. Lema sendiri merupakan pola suatu kata atau frasa di dalam kamus [2]. Tidak hanya memungkinkan dalam menampilkan semua kalimat dalam dokumen tertentu yang mengandung kata-kata yang memiliki persamaan unsur leksikal, dengan konkordansi dapat juga diperoleh informasi tambahan dari suatu kata tertentu tersebut setelah dilakukan pengolahan.

Oleh karena itu pada Tugas Akhir ini menggunakan bahasa Arab-Indonesia dengan menggunakan tiga macam konkordansi, yaitu lema, kata dan sinonim. Hasil dari konkordansi juga akan dikombinasikan dengan pemrosesan tambahan seperti concordance plot, context word, dan N-Gram. N-Gram dapat diartikan berfungsi dalam pengambilan potongan n karakter dalam suatu string atau kalimat tertentu [5]. Penggunaan N-Gram dibutuhkan dalam pengolahan untuk memberi informasi tambahan yang masih terkait dengan kata kunci yang dimasukkan user

seperti halnya untuk mengetahui frekuensi kemunculan, range dan kombinasi cluster yang mungkin dari suatu kata tertentu dari ayat-ayat Al-Quran. Informasi-informasi tambahan ini nantinya akan ditampilkan dengan semua ayat Al-Quran yang mengandung unsur leksikal sama dengan kata kunci beserta jenis part of speechnya, sehingga user akan lebih mudah dalam pengkajian Al-Quran khususnya dalam menemukan informasi secara menyeluruh.

2. Perancangan Sistem

2.1 Dataset (Corpus)

Data set atau corpus yang digunakan dalam aplikasi ini dibangun secara manual dengan megambil data-data yang dibutuhkan dari berbagai dokumen yang terkait, seperti halnya ayat Al-Quran, ayat per kata, terjemahan Al-Quran, terjemahan per kata, label POS-tag, dan lema. Corpus berbentuk tabel dengan jumlah data 33.081 baris, yang mewakili dari 13 juz Al-Quran, yaitu juz 1, 2, 3, …, 11, 29, dan 30.

Tabel 2-1: Template Kolom Dataset

Index	Kata	Lema/Root	Terj Kata	Ayat	Terj Ayat	Satu Kata	POS-tag	
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...

2.2 Konkordansi

2.2.1 Berdasarkan Lema

Pada tahap ini dilakukan proses pencarian kata pada setiap ayat Al-Quran yang memiliki kesamaan lema. Adapun tahap-tahapnya adalah sebagai berikut.

i. Kata yang terpilih akan dimasukkan ke dalam variabel input
ii. Berdasar informasi indeks yang didapat dari kata didalam variabel input, dicarilah lema dari kata tersebut pada korpus
iii. Dicari semua kata yang memiliki lema yang sama dengan lema pada kata di variabel input
iv. Ditampilkan semua ayat yang mengandung kata hasil dari poin 3 dan juga terjemahan ayat tersebut

2.2.2 Berdasarkan Kata

Pada tahap ini dilakukan proses pencarian kata pada setiap ayat Al-Quran yang memiliki kesamaan kata terjemahan. Adapun tahap-tahapnya adalah sebagai berikut.

i. Kata yang terpilih akan dimasukkan ke dalam variabel input
ii. Berdasar informasi indeks yang didapat dari kata didalam variabel input, dicarilah terjemahan kata dari kata tersebut pada korpus
iii. Dicari semua kata yang memiliki kandungan terjemahan kata yang sama dengan kata input
iv. Ditampilkan semua ayat yang mengandung kata hasil dari poin 3 dan juga terjemahan ayat tersebut

2.2.3 Berdasarkan Sinonim

Pada tahap ini dilakukan proses pencarian kata pada setiap ayat Al-Quran yang memiliki kesamaan makna (sinonim). Adapun tahap-tahapnya adalah sebagai berikut.

i. Kata yang terpilih akan dimasukkan ke dalam variabel input
ii. Berdasar informasi indeks yang didapat dari kata didalam variabel input, dicarilah terjemahan kata dari kata tersebut pada korpus
iii. Dengan API tesaurus, didapat list sinonim dari kata tersebut.
iv. Dicari semua ayat yang mengandung kata yang ada di list sinonim
v. Ditampilkan semua ayat yang mengandung kata hasil dari poin 3 dan juga terjemahan ayat tersebut

2.3 Concordance Plot

Pada tahap ini dilakukan proses penggambaran grafik yang merepresentasikan letak ayat yang mengandung kata yang sama dengan kata pada variabel *input.* Adapun tahap-tahapnya adalah sebagai berikut.

i. Kata yang ada di variabel *input* digunakan kembali untuk mencari terjemahan dari kata tersebut.
ii. Dicari semua terjemahan ayat yang mengandung kata yang sama dengan kata pada variabel *input.*
iii. Jika ditemukan terjemahan ayat yang dimaksud pada poin (ii), maka akan dicek indexnya untuk penggambaran letak ayat tersebut pada grafik.
iv. Proses pada poin (iii) diulang terus hingga tidak ditemukan lagi terjemahan ayat yang mengandung kata yang mengandung kata yang sama dengan kata pada variabel *input.*
v. Ditampilkan hasil akhir grafik beserta terjemahan ayatnya.

2.4 Context Word

Pada tahap ini dilakukan proses pencarian kombinasi kata input dengan kata lain pada ayat yang ada di list hasil konkordansi. Seperti pada gambar diatas, yang menunjukkan keluaran berupa list ayat-ayat yang mengandung kombinasi kedua kata *input.* Adapun tahap-tahapnya adalah sebagai berikut.

i. Kata yang ada di variabel *input* digunakan kembali dan dilakukan tokenisasi terhadap terjemahan ayat.
ii. Di cek apakah kata inputan kedua ada di hasil tokenisasi, jika ada maka ditampilkan kedalam list *output*an, jika tidak ada maka dihapus dari list.
iii. List diperbarui dan disusun ulang berdasar urutan tertentu.
iv. Proses akan diulangi terus hingga semua list selesai diolah.

2.5 N-GRAM

Pada tahap ini dilakukan proses pencarian kombinasi kata input dengan kata lain pada ayat yang ada di list hasil konkordansi. Adapun tahap-tahapnya adalah sebagai berikut.

i. Dilakukan pengecekan pada ketentuan tambahan yang dituliskan *user* pada form lanjutan
ii. Dilakukan pembaruan list ayat
iii. Instansiasi *range* = *range* minimal
iv. Dilakukan pengecekan apakah range dari suatu kombinasi kata tertentu masih kurang dari rangeMax, selama kondisi tersebut terpenuhi, maka kata tertentu tersebut ditampilkan di list.
v. Ketika tidak terpenuhi lagi, maka ditampilkan kumulatif nilai frekuensi kemunculan, rank, *range* dan *cluster*.

2.6 Pengujian

Untuk mengukur kualitas dan efektifitas dari hasil kategorisasi, dilakukan pengukuran *precision, recall* dan *accuracy*. Metode ini dilakukan dengan membandingkan antara hasil pengelompokan *Gold Standart* dengan pengelompokan oleh sistem yang hendak diuji. *Gold Standart* merupakan hasil yang dijadikan acuan dalam pengujian, dan biasanya disusun secara manual.

Tabel 2-6 : Klasifikasi untuk Pengujian Performansi

Sistem	***Gold Standart***	
	Mengandung *Keyword*	**Tidak Mengandung *Keyword***
Mengandung *Keyword*	*True Positif*	*False Positif*
Tidak Mengandung *Keyword*	*False Negatif*	*True Negatif*

Dari perbandingan tersebut, maka pengukuran performasi yang akan dilakukan meliputi : *Precision, Recall* dan Akurasi.

a. Akurasi
   Ukuran yang digunakan untuk mengetahui seberapa besar kebenaran yang didapatkan dari akumulasi keseluruhan data yang ada.
b. *Precision*
   Ukuran yang digunakan untuk mengetahui seberapa besar dari hasil yang telah terpilih itu benar

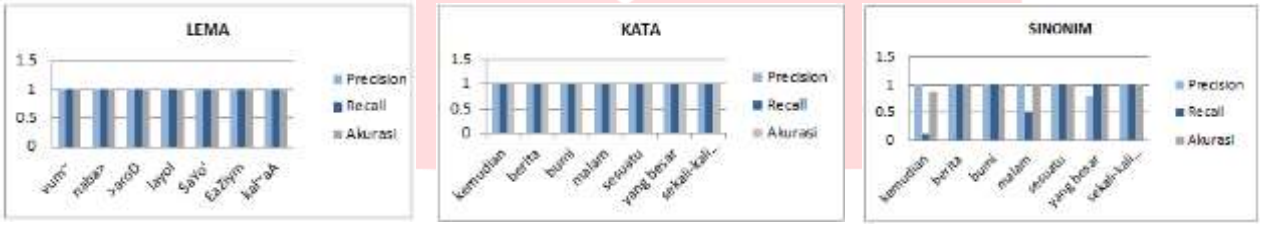
c. *Recall*

Ukuran yang digunakan untuk mengetahui seberapa besar yang benar itu terpilih.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \qquad Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

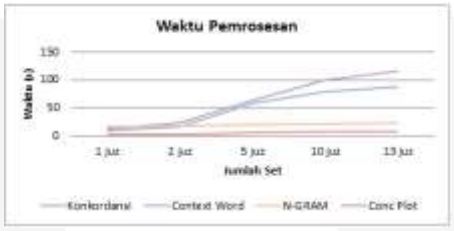
3. Pembahasan

Pada Tugas Akhir ini, pengujian dilakukan dengan cara antara lain: *blackbox testing*, performansi dan kecepatan eksekusi.



(a) (b) (c)

Gambar 3-1 : Hasil perhitungan precision, recall, dan akurasi pengujian performansi pada: a) konkordansi berdasarkan lema, b) konkordansi berdasarkan kata terjemahan, dan c) konkordansi berdasarkan sinonim.



Gambar 3-2 : Grafik hasil analisis kecepatan pemrosesan

3.1 Pengujian Performansi

Pada pengujian ini menggunakan batasan domain juz 30 dimana datanya sebanyak 2308 baris dengan 53.834 kata. *Gold standart* didapat dari perhitungan manual untuk konkordansi lema dan kata terjemahan, sedangkan konkordansi sinonim mengacu pada tesaurus Bahasa Indonesia pada web www.sinonimkata.com.

Tabel 3-1: Tabel Perhitungan Performansi Konkordansi Lema

Lema	TP	FN	FP	TN		P	R	A
yum~	17	0	0	547		1	1	1
naba>	1	0	0	563		1	1	1
>aroD	12	0	0	552		1	1	1
layol	9	0	0	555		1	1	1
$aYo'	4	0	0	560		1	1	1
EaZiym	3	0	0	561		1	1	1
kal~aA	18	0	0	546		1	1	1

Tabel 3-2: Tabel Perhitungan Performansi Konkordansi Kata

Kata	TP	FN	FP	TN		P	R	A
kemudian	17	0	0	547		1	1	1
berita	4	0	0	560		1	1	1
bumi	17	0	0	547		1	1	1
malam	17	0	0	547		1	1	1
sesuatu	4	0	0	560		1	1	1
yang besar	8	0	0	556		1	1	1
sekali-kali tidak	19	0	0	545		1	1	1

Tabel 3-3: Tabel Perhitungan Performansi Konkordansi Sinonim

Sinonim	TP	FN	FP	TN		P	R	A
kemudian	11	79	0	553		1	0.12	0.88
berita	2	0	0	562		1	1	1
bumi	15	0	0	549		1	1	1
malam	1	1	0	563		1	0.5	1
sesuatu	11	0	0	553		1	1	1
yang besar	11	0	3	553		0.786	1	0.99
sekali-kali tidak	0	0	0	564		1	1	1

Dari gambar 3-1 diatas, grafik (a) dan (b) diatas dapat dilihat bahwa konkordansi berdasarkan lema dan terjemahan kata dapat mencapai hasil yang memuaskan, namun berbeda dengan konkordansi berdasarkan sinonim atau kesamaan makna yang digambarkan pada grafik (c).

Pada konkordansi berdasarkan kesamaan makna, nilai recall sangat rendah, hal ini dikarenakan daftar sinonim yang digunakan saat pemrosesan kurang lengkap, akibatnya banyak ayat yang mengandung terjemahan kata bersinonim dengan kata kunci yang tidak ditampilkan.

3.2 Pengujian Kecepatan

Pengujian dilakukan untuk mengetahui kecepatan proses eksekusi perangkat lunak yang telah dibangun dibandingkan jumlah data dan jenis file dataset yang digunakan.

Tabel 3-4: Tabel Informasi Dataset

	Implementasi	Jumlah Kata				
		1 juz	2 juz	5 juz	10 juz	13 juz
.txt	conc plot dan n-gram	4170	9473	24885	49870	64443
.xls	konkordansi dan context word	53834	$2{,}4 \times 10^{11}$	$6{,}3 \times 10^{11}$	$1{,}2 \times 10^{12}$	$1{,}6 \times 10^{12}$

Tabel 3-5: Tabel Hasil Pengukuran Kecepatan Eksekusi

Data set	Konkordansi	Context Word	N-GRAM	Conc Plot
1 juz	12.37	9.07	15.36	2
2 juz	22.34	16.23	17.22	2.09
5 juz	64	57.66	19.66	4.66
10 juz	100.5	78.37	21.54	7.05
13 juz	116.07	87.34	23.06	8.03

Dari gambar grafik 3-2 dapat dilihat bahwa baik pada dataset berformat *.xls maupun *.txt, keduanya sama-sama berbanding lurus antara jumlah data dengan kecepatan eksekusinya.

Tidak hanya itu, jenis format file yang digunakan (baik berformat *.xls maupun *.txt) tidak berpengaruh terhadap kecepatan eksekusi, hal ini terlihat pada saat keduanya mengandung jumlah data yang hampir sama. Yaitu pada saat dataset berformat *.xls mengandung 53.834 kata (1 juz) dan dataset berformat *.txt hanya mengandung 53.791 kata (10 juz).

4. Kesimpulan

Berdasarkan hasil pengujian dapat disimpulkan bahwa pada konkordansi sinonim, nilai *precision* dipengaruhi oleh ketepatan sistem dalam mencari kata yang mengandung sinonim dari kata kunci. Semakin tepat dan lengkap data yang didapat maka nilainya akan semakin besar, semakin banyak yang tidak sesuai ikut dimunculkan sistem maka nilai *precision*-nya juga akan semakin kecil. Selain itu, pada kasus yang sama, nilai *recall* dipengaruhi oleh kelengkapan data sinonim kata berbahasa Indonesia. Semakin lengkap list dalam data sinonim maka akan semakin besar nilai *recall* yang didapat, begitu pula sebaliknya, list data sinonim yang kurang lengkap akan mengakibatkan rendahnya nilai *recall*. Tidak hanya itu, jenis format file yang digunakan pada dataset juga tidak berpengaruh terhadap kecepatan proses eksekusi. Kecepatan dipengaruhi oleh banyaknya dataset dan algoritma yang digunakan untuk pemrosesan

Untuk pengembangan lebih lanjut, sebaiknya menggunakan konsep similarity dengan memperhatikan kedudukan makna kata tersebut pada terjemahan ayat dalam penggunaan konkordansi berdasar sinonim agar

kemiripan makna benar-benar bisa relevan. Selain itu juga membangun dataset sinonim-hiponim yang lengkap dan disesuaikan dengan konteks bahasa Al-Quran untuk meningkatkan presentase nilai recall, mencari algoritma yang lebih efektif untuk mempercepat waktu eksekusi program, serta membangun *corpus* yang mengandung suatu index khusus yang memuat data konkordansi sehingga menghemat waktu eksekusi.

Daftar Pustaka

[1] Taniran Kencanawati, Penerjemahan Berdasar Makna: Pedoman untuk Pemadanan Antarbahasa. Jakarta: Penerbit Arcan, 1989.
[2] Departemen Pendidikan Nasional, Kamus Besar Bahasa Indonesia Edisi Ketiga. Jakarta: Balai Pustaka, 2010.
[3] Imelda S., "Sistem Pencarian Ayat Al-Qur'an Berdasarkan Terjemahan Bahasa Indonesia Dengan Pemodelan Ruang Vektor," M.S. skripsi, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekan Baru, 2013.
[4] Laurent Anthony, "Concordancing with AntConc: An Introduction to Tools and Techniques in Corpus Linguistics," presented at Proceedings of the JACET 45th Annual Convention, pp. 218-219, 2006.
[5] Daniel Jurafsky dan James Martin, Speech and Language Processing Second Edition. New Jersey: Pearson Prentice Hall,2016.
[6] Bachrul Ilmy, Pendidikan Agama Islam. Bandung: Grafindo Media Pratama, 2007.
[7] Hanna E, A Concordance of The Qur'an. Los Angles: University of Calofornia Press, 1983.
[8] Jati Sasongko, "Aplikasi untuk Membangun Corpus dari Data Hasil Crawling dengan Berbagai Format Data Secara Otomatis," M.S. skripsi, Universitas Stikubank, 2010.
[9] Tony McEnery and Andrew Wilson, Corpus Linguistics: An Introduction. Jerman: Edinburgh University Press, 2001.
[10] Yuswanto, Algoritma dan Pemrograman dengan Visual Basic Net 2005. Jakarta: Cerdas Pustaka Publisherm, 2008.
[11] Eko B Purwanto, Perancangan dan Analisis Algoritma. Yogyakarta: Graha Ilmu, 2008.
[12] Dony Wiranata, "Quranic Concepts Similarity Based on Lexical Database," M.S. skripsi, Universitas Telkom, Bandung, 2016.
[13] Asian Jelita, Williams Hugh, and Tahaghoghi S.M.M., Stemming Indonesia. M.S. journal, RMIT University, Australia, 2005.
[14] Faruq Tataran, "Aplikasi Panduan Kata dalam Mencari Ayat Al-Quran Juz 30 Berbasis Java Mobile," M.S. skripsi, Universitas Islam Negri Syarif Hidayatullah, Jakarta, 2010.
[15] A. M. H. B. a. I. K. Fatima Zahra Lahlou, "A Text Classification Based Method for Context Extraction from Online Reviews," Intelligent Systems: Theories and Applications (SITA), pp. 1-5, 2013.
[16] Halimah Zaman, Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access. Berlin: Springer. 2003
[17] John Sinclair, Guide to Good Practice Corpus and Text — Basic Principles. Los Angles: Tuscan Word Centre, 2004.