Pencarian Ayat Al-Qur'an yang Tidak Utuh Berdasarkan Kemiripan Fonetis

Putri Cendikia[1],Moch Arif Bijaksana [2], Kemas Muslim Lhaksmana

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]Divisi Digital Service PT Telekomunikasi Indonesia
[1]putricendikiaa@students.telkomuniversity.ac.id, [2]arifbijaksana@telkomuniversity.ac.id,
[3]kemasmuslim2@telkomuniversity.ac.id,

Abstrak Al-Qur'an merupakan kitab suci umat islam yang memiliki banyak kata didalamnya dengan penggunaan aksara arab. Aksara arab ini tidak berkesinambungan dengan aksara latin terutama latin Indonesia. Seiring perkembangan teknologi, di kembangkan sebuah sistem pencarian ayat Al-Qur'an berdasarkan kemiripan fonetis salah satunya adalah lafzi+ yang merupakan perkembangan dari sistem lafzi. Namun lafzi+ belum bisa menangani dengan baik ketika pengguna ingin mencari ayat yang tidak leng- kap atau tidak utuh dikarenakan kueri yang dimasukkan oleh pengguna tidak sama dengan kueri dalam korpus Al-Qur'an. Maka dari masalah tersebut dibuatlah pengembangan dari sistem lafzi+. Dengan me- lakukan pengindeksan trigram untuk kecocokan string antara kueri dengan transliterasi ayat Al-Qur'an serta menghitung pemeringkatan dokumen dengan batas ambang maka hasil yang di munculkan mampu mengeluarkan tujuan standar ayat yang ingin dihasilkan serta kemungkinan-kemungkinan ayat dari kueri masukkan. Pengujian menunjukkan bahwa sistem ini menghasilkan nilai lebih besar dari sistem-sistem sebelumnya yaitu dengan nilai Recall dan MAP sebesar 99,92% dan 91,40%. Sistem ini dapat menghasilkan . Sedangkan sistem sebelumnya mendapatkan hasil 16,74% dan 18%.

Kata kunci :Al-Qur'an, sistem pencarian, kemiripan fonetis, trigram

Abstract
The Qur'an is the Muslim holy book which has many words in it with the use of Arabic script. This Arabic script is not sustainable with Latin script, especially Latin Indonesia. As technology develops, a system for searching verses of the Qur'an is developed based on phonetic similarities, one of which is lafzi + which is the development of the lafzi system. However, lafzi+ cannot handle it well when users want to look for verses that are incomplete or incomplete because the query entered by the user is not the same as the query in the corpus of the Qur'an. Then from this problem the development of the lafzi+ system was made. By doing trigram indexing for matching strings between the query and transliteration of the verses of the Qur'an and calculating the ranking of documents with a threshold, the results that appear are able to issue the standard purpose of the verse to be produced as well as the possibilities of verses from the entered query. Tests show that this system produces a greater value than previous systems, namely the Recall and MAP values of 99.92 % and 91.40 %. This system can produce. Whereas the previous system got results of 16.74 % and 18 %.
Keywords: Al-Qur'an, search system, phonetic similarity, trigram.

1. Pendahuluan

Latar Belakang Al-Qur'an merupakan pedoman hidup bagi seluruh muslim yang ada di dunia salah satunya adalah Indonesia. Al-Qur'an memiliki arti sebagai "bacaan". Al-Qur'an tersusun dari puluhan ribu kata dan ratusan ribu huruf dengan penulisannya menggunakan aksara arab[2]. Begitu banyaknya ayat dalam Al-Qur'an membuat sebagian dari para pembaca kesulitan untuk mencari ayat yang ingin ditemukan dan penggunaan askara arab yang tidak berkesinambungan dengan aksara latin Indonesia.Pada saat seorang pembaca ingin menemukan sebuah ayat, namun pencarian tersebut tidak menggunakan aksara arab atau melakukan pencarian dengan menggunakan latin Indonesia. Maka dengan adanya kemajuan teknologi pada saat ini, dibuatlah sistem yang dapat melakukan pencarian ayat Al-Quran berbasis kemiripan fonetis. Hal ini berguna ketika pengguna ingin menemukan suatu ayat yang ingin dicari dengan menggunakan teks latin sesuai pelafalan Indonesia. Sistem tersebut adalah lafzi.

Lafzi merupakan sistem pencarian ayat berdasarkan kemiripan fonetis yang lebih sesuai dengan representasi pelafalan orang Indonesia. Sistem lafzi juga dapat membantu untuk menemukan ayat terpotong atau tidak lengkap. Namun, hanya beberapa pencarian saja atau ayat yang terpotong di bagian tertentu yang dapat di munculkan oleh sistem Lafzi,maka dari itu dibuatlah sistem pengembangan dari sistem lafzi dengan memiliki topik khusus yaitu menangani pencarian ayat yang tidak utuh atau terpotong yaitu lafzi+. Namun dalam sistem lafzi+ hanya bisa menangani ayat terpotong pada bagian tertentu saja seperti ”zalikal kitabu fihi” yang memiliki ayat lengkap ”zalikal kitabu la raibafihi”.Pada kasus dimana penghapusan 2 atau 3 kali lebih pada kueri lengkap tidak bisa di temukan pada sistem lafzi+. Dalam artian sistem lafzi dan lafzi+ belum bisa di implementasikan pada seluruh Al-Qur’an. Sistem tidak dapat mengeluarkan hasil pencarian dikarenakan kata yang hilang tersebut membuat kueri masukkan tidak selaras dengan korpus Al-Qur’an. Maka dalam penelitian ini dibuat sistem pencarian ayat Al-Qur’an yang tidak lengkap dimana seluruh inputan yang memiliki tujuan standar pencarian dapat ditemukan oleh sistem serta dapat memunculkan hasil pencarian yang memiliki persentase kemiripan yang tinggi.

Topik dan Batasannya Topik pada penelitian ini adalah mengembangkan sistem pencarian ayat Al-Qur’an berdasarkan kemiripan fonetis yang bisa menangani pencarian ayat yang tidak lengkap. Sebagai contoh, kata ”kitabu raifihi” yang memiliki kueri lengkap ”kitabu la raibafihi” seharusnya dapat ditemukan di surah Al-Baqarah ayat 2, namun pada sistem lafzi dan lafzi+ tidak ditemukan hasil tujuan standar yang seharusnya bisa ditampilkan dan sistempun tidak mengeluarkan opsi ayat lain yang mungkin memiliki kemiripan dengan contoh kueri tersebut.

Penelitian ini memiliki batasan agar terfokus pada tujuan. Adapun batasannya sebagai berikut:

1. Query Tidak mencakup lintas ayat yang query nya terdiri dari 2 ayat yang berbeda.

Tujuan Tujuan yang dicapai dalam tugas akhir ini adalah sebagai berikut:

1. Mengembangkan sistem pencarian ayat Al-Qur’an yang dapat menyelesaikan permasalahan inputan kueri ayat tidak lengkap yang dapat diimplementasikan dalam seluruh Al-Qur’an.
2. Melakukan perhitungan untuk nilai MAP dan Recall pada sistem yang dikembangkan guna mengetahui seberapa tingkat ke akurasian dari sebuah sistem dalam menemukan pencarian dan membandingkan nilai-nilai tersebut dengan sistem sebelumnya.

Organisasi Tulisan Organisasi tulisan yang ada pada tugas akhir ini adalah pendahuluan, studi terkait, sistem yang dibangun, evaluasi, kesimpulan dan daftar pustaka.Dalam pendahuluan menjelaskan latar belakang yang diangkat dan solusi dari permasalahan tersebut, serta berisikan rumusan masalah, topik dan batasan, serta tujuan dari tugas akhir ini. Pada studi terkait berisikan tentang studi atau literatur yang pernah digunakan untuk mendukung penelitian ini yaitu mengenai alih aksara arab ke latin, pencocokan string berdasarkan kemiripan fonetis, dan N-Gram. Gambaran sistem yang di bangun berisikian rancangan atau gambaran sistem secara umum dari pemrosesan query hingga mendapatkan hasil. Pada evaluasi berisikan tentang teknik pengeksekusian dari sistem, skenario pengujian, hasil dari sistem yang telah diuji, dan menganalisis hasil pengujian tersebut. Pada kesimpulan berisikan gagasan yang telah di capai pada penelitian ini. Daftar pustaka berisikan literatur penunjang dari penelitian tugas akhir ini.

2. Studi Terkait

2.1 Lafzi

Sistem lafzi dibuat sesuai dengan pelafalan orang Indonesia untuk melakukan pencarian ayat Al-Qur’an, dimana menggunakan pengkodean fonetis yang didasarkan pada pemadanan aksara Arab-Latin yang dipakai oleh orang Indonesia [6]. Pada sistem ini kueri masukan akan dibentuk kedalam sebuah pengkodean fonetis hingga pembentukan tokenisasi trigram. Misalkan teks latinnya adalah ”arrahman” lalu mengubah teks latin tersebut ke kode fonetis menjadi ”ARAHMAN” dengan trigramnya [ARA, RAH, AHM, HMA, MAN]. Lalu semua trigram akan dicari kedalam database hingga menemukan ayat yang mungkin menjadi tujuan standar pencarian.Terdapat sebuah variabel guna menyimpan data dan menampung ayat beserta urutan trigram yang muncul dalam ayat. Setelah itu menghitung selisih dari trigram dan urutan terpanjang pada kandidat yang di tampung pada variabel.Skor yang didapatkan akan diurutkan berdasarkan skor paling terbesar.

2.2 Lafzi+

Sistem lafzi+ adalah sistem kelanjutan dari lafzi. Sebagian besar sistem ini hampir sama dengan lafzi, namun pada lafzi+ untuk melakukan pencocokan terhadap trigram menguunakan metode KNN.Setelah membuat kode

fonetis dan trigramnya dari kueri masukan, maka akan melakukan pencocokan kueri dengan KNN antara kueri masukan dengan korpus Al-Qur'an. Algoritma K-NN bekerja berdasarkan jarak terpendek atau jarak terdekat. Setelah itu melakukan perhitungan selisih antara urutan trigram dan urutan panjang, maka kandidat ayat yang didapat tersebut akan diukur ke LCS agar mendapatkan kandidat terbaik [10].Perhitungan LCS ini dilakukan sebanyak 2 kali agar kandidat yang didapat memiliki nilai besar. Contohnya adalah kueri "zalikal kitabu fihi". Maka akan dilakukan perhitungan, setelah mendapatkan hasil tidak semua kueri yang dapat ditemukan.Contoh kueri yang di temukan adalah "zalikal kitabu" dan yang tidak ditemukan adalah "fihi". Maka kueri yang telah ditemukan tersebut akan disimpan dan kueri lainnya yang tidak ditemukan akan dilakukan perhitungan sampai semua kueri berhasil di temukan. LCS dapat memperhitungkan nilai kerapatan.

2.3 N-Gram

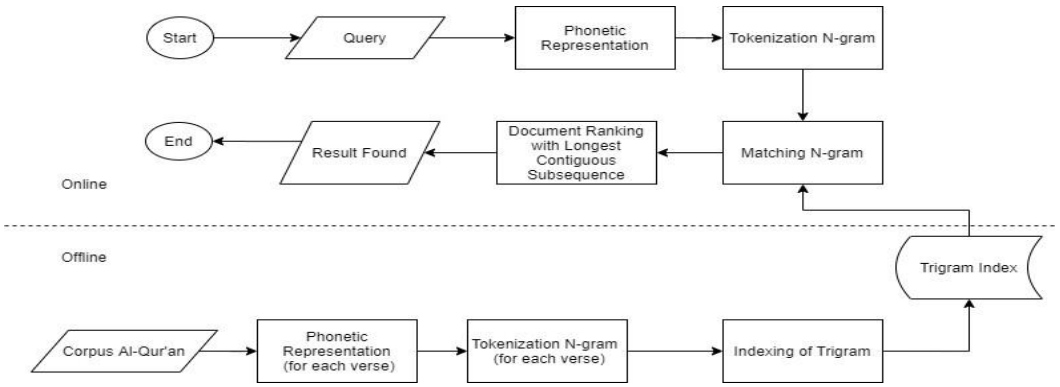
Biasanya, N-gram kata dapat digunakan untuk memperkirakan kata dalam urutan tertentu. Lebih tepatnya N-gram adalah wadah yang menampung kata dan memiliki karakter sebanyak n kata [4]. Namun pada pembangkitan karakter, N-gram dapat diartikan sebagai metode yang digunakan untuk pengambilan potongan huruf dari sejumlah n kata yang secara kesinambungan dibaca dari teks sumber sampai akhir dokumen yang di maksud. N-gram memiliki simbol pada awal dan akhir string guna mmbentuk N-gram yang utuh. Contohnya adalah "STORY" diuraikan menjadi beberapa N-gram:

- uni-gram : _S,T,O,R,Y_
- bi-gram : _ST,TO,OR,RY_
- tri-gram : _STR,TRO,ROY_
- quad-gram : _STRO,TROY_

Pada umumnya, string yang memiliki panjang *j* serta memiliki penanda, akan memiliki panjang j+1 bigram, j+1 trigram dan seterusnya.Manfaat lain pada N-gram yang bisa didapat adalah berdasarkarkan karakter N-gram sebagai komponen dari string [3], maksudnya adalah karena setiap string terpecah menjadi komponen kecil dan hal tersebut tidak terlalu terpengaruh terhadap kesalahan penulisan pada dokumen.

3. Sistem yang Dibangun

Pada bagian ini, alur sistem pencarian kueri latin tidak lengkap tergambarkan pada Gambar 1. Alur sistem ini di bagi menjadi dua bagian yaitu offline dan online. Proses offline merupakan proses dari korpus Al-Qur'an yang diubah kedalam pengkodean fonetis lalu melakukan tokenisasi menggunakan trigram yang dibuat sebagai acuan dokumen yang digunakan pada sistem pencarian ketika melakukan pencocokan terhadap kueri masukan dengan korpus Al-Qur'an. Pemrosesan online telah dibuat secara keseluruhan. Alur dari kedua bagian ini memiliki langkah yang sama, hanya saja alur yang sesungguhnya ada pada pemrosesan online yaitu pada pencocokan trigram dan pemeringkatan dokumen dengan *longest contiguous subsequence*.Alur sistem ini menggambarkan sistem dari kueri masukkan oleh pengguna hingga sistem menemukan hasil dari pencarian kueri tersebut. Pada saat kueri di masukkan, kueri tersebut akan di proses menjadi kode fonetis. Setelah itu akan dilakukan proses pencocokan tigram, dimana pencocokan tersebut dilakukan antara tokenisasi yang ada pada dataset dengan kueri masukan agar dapat diukur nilai kesamaannya. Setelah mengukur nilai kesamaan, lalu melakukan pemeringkatan dokumen dengan nilai kesamaan, nilai tertinggi akan menjadi hasil pencarian urutan pertama. Ketika hasil telah ditemukan maka akan di hitung nilai MAP dan recall terhadap masing-masing hasil pencarian kueri.



Gambar 1. alur sistem pencarian ayat Al-Qur'an berdasarkan kemiripan fonetis

3.1 Pengkodean Fonetis pada Kueri

Pengkodean fonetis dibuat guna menggambarkan aksara Arab dan penulisan pengucapan dalam aksara Latin ke suatu kode yang sama apabila pengucapannya serupa. Pengokodean fonetis juga memperhitungkan tanda baca Al- Qur'an (tajwid) yang sedikit berbeda dengan cara membaca teks berbahasa Arab biasa. Kueri masukan pada sistem pencarian ini berupa teks latin yang harus diubah kedalam kode fonetik agar dapat disesuaikan dengan pengkodean fonetis pada teks Al-Qur'an. Langkah-langkah prosedur praproses yang harus dilakukan secara berurutan pada kueri masukan sebagai berikut [6]:

1. Subsitusi Huruf Vokal
   Pada aksara Arab, hanya ada tiga kategori huruf vokal yaitu A, I dan U saja [1]. Sedangkan pada aksara latin terdapat penambahan vokal lainnya yaitu E dan O. Dalam hal ini vokal harus disubstitusi. Huruf vokal O, semisal pada "QOYYIMAL" diganti menjadi A sehingga kata tersebut menjadi "QAYYIMAL". Pada huruf vokal E, semisal pada "MUTTAQIEN" diganti menjadi I sehingga kata tersebut menjadi"MUTTAQIIN".

2. Menghilangkan Pengulangan Karakter
   Pada huruf yang ganda atau sama dan posisinya berdampingan maka huruf tersebut dijadikan satu. Tahap ini setara dengan menghilangkan penekanan pada huruf konsonan dan menghilangkan bacaan panjang pada huruf vokal atau penggabungan vokal.

3. Substitusi Diftong
   Dua Huruf vokal yang diucapkan sekaligus atau membentuk diftong disesuaikan dengan penyusunan diftong pada bahasa Arab. Caranya adalah dengan mengganti AI menjadi AY, serta AU diganti menjadi AW.

4. Substitusi Bacaan Iqlab
   Tata cara pada tajwid iqlab adalah melakukan pergantian bacaan huruf nun mati atau tanwin menjadi huruf mim apabila bertemu dengan huruf ba (ب)

5. Substitusi Bacaan Idgham
   Tata cara pada tajwid idgham yaitu pada saat huruf nun mati atau tanwin (ن) bertemu dengan huruf-huruf idgham, yaitu ي (ya), و (wau) , م (mim) , ن (nun) , ل (lam) , dan ر (ra). Dalam hal ini, Idgham diproses dengan cara menghilangkan huruf N apabila bertemu dengan huruf idgham.

6. Substitusi Bacaan Ikhfa

   Tata cara pada tajwid ikhfa yaitu pada saat huruf nun mati atau tanwin bertemu dengan huruf-huruf ikhfa yaitu ت (ta), ط (tha), د (dal), ف (fa), ق (qaf), ك (kaf), ص (shad), ض (dhad), ذ (dzal), ث (tsa), ج (jim), ش (syin), س (sin), ز (za), atau ظ (dza).menyembunyikan huruf N, terkadang bisa dituliskan atau dibaca dengan bunyi NG contohnya pada مَنْك (kuntum). Pada kasus ini, huruf G harus dihilangkan agar hasil dapat disetarakan dengan pengkodean pada teks Al-Qur'an.

7. Pemadanan Huruf
   Pada beberapa huruf arab yang memiliki 2 atau lebih huruf konsonan dari teks latinnya perlu dipertimbangkan agar dapat melakukan pemadaman huruf, aturan ini dilakukan agar dapat disesauikan dengan pengkodean teks Al-Qur'an. Misalkan huruf tsa dituliskan TS, dan huruf dza ditulisakan DZ [5]. Pemadanan huruf dituliskan dalam tabel 1.

Tabel 1. aturan pemadanan aksara latin

Aksara Latin	Padanan
SH, TS, SY	S
KH, CH	H
ZH, DZ	Z
DH	D
TH	T
GH	G
NG ('ain)	X
F, V, P	F
Q, K	K
J, Z	Z
',' (apostrof)	X

8. Penghilangan Spasi
   Seluruh spasi yang ada harus dihilangkan. Hal ini dilakukan agar data dapat disetarakan dengan hasil prosedur pada pengkodean teks Al-Qur'an.

Pengkodean fonetis pada teks latin memiliki prosedur yang dilakukan secara berurutan. Contoh penerapan prosedur pengkodean fonetis ini, tercantum pada tabel 2 pada kueri "kitabu raifihi" yang telah dirubah dalam huruf kapital. Pengkodean fonetis pada kueri ini dilakukan agar hasil yang didapat setara dengan pengkodean pada teks Al-Qur'an.

Tabel 2. contoh pengkodean fonetis kueri

Langkah	Hasil	Deskripsi
Teks Asli	kitabu raifihi	tidak melakukan substitusi vokal E dan O
1	KITABU RAIFIHI	tidak melakukan substitusi vokal E dan O
2	KITABU RAIFIHI	tidak melakukan penghilangan karakter ganda
3	KITABU RAYFIHI	melakukan substitusi diftong AI MENJADI AY
4	KITABU RAYFIHI	tidak melakukan substitusi bacaan idgham
5	KITABU RAYFIHI	tidak melakukan substitusi bacaan iqlab
6	KITABU RAYFIHI	tidak melakukan substitusi bacaan ikhfa
7	KITABU RAYFIHI	tidak melakukan pemadanan huruf
8	KITABURAYFIHI	spasi dihilangkan

3.2 Tokenisasi Trigram

Hasil dari pengkodean fonetis yaitu berupa string pada Al-Qur'an maupun kueri masukkan akan dilakukan tokenisasi untuk pembentukan trigram. Dalam pembentukan trigram tidak membutuhkan penanda awal atau akhir pada string, karena kueri masukkan adalah substring dari teks Al-Qur'an [6]. Pada contoh tabel 2, mendapatkan hasil tokenisasi sebanyak 11 trigram yaitu berupa {KIT,ITA,TAB,ABU,BUR,URA,RAI,AIF,IFI,FIH,IHI} sesuai dengan string "kitabu raifihi" yang telah dirubah kedalam kode fonetis.

3.3 Pencarian Indeks pada Trigram

Kueri yang telah sampai ke tahap tokenisasi trigram, maka trigram yang didapat tersebut akan dicari untuk setiap *inverted index*[9]. Trigram yang terindeks tersebut telah tersedia pada database.Informasi ini membuat perhitungan trigram menjadi lebih cepat. Seperti pada tabel berikut. Pada proses ini terdapat paling sedikit satu

Tabel 3. *inverted index* pada setiap trigram

Trigram	***Postings List***
TAB	*{*9" : [10],"23" : [28],"44" : [34],"45"[54],"51"[61],..*}*
ABU	*{*9" : [11],"14" : [72],"17" : [48],"37"[11],..*}*

trigram dari kueri masukkan. Trigram dari dokumen dan trigram dari kueri dibandingkan kemudian jumlah trigram yang sama dihitung[6]. Trigram yang memiliki jumlah sama paling banyak kemudian dikalikan dengan nilai *treshold* dimana hasilnya digunakan sebagai minimal skor.

3.4 Pemeringkatan Dokumen

Pada pemeringkatan dokumen, dilakukan pemberian skor pada hasil dari sebuah pencarian. Skor maksimum akan didapat ketika jumlah trigram pada kueri selaras dengan jumlah trigram pada dokumen. Sedangkan skor minimum didapat ketika jumlah trigram yang ditemukan paling sedikit hanya 1 trigram dalam dokumen. Dalam hal ini digunakan metode *Longest Contiguous Subsequence*. Metode*Longest Contiguous Subsequence A*[*k*1...*k*2] mendukung fungsi sorotan hasil pencarian yang lebih terkelompok. Skor kemunculan tidak hanya dilihat dari posisi kemunculan pertama query pada dokumen, namun seluruh posisi kemunculan juga diperhitungkan.

$$\text{skor} = \text{skor}(i < [i_1 \ldots i_2] - \begin{cases} < \text{maxGap}, \text{if } (i+1) - i > \text{maxGap}) \text{ if } \text{gap} = i+1 \\ < \text{maxGap}, \text{if } (\text{gap} > \text{maxgap}) \text{ if } \text{gap} \end{cases} \quad (1)$$

Membandingakan nilai maksimum gap atau kompetensi dengan list posisi angka setelah kurang posisi angka saat ini. Hal ini akan diulangi sebanyak jumlah list angka kurang 1. Jika perbandingannya melebihi nilai maksimum gap, maka posisi tersebut menjadi posisi awal. Misalnya diambil contoh pada tabel 3 setelah melakukan pencarian dan mendapatkan *postings list*,setiap trigram, maka data ayat tersebut akan disimpan pada sebuah variabel yang menampung sebuah ayat bersama dengan urutan di mana trigram muncul dalam ayat tersebut. Terdapat list pada

Tabel 4. indeks trigram pada sebuah ayat

Ayat	**Sequence**	**Sorting**
9	*{*10,11,21,8,9,16,22*}*	*{*8,9,10,11,16,21,22*}*

tabel 4 dengan ayat 9 yang menandakan bahwa kueri tersebut ada diurutan ke-9 dari seluruh ayat yang ada pada Al-Qur'an. Dimisalkan saja nilai max Gapnya 3, lalu setiap angka yang ada didalam list akan di kurangi sesuai urutan *sequence* tersebut semisal 9-8 = 1, 10-9= 1, 16-11= 4 maka angka yang diambil dimulai dari angka yang memiliki hasil yang lebih dari nilai max Gapnya. Maka kandidat list yang didapat adalah (16, 21, 22). Metode ini digunakan untuk mengurangi list angka dari inputan yang ada agar pencarian ayat mudah untuk ditemukan hasilnya. Kelemahan dari metode ini adalah tidak mengambil himpunan list sebelumnya yang lebih lengkap dari himpunan didapat setelahnya. Contohnya adalah (1,2,3,4,5,8,9,10) maka yang di ambil sebagai kandidat list adalah (8,9,10). Hal ini membuat penuruan pada tingkat akurasi walaupun sebenarnya ayat tersebut relevan dengan kueri masukan.

4. Evaluasi

Pengujian dilakukan dengan menggunakan data uji transliterasi latin seluruh Al-Qur'an. Memasukkan kueri latin yang beraneka ragam, dari kata acak sampai dengan kueri yang tidak utuh atau tidak lengkap. Pada pencarian tersebut ayat yang dihasilkan akan di hitung nilai *Mean Average Precision*MAP dan Recall guna mengetahui seberapa besar akurasi dari sistem yang dibangun. Recall digunakan sebagai ukuran dokumen yang relevan yang

dihasilkan oleh sistem[7]. Dimana jumlah dokumen relevan yang diambil dibagi dengan sejumlah dokumen yang relevan dalam korpus.Nilai dari recall menentukan tingkat keberhasilan sistem dalam menemukan hasilpencarian

atau tujuan standar dari kueri masukan. Nilai minimum dan maksimum recall adalah 0 dan 1. Jika nilai recall sistem adalah 1 ini menandakan bahwa sistem berhasil melakukan pencarian sesuai dengan dokumen relevan dalam korpus yaitu tujuan standar dari hasil pencarian. Recall dihitung dengan persamaan.

$$Recall = \frac{TP}{TP + FN} = \frac{relevant\ items\ retrieved}{relevant\ items} = P(retrieved \mid relevant) \quad (2)$$

FN adalah semua item relevan namun tidak dihasilkan oleh sistem dan TP adalah semua item relevan yang dihasilkan oleh sistem.

Nilai *mean average precision* (MAP) merupakan nilai rata-rata dari *average precision* (AP). AP merupakan perhitungan untuk item relevan yang di ambil dengan patokan dokumen relevan dalam korpus[8]. Setiap item relevan yang tidak dihasilkan oleh sistem akan mendapatkan nilai 0 dan nilai presisi dihitung dari keterurutan item dari daftar hasil sistem.

Tabel 5. Tabel contoh untuk menghitung AP

Output ke-	Benar/Salah	Presisi	Keterangan
1	Benar	1/1	Presisi pada urutan 1
2	Salah	-	Tidak presisi
3	Benar	2/3	Presisi pada urutan 3
4	Benar	3/4	Presisi pada urutan 4
5	Salah	-	Tidak presisi

Dari tabel contoh diatas, maka didapatkan skor AP yaitu $\frac{1+\frac{2}{3}+\frac{3}{4}}{3} = 81$

$$A\blacksquare = \sum_{\blacksquare=1}^{N} \left(\frac{\blacksquare\blacksquare_{\blacksquare}}{\blacksquare\blacksquare\blacksquare a\blacksquare\blacksquare_{\blacksquare}}\right) \quad (3)$$

$$\blacksquare A\blacksquare = \frac{1}{|\blacksquare|} \sum_{\blacksquare=1}^{Q} A\blacksquare(\blacksquare_1) \quad (4)$$

Lafzi++ (tugas yang dibuat) adalah sistem yang telah diperbarui dan ada penambahan dari sistem sebelumnya. Melakukan pengujian terhadap sistem lafzi, lafzi+ dan lafzi++ dengan data uji yang dibuat. Menghitung dan membandingkan nilai akurasi dari hasil pencarian dari sistem-sistem yang diuji.

4.1 Hasil Pengujian

Hasil pengujian ini menunjukkan tingkat akurasi dari masing-masing sistem dengan melakukan pengujian kueri yang telah dibuat oleh peneliti.

Tabel 6. Tabel hasil pengujian menggunakan evaluasi *recall* dan *MAP* dengan nilai ambang batas 0.62

Sistem	Kueri Tidak lengkap (skenario pertama)		Kueri Lengkap (skenario kedua)	
	Recall	MAP	Recall	MAP
Lafzi	46,9%	49%	95,25%	94,97%
Lafzi+	16,74%	18%	91,74%	93,37%
Lafzi++ (this work)	99,92%	91,40%	100%	100%

Pada hasil pengujian ini terlihat bahwa nilai dari masing-masing recall dan MAP sistem sangat berbeda. Lafzi++ mampu menemukan tujuan standar dari pencarian atau menemukan kemungkinan hasil dari pencarian lebih banyak dari sistem-sistem sebelumnya. Hal ini di karenakan pemeringkatan dokumen dan nilai *treshold* yang digunakan baik untuk pencarian tersebut.Dalam pengujian pertama yang dilakukan dengan kueri tidak lengkap

terhadap sistem. Pengujian mendapatkan hasil nilai MAP 91,40% dan recall 99,92%. Pada MAP yang Jika melakukan perbandingan dengan sistem lafzi yang memiliki nilai MAP dan recall sebesar 49% dan 46,9% dan sistem lafzi+ yang memiliki nilai MAP 18% dan recall 16,74% maka terlihat bahwa sistem lafzi++ lebih mengungguli dari sistem-sistem sebelumnya. Hal ini dikarenakan sistem lafzi dan lafzi+ belom bisa menangani permasalahan kueri masukan tidak lengkap secara keseluruhan, terutama ketika dilakukan penghapusan sebanyak 2 atau 3 kali lebih pada kueri yang relevan dalam korpus.

Sebelumnya telah melakukan pengujian terhadap nilai treshold untuk mencari nilai treshold optimal pada sistem.Data yang di gunakan untuk pengujian adalah data yang sama dengan data uji sebelumnya. Ada tiga nilai yang diambil untuk dijadikan pengujian yaitu nilai 0,8 dengan mendapatkan nilai MAP dan recall sebesar 49.80% dan 44.64%, dengan percobaan nilai 0.7 mendapatkan nilai MAP dan recall sebesar 66.28% dan 67.08%. Terhadap nilai-nilai yang didapatkan terlihat bahwa ketika nilai semakin rendah maka nilai kerapatan dan akurasi dari hasil sistem semakin besar. Hal inilah yang digunakan sebagai acuan mengapa di ambilnya nilai optimal treshold pada permasalahan ini sebesar 0.62.

Tabel 7. Tabel Hasil Percobaan untuk Mencari Nilai Treshold yang Optimal

Nilai Ambang Batas	Recall	MAP
0.8	44.64%	49.80%
0.7	67.08%	66.28%
0.62	99.92%	91.40%

4.2 Analisis Hasil Pengujian

Pengujian dilakukan dengan 100 kueri dengan sebanyak 50 kueri tidak lengkap sebagai skenario pertama dan 50 kueri yang lengkap sebagai skenario kedua. Dokumen yang di jadikan sebagai acuan atau patokan untuk melakukan perbandingan atau pencocokan pada kueri masukan di ambil dari *Corpus Al-Qur'an* dengan melakukan pencarian dari kata per kata. Setelah itu mencari ayat yang memiliki kesamaan pada contoh kueri yang akan di uji dalam sistem. Ayat yang diambil hanya ayat yang memiliki kemiripan 100% dengan contoh kueri yang di ujikan. Diberikan beberapa contoh kueri yang dijadikan sebagai data uji pada sistem. Pada tabel 6 menunjukkan skenario pertama yaitu pengujian dengan kueri tidak lengkap. Tabel skenario pertama dan kedua berisi aksara Arab beserta teks latin dari aksara terebut dan keterangan (surah:ayat) dari hasil pencarian ayat dengan sistem Lafzi++.

Pada tabel 6 terlihat nilai MAP dan recall dari sistem lafzi++. Untuk nilai MAP mendapatkan 91,23% dika renakan beberapa kueri yang di ujikan kepada sistem memiliki tingkat akurasi dibawah nilai maksimal yaitu 1. Hal ini dipicu oleh perbedaan atau tidak relevannya antara hasil pencarian dengan korpus Al-Qur'an akibatnya urutan dari hasil pencarian tersebut menjadi berbeda dengan korpus. Contohnya adalah kueri "illaha samin alim" yang sebenarnya terdapat pada surah dan ayat (8 : 17),(49 : 1),(2 : 227),(2 : 181). Namun dalam hasil pencarian yang didapat adalah (49 : 1),(2 : 181),(2 : 227),(2 : 244),(8 : 17),(8 : 53),(17 : 82),(8 : 42),(12; 51),(22 : 25),(2 : 158), (51 : 28), (2 : 255), (40 : 40), (2 : 256), (7 : 158), (35 : 3), (11; 82), (15 : 74). Terlihat bahwa (2:244) tidak ada dalam korpus Al-Qur'an sedangkan dalam pencarian dimunculkan, hal ini lah yang menyebabkan nilai akurasi dari kueri ini tidak mencapai batas maksimal dan hanya mendapatkan nilai 0,95.

Terhadap nilai Recall yang didapatkan yaitu sebesar 99,97% hal ini dikarenakan ada hasil pencarian kueri yang tidak ditemukan oleh sistem namun sebenarnya ayat tersebut ada pada korpus Al-Qur'an yaitu "inalazi kafaru" yang seharusnya ada pada (2 : 6),(2 : 161),(2 : 213),(3 : 4),(3 : 10),(3 : 21),(3 : 55),(3 : 90),(3 : 91),(3 :116), (3 : 127), (4 : 56), (4 : 137), (4 : 167), (4 : 168), (5 : 36), (5 : 78), (8 : 36), (8 : 65), (13 : 33), (22 : 25), (40 : 10), (41 : 41), (47 : 32), (47 : 34), (57 : 15), (98 : 6) namun dalam pencarian hanya mendapatkan (41 : 41), (4 : 168), (2 : 6), (2 : 161), (3 : 4), (3 : 10]), (3 : 90), (3 : 91), (3 : 116), (4 : 56), (4 : 167), (5 : 36), (5 : 78), (8 : 36), (22 :25), (40 : 10), (47 : 32), (47 : 34), (98 : 6), (41 : 27), (5 : 103), (3 : 127), (57 : 15), (5 : 17), (5 : 72), (5 : 73), (3 : 55), (8 : 65), (2 : 213), (13 : 33), (4 : 97]), (2 : 101), (3 : 183), (3 : 77), (2 : 144), (2 : 147), (2 : 218), (4 : 137). Dari jumlah yang didaptkan oleh korpus sebanyak 27 surah dan ayat sistem hanya bisa menemukan 26 surah dan ayat,hal ini mengakibatkan nilai recall dari kueri tersebut menjadi 0,96 yang nilai tersebut kurang dari nilai maksimal.

Ada beberapa kueri yang tidak dapat ditemukan oleh sistem-sistem sebelumnya salah satunya adalah kueri "ki tabu raifihi" seharusnya dapat ditemukan pada surah Al-Baqarah ayat 2. Penghapusan beberapa huruf dalam kueri masukan membuat hasil trigram menjadi berbeda dengan trigram yang ada pada Al-Qur'an. Hal ini menyebabkan kueri masukkan tidak dapat ditemukan oleh sistem lafzi dan lafzi+. Setelah melakukan pemeringkatan dokumen dan menentukan optimal nilai treshold yang tepat untuk sistem, lafzi++ memiliki kenaikan pada nilai recall dan MAP. Hal ini juga membuat kemunculan hasil dari pencarian menjadi lebih banyak dari sistem-sistem sebelumnya.

Tabel 8. contoh kueri untuk tahapan pertama

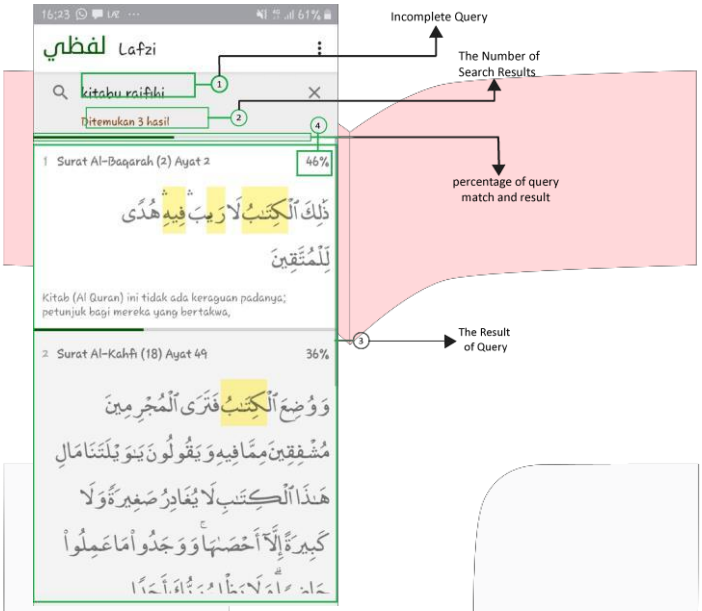
No	Aksara Arab	Teks Latin	Kemunculan pada Sistem Lafzi++ (surah:ayat)
1	السفهاء ولكن يعلمون	sufaha walakil yalamun	2:13, 14:11, 2:235, 7:155, 28:57, 12:68, 2:251, 2:102, 4:5, 6:94, 12:21, 30:56, 39:49, 4:12, 4:36, 9:40, 7:150, 7:33, 2:282, 10:5, 9:123, 9:36, 2:213, 60:1
2	الكتب ريب فيه	kitabu raifihi	2:2, 18:49, 6:91
3	هو يزكيكم	huwa ziyusa rukum	3:6, 10:22, 33:43, 8:72, 6:60, 2:85, 2:25, 40:28, 4:12, 10:4, 5:60, 4:11
4	في الحمد فياشا	fil hami fayasa	3:6,35:1,9:7,7:129,5:64, 5:91, 2:196,20:97,23:33, 13:13,30:9, 35:44,40:21, 40:82,2:282,5:31, 40:78,2:144,2:217,2:27,7:53, 22:5,24:33
5	ملكت ايمانكم	malakat makum	30:28, 22:36, 33:71, 6:50, 24:15, 5:3, 2:134, 2:141, 2:187, 7:10, 7:39, 7:11, 6:119, 27:60, 33:53, 57:10, 33:49, 16:71, 33:55, 8:72, 4:3,4:24, 4:25, 4:36, 4:88,7:59, 7:65,7:73, 7:85, 9:38, 10:35, 11:50,11:61, 11:84, 14:44, 23:23, 24:31, 24:33, 24:58, 29:25, 32:4, 33:50, 33:52, 47:30, 57:8, 71:13, 4:171

Tabel 9. contoh kueri untuk tahapan Kedua

No	Aksara Arab	Teks Latin	Kemunculan pada Sistem Lafzi++ (surah:ayat)
1	السفهاء ولكن لّا يعلمون	sufaha walakil yalamun	2:13,34:36,2:272,7:155,12:21, 2:251,2:102,2:235, 2:282,9:40
2	ألكتب لَّا ريب فيه	kitabu la raiba fihi	2:2,10:37,32:2,3:9,3:25, 4:87,6:12,17:99,22:7,40:59,42:7, 45:26,45:32,18:21,7:54
3	هو الذي يصوركم	huwallazi yusawwirukum	3:6,10:22,33:43,6:60,2:264, 2:282,59:24,39:5
4 yasya	في الرحام كيف يشاء	fil ar ḥami kaifa	3:6,22:5,5:64,12:109,30:9, 35:44,40:21,40:82,47:10, 7:129,2:228,5:31,73:20
5	ملكت ايمانكم	malakat aimanukum	30:28,24:33,4:36,24:58,4:3,4:24, 4:25,16:71,33:55,23:6,24:31, 33:50,70:30,4:33,2:225,5:89, 9:13,4:135,16:92,16:94, 5:53,66:6,66:8,2:246,9:12,22:5, 18:19,41:44,2:249,5:3,11:27,2:282

Pengujian kedua dilakukan dengan kueri masukkan yang lengkap sesuai dengan lafaz aslinya. Kueri yang dimaksudkan adalah kueri yang tidak ada penghapusan ditengah kata, huruf ataupun kalimatnya. Sebagai contoh adalah ألكتب لَّا ريب فيه(kitabu la raiba fihi). Skor pemeringkatan dokumen menjadi terurut sesuai dengan seberapa kemungkinan kueri muncul dalam ayat tersebut. Urutan awal pada dokumen hasil pencarian pada sistem

sebelumnya yaitu lafzi+ adalah ayat yang memiliki *highligh* terpanjang, namun hal ini tidak menentukan bahwa seberapa kemungkinan kueri muncul dalam ayat tersebut. Urutan awal pada dokumen hasil pencarian pada sistem sebelumnya yaitu lafzi+ adalah ayat yang memiliki *highligh* terpanjang, namun hal ini tidak menentukan bahwa urutan awal adalah tujuan standar yang di cari oleh pengguna karna tidak ada alat ukur untuk mengetahui seberapa besar kemiripan antara kueri dengan hasil pencarian. Maka dari itu sistem lafzi++ membuat sebuah persentase keterurutan ayat dari hasil pencarian serta menentukan *highligh* yang benar sesuai dengan kueri masukkan. Urutan pertama pada ayat adalah ayat yang memiliki kedekatan dengan kueri paling besar nilai maksimal persentasenya adalah 100%, dalam artian keseluruhan dari kueri masukkan ada pada ayat tersebut lalu diikuti oleh pengurutan persentase lainnya sampai kemungkinan paling terkecil akan muncul dalam hasil pencarian. Urutan awal mungkin saja sebagai tujuan standar dari kueri yang dimasukkan. Diberikan contoh kueri yang dijadikan sebagai data uji pada sistem yang memiliki tujuan standar dari pencarian dan menunjukkan jumlah hasil dari pencarian.



Gambar 2. Aplikasi Lafzi++

Tabel 10. contoh kueri untuk percobaan aplikasi

Kueri Tidak Lengkap	Tujuan Standar	Junlah Hasil Pencarian (surah:ayat)
Kitabu raifihi	Kitabu la raifihi	3

Pada Gambar 2 menunjukkan sistem Lafzi++ ketika melakukan pencarian ayat. Sistem memiliki fitur kueri masukan, jumlah hasil pencarian, akurasi kemiripan, dan highligh serta hasil dari pencarian tersebut yang berisikan nama surah dan ayat Aksara Arab beserta terjemahan dari ayat tersebut. Pada kueri ”kitabu raifihi” yang memiliki jumlah hasil pencarian sebanyak 3 surah dan ayat. Pada hasil pencarian pertama menunjukkan tingkat kemiripan pada kueri sebanyak 46%. jumlah dari besaran persentase tersebut merupakan implementasi dari highligh yang ada. Highligh berguna untuk menunjukkan kepada pengguna bahwa kueri yang di masukan ada pada hasil pencarian yang di maksudkan

5. Kesimpulan

Dari hasil analisis pengujian maka dapat di tarik kesimpulan :

1. Sistem lafzi++ dapat mengeluarkan tujuan standar pada seluruh kueri dalam data uji daripada lafzi dan lafzi+ serta dapat memunculkan presentase kemiripan dari ayat yang ditemukan dengan kueri yang di masukan.
2. Pada sistem lafzi++ ini memiliki nilai recall 99,78% dan MAP 91,23%. Nilai tersebut jauh lebih baik dari sistem pencarian ayat sebelumnya.

Saran untuk penelitian selanjutnta adalah mencari metode untuk pemeringkatan dokumen agar nilai yang di keluarkan dari MAP jauh lebih baik atau sempurna. Menambahkan fitur *suggestion* atau fitur saran ketika pengguna melakukan kesalahan dalam pengetikkan.

Daftar Pustaka

[1] S. Hadi, S. Chamamah, M. Ramlan, and I. D. P. Wijana. Perubahan fonologis kata-kata serapan dari bahasa arab dalam bahasa indonesia. *Jurnal Humaniora*, 15(2):121–132, 2003.

[2] B. Hammo, A. Sleit, and M. El-Haj. Effectiveness of query expansion in searching the holy quran. In *The Second International Conference on Arabic Language Processing CITALA'07*, pages 1–10, 2007.

[3] A. Hamzah. Deteksi bahasa untuk dokumen teks berbahasa indonesia. In *Seminar Nasional Informatika (SEMNASIF)*, volume 1, 2015.

[4] A. Hanafi. Pengenalan bahasa suku bangsa indonesia berbasis teks menggunakan metode n-gram. *IT TEL-KOM*, 2009.

[5] A. M. Ismail and M. U. Nawawi. Pedoman ilmu tajwid. *Surabaya: Karya Abditama*, 1995.

[6] M. A. Istiadi. Sistem pencarian ayat al-quran berbasis kemiripan fonetis. *Final Project IPB. Bogor*, 2012.

[7] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[8] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and trends in Information Retrieval*, 3(1—2):1–224, 2009.

[9] H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[10] A. Zafran, M. A. Bijaksana, and K. M. Lhaksmana. Truncated query of phonetic search for al qur'an. In *2019 7th International Conference on Information and Communication Technology (ICoICT)*, pages 1–4. IEEE, 2019.