Preprocessing Pendukung Information Retrieval Melalui Pembentukan Korpus Al-Quran Terjemahan Bahasa Indonesia

Irfan Humaini, Witta Listiya Ningrum
Universitas Gunadarma, Sistem Informasi
Jl. Margonda Raya 100, 16424, Depok, Indonesia
Irfan_humaini@staff.gunadarma.ac.id

***Abstrak*—Al Quran merupakan kitab suci umat Islam, merupakan pedoman bagi seluruh umat Islam. Al Quran sudah diterjemahkan ke berbagai bahasa salah satunya adalah bahasa Indonesia. Data set Al Quran agar layak digunakan dalam proses *information retrieval* perlu dilakukan *preprocessing* karena pada proses *information retrieval* proses dilakukan berdasarkan kata yang dianggap penting saja dalam suatu dokumen yang berbentuk token dan merupakan kata dasar. Proses tersebut secara umum terdiri dari *tokenizing, stopword removal* dan *stemming*, dengan proses yang dilakukan tersebut hasilnya berupa korpus Al Quran terjemahan bahasa Indonesia yang layak dalam proses *information retrieval* Al Quran Terjemahan Bahasa Indonesia.**

Kata kunci - Pre processing; tokenizing; stopword removal; stemming;

I. PENDAHULUAN

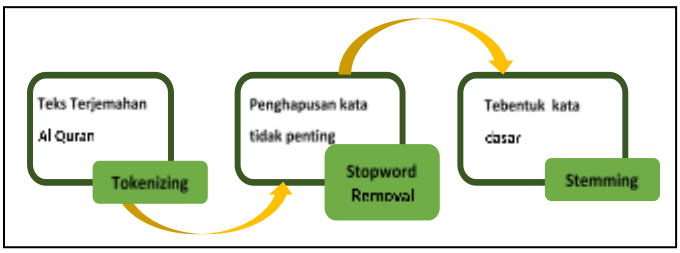
Al Quran adalah kitab suci umat Islam, *"Kitab ini tidak ada keraguan padanya, petunjuk bagi mereka yang bertakwa"* (Al Quran. Al-Baqarah:2). Hal-hal yang terkandung di dalam Al Quran berhubungan dengan keimanan, ilmu pengetahuan, hukum, peraturan-peraturan yang mengatur tingkah laku dan tata cara hidup manusia, kisah-kisah umat sebelumnya, ibadah serta tauhid (pengesaan Allah). Al Quran terdiri dari 30 Juz, 114 Surat dan 6326 Ayat, sehingga untuk mencari ayat yang sesuai dengan yang diinginkan pengguna berdasarkan kata kunci akan sulit dan memakan waktu. Berdasarkan hal tersebut maka dikembangkan sistem *information retrieval* untuk mempermudah pencarian ayat yang sesuai dengan keinginan pengguna.

Text operations (pengoperasian teks) adalah proses transformasi dokumen dan *query* menjadi kata-kata indeks. Dalam suatu dokumen terdapat beberapa kata yang memiliki makna lebih penting dibandingkan kata-kata lainnya, sehingga *preprocessing* terhadap teks dalam suatu koleksi dokumen dianggap perlu dalam menentukan kata yang akan digunakan sebagai *index terms* [2]. Tahap *preprocessing* mencakup pengoperasian teks seperti penghapusan *markup*, penghapusan *stopwords*, dan *stemming* (pembentukan kata dasar)[3][4]. *Stem* adalah bagian kata yang tinggal setelah menghilangkan imbuhannya (awalan dan akhiran) [1].

Dataset yang digunakan pada penelitian ini bersumber dari terjemahan kementrian Agama Republik Indonesia. Pada dataset tersebut selanjutnya dilakukan *pre-processing*. Proses *pre-processing* meliputi *tokenizing, filtering,* pembentukan *Inverted Index* dan *stemming* [4]. Hal ini dilakukan agar proses *information retrieval* dapat dilakukan dengan baik atau informasi yang dihasilkan sesuai dengan yang diharapkan, dimana hasil akhir information retrieval akan menampilkan atau menghasilkan ayat-ayat Al Quran berdasarkan kata dasar dari kata kunci yang dimasukkan, sehingga hasil information retrieval menjadi lebih luas dan lengkap dibanding jika tidak menggunakan kata dasar.

II. METODE

Pada tahapan ini data set Al Quran agar layak dalam proses *information retrieval* dilakukan *preprocessing* karena pada proses *information retrieval* proses yang dilakukan berdasarkan kata yang dianggap penting saja dalam suatu dokumen dan dalam bentuk kata dasar. sebagai contoh jika pengguna ingin mencari ayat yang berhubungan dengan kata *"Beriman"* jika tidak melalui *preprocessing* hanya ayat yang mengandung kata *"Beriman"* saja yang ditampilkan sebagai hasil pencarian *information retrieval* sedangkan ayat yang mengandung kata *"iman", "mengimani", "diimani"*, dan lain-lain tidak menjadi hasil pencarian karena kata yang dicari adalah "Beriman", jika telah melalui *preprocessing*, maka hasil pencarian *information retrieval* akan menampilkan seluruh ayat yang memiliki kata dasar *"iman"*



Gambar 1. Tahapan Preprocessing

Tahapan pada gambar 1 dijelaskan sebagai berikut:

1) Pada proses tokenizing yang dilakukan terhadap seluruh ayat Al Quran adalah menghilangkan seluruh tanda baca dan seluruh kata diproses menjadi huruf kecil semua [7][8].

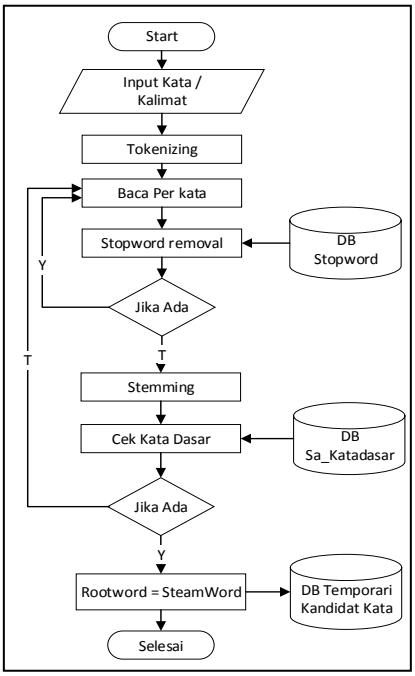
2) Proses stopword removal, menghilangkan kata-kata yang dianggap tidak penting, proses ini dilakukan berdasarkan kamus stopword yang terdapat dalam basis data stopword [5].

3) Proses Stemming adalah menjadikan seluruh kata menjadi kata dasar, algoritma stemming yang digunakan pada penelitian ini adalah algoritma Stemming ECS. Lebih jelasnya preprocessing dapat dilihat pada gambar 2 dan kemudian dijelaskan lebih rinci pada gambar 3 dan gambar 5.

III. HASIL DAN DISKUSI

A. Tahapan Preprocessing

Preprocessing dimulai dari tahapan input kata atau kalimat yang kemudian masuk ke dalam proses *tokenizing*, kemudian akan dilakukan proses baca kata per-kata sebelum masuk ke dalam proses stopword removal untuk membuang kata yang ada dalam *stopword* dengan membandingkan kata yang ada pada database *Stopword* hingga didapat kata-kata yang tidak terdapat pada DB_*Stopword* untuk masuk ke dalam proses *stemming* dan lanjut ke tahap pengecekan kata dasar yang dibandingkan dengan database Sa_Katadasar dan jika pada hasil akhir ditemukan, maka akan disimpan pada database Temporari Kandidat Kata Dasar. Proses ini dapat dilihat pada gambar 2.

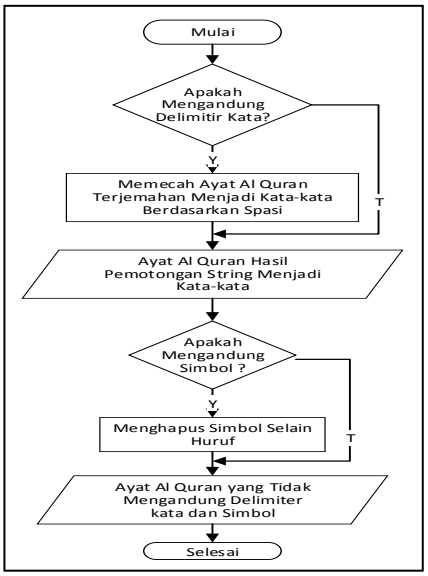


Gambar 2. Skema Preprocessing

B. Tokenizing

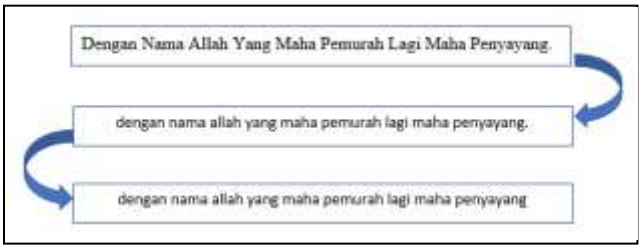
Tokenizing merupakan proses pemisahan sebuah teks menjadi kata, *frasa*, *symbol* atau elemen bermakna lain yang disebut *Token*. Proses untuk melakukan *Tokenizing* adalah sebagai berikut :

- Setelah input kalimat diterima, sistem akan mengubah semua karakter huruf besar menjadi huruf kecil.
- Sistem selanjutnya akan menghilangkan tanda baca yang ada di dalam kalimat.
- Akan dihasilkan kumpulan kata penyusun kalimat atau Terms.
- Proses Tokenizing selesai.



Gambar 3. Proses Tokenizing

Gambaran hasil proses *Tokenizing* dapat dilihat pada gambar 4 berikut.

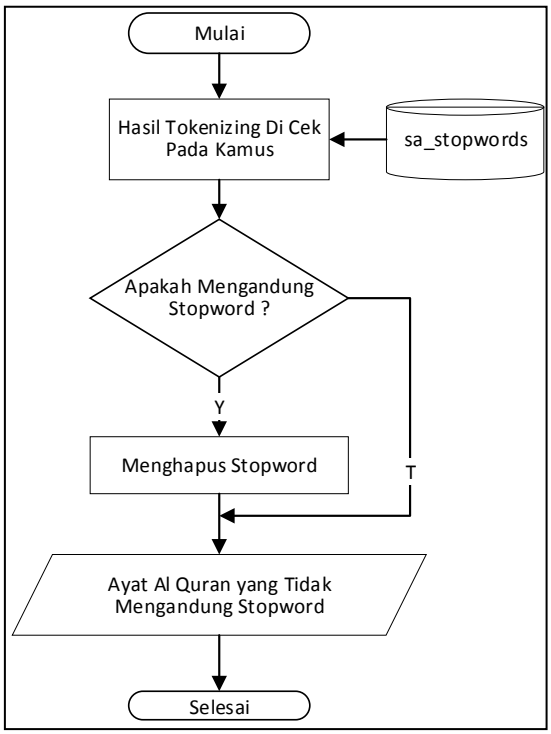


Gambar 4. Hasil Proses *Tokenizing*

C. Stopword Removal

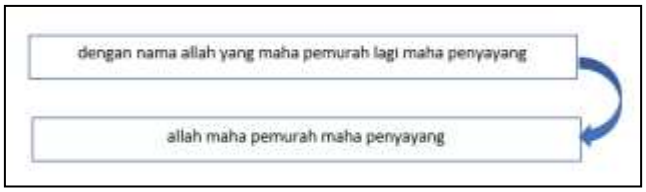
Proses *stopword removal* atau *filtering* akan menghapus, menghilangkan atau membuang kata-kata yang tidak penting dan tidak memiliki makna pada kalimat input. Langkah-langkah untuk melakukan proses *stopword removal* dapat dilihat pada gambar 5. *Output* kata-kata dari proses *Tokenizing* akan digunakan sebagai input, selanjutnya sistem akan membandingkan setiap kata-kata tersebut dengan setiap *stopword* yang terdapat di dalam *database*. Apabila terjadi kesamaan kata dengan kata yang terdapat di dalam *stopword* database, maka kata tersebut dihilangkan. Apabila berbeda,

maka kata tersebut akan disimpan dan untuk selanjutnya akan digunakan sebagai *input* pada proses selanjutnya.



Gambar 5. Proses Stopword Removal

Gambaran hasil proses *Stopword Removal* seperti gambar 6 berikut.



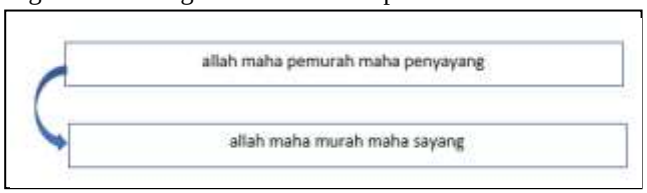
Gambar 6. Hasil Proses Stopword Removal

D. Stemming

Proses *stemming* akan menemukan kata dasar dari sebuah kata *(root word)* dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan *confixes* (kombinasi dari awalan dan akhiran) pada kata turunan. Pada dasarnya, algoritma *stemming* dengan beberapa cara berikut:

- *Particle* (P) atau partikel, termasuk di dalamnya adalah partikel "-lah", kah", "-tah", dan "-pun".
- *Possessive Pronoun* (PP) atau kata ganti kepunyaan, termasuk di dalamnya adalah "-ku" , "-mu", dan "-nya".
- *Derivation Suffixes* (DS) yakni kumpulan akhiran yang secara langsung dapat ditambahkan pada kata dasar. Termasuk di dalam tipe ini adalah akhiran "-i", "-kan", dan "-an".
- *Derivation Prefixes* (DP) yakni kumpulan awalan yang dapat langsung diberikan pada kata dasar murni, atau pada kata dasar yang sudah mendapatkan penambahan sampai dengan dua awalan. Termasuk di dalamnya adalah awalan yang dapat bermorfologi ("me-", "be-", "pe-", dan "te-") dan awalan yang tidak bermorfologi ("di-", "ke-" dan "se ").

Gambaran hasil proses *Stopword Removal* dapat dilihat pada gambar 7 sebagai hasil akhir dari penelitian.



Gambar 7. Hasil Proses Stopword Removal

IV. KESIMPULAN

Berdasarkan penelitian yang dilakukan, penyusunan korpus Al Quran terjemahan Bahasa Indonesia sudah berhasil dilakukan dengan rancangan proses dari *preprocessing*. Hal ini berfungsi untuk mendukung *information retrieval* sehingga data atau korpus Al Quran terjemahan bahasa Indonesia dapat diproses dengan baik. *Preprocesing* yang baik akan dapat meningkatkan kualitas informasi sehingga hasil *information retrieval* menjadi lebih relevan dan presisi.

DAFTAR PUSTAKA

[1] Agusta, Ledy, "Perbandingan Algortima Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia", Universitas Kristen Satya Wacana, 2009.

[2] Akram Roshdi, Akram Roohparvar, "Review: Information Retrieval Techniques and Applications", International Journal of Computer Networks and Communications Security, VOL. 3, NO. 9, SEPTEMBER 2015, 373–377.

[3] Bunyamin, Hendra, "Algoritma Umum Pencarian Informasi Dalam Sistem Temu Kembali Informasi Berbasis Metode Vektorisasi Kata dan Dokumen", Jurnal Informatika UKM, No. 2, Vol. I, hal 85-87, 2005.

[4] Broto Poernomo T.P , Ir. Gunawan, "Sistem Information Retrieval Pencarian Kesamaan Ayat Terjemahan Al Qur'an berbahasa Indonesia dengan Query Expansion dari tafsirnya". IDeaTech 2015, ISSN: 2089-1121, 2015.

[5] Manning, Christopher D., Prabhakar Raghavan, "Introduction to Information Retrieval". Cambridge University Press, Cambridge, England, 2009.

[6] Surya Agustian, Imelda Sukma Wulandari, "Sistem Qur'an Retrieval Terjemahan Bahasa Indonesia berbasis Web dengan reorganisasi Korpus", KNSI 2013, ISBN 978-602-17488-0-0, 2013.

[7] Miftah, Andriansyah, "Developing Indonesian corpus of pornography using simple NLP-text mining (NTM) approach to support government anti pornography program", Conference Paper, 2017.

[8] Tony McEnery & Costas Gabrielatos, "English Corpus Linguistics", 2006.