Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

Team project ©2017
Dony Pratidana S. Hum | Bima Agus Setyawan S. IIP

BAB I

PENDAHULUAN

1.1 Latar Belakang

Information retrieval (IR) adalah ilmu yang mempelajari pencarian dokumen untuk memenuhi kebutuhan informasi dari dalam koleksi besar media penyimpanan komputer (Manning, Raghavan, & Schütze, 2008). Berkembangnya teknologi informasi membuat IR tidak lagi digunakan oleh beberapa orang yang terlibat dalam pustakawan saja tetapi sekarang jutaan orang ikut terlibat dalam IR setiap harinya ketika mereka menggunakan *web search engine*. Lebih dari 80% pengguna Internet memanfaatkan *search engine* untuk mencari informasi yang mereka butuhkan (Sarwono, 2010).

Universitas Multimedia Nusantara (UMN) memiliki situs sebagai sarana *online presence* sesuai dengan tujuannya menjadi universitas unggulan di bidang teknologi informasi dan komunikasi baik di tingkat nasional maupun internasional (Universitas Multimedia Nusantara, 2010). Dengan tujuan tersebut perlu diterapkan *information retrieval* agar pengunjung situs dapat dengan mudah memperoleh informasi yang mereka cari. Pada Agustus 2011, William Riyanto telah berhasil membangun *search engine* untuk situs UMN (Riyanto, 2011). Penelitiannya telah berhasil mengurutkan hasil pencarian berdasarkan peringkat relevansi dengan menggunakan algoritma *PageRank*. Berdasarkan hasil observasi, *search engine* yang dibangun tidak dilengkapi dengan proses menginterpretasi *query* pengguna layaknya *search engine* mayor, seperti Google, Yahoo, dan AOL,

Web Search Engine..., Stephani Marcelli Djojo, FTI UMN, 2012

sehingga hasil pencarian terbatas.

Dalam mendapat informasi yang diinginkan rata-rata manusia menggunakan suatu *query* mengandung 2,4 kata (Spink & Jansen, 2004). Dengan kata lain, umumnya *query* pencarian yang dimasukkan oleh pengguna terdiri dari 2-3 jumlah kata per *query*. Kondisi yang ada adalah pengguna tidak selalu merumuskan *query* pencarian menggunakan dan menyusun kata yang tepat.

Google dikenal sebagai *search engine* yang baik dalam menginterpretasikan *query*. Google menggunakan algoritma dan perhitungan untuk menginterpretasi *query* (Google Guide, 2009), antara lain metode Boolean, operator *special character*, algoritma *stemming*, dan *stop words* yang menjadi fokus penelitian. Kombinasi *query* dapat dibuat dengan operator Boolean seperti AND dan OR. Ada juga *special character* seperti tanda kutip (“ ”) dan *tilde* (~) digunakan Google sebagai operator menginterpretasikan *query*. Algoritma *stemming* digunakan Google sejak tahun 2003 (Notess, 2003) untuk memungkinkan *search engine* secara otomatis menambahkan bentuk kata similiar pada *query* pencarian agar mengembalikan hasil relevan tambahan (Google, 2007). Dengan *stemming*, Google melakukan pencarian sampai ke *root* kata-kata yang dimasukkan sebagai *query* sehingga jumlah hasil pencarian lebih banyak untuk *query* yang terdiri lebih dari satu kata. Google menginterpretasikan *stemming* pada *query* yang terdiri lebih dari satu kata (Google Guide, 2009). Dalam mengelola *query* diperlukan metode untuk menghapus kata-kata umum yang tidak menambah makna pada pencarian, seperti *di*, *ke*, *dari*, *untuk*. Hal ini dikenal dengan *stop words*.

UNIVERSITAS MULTIMEDIA NUSANTARA

Morfoligi bahasa menjadi permasalahan tersendiri dalam *search engine* karena setiap bahasa memiliki aturan yang berbeda-beda dalam fungsi perubahan-perubahan bentuk kata. *Stemming* Google sudah canggih karena menggunakan *morphological analyzer* multibahasa. Ada beberapa algoritma *stemming* bahasa Indonesia yang dikenal, antara lain algoritma Nazief-Andriani, algoritma Arifin-Setiono, dan algoritma Vega. Dari algoritma *stemming* bahasa Indonesia yang ada, algoritma Nazief-Andriani telah terbukti sebagai algoritma *stemming* yang memiliki keakuratan paling baik (Asian, Williams, & Tahaghoghi, 2004) dan meningkatkan performa dari IR *search engine* (Agusta, 2009).

Mengacu pada konten dalam situs UMN yang secara umum menggunakan bahasa Indonesia, penggunaan algoritma *stemming* Nazief-Andriani cocok untuk diimplementasikan dalam *search engine*. Kondisi saat ini masih terbuka peluang mengembangkan model *web search engine* yang menggunakan algoritma Nazief-Andriani dan operator pencarian dalam rangka meningkatkan relevansi pencarian.

1.2 Identifikasi Masalah

Permasalahan yang dihadapi adalah hasil pencarian *search engine* pada situs UMN terbatas karena tidak menginterpretasikan *query* pencarian.

UNIVERSITAS
MULTIMEDIA
NUSANTARA

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, berikut rumusan masalah.

a. Bagaimana mengimplementasikan algoritma *stemming* Nazief-Andriani, pada sebuah model *web search engine*?
b. Bagaimana mengimplementasikan operator pencarian, yaitu operator Boolean dan operator *special character* pada *query* pencarian?
c. Bagaimana mengimplementasikan *stop words* pada *query* pencarian?
d. Bagaimana menyajikan hasil pencarian?

1.4 Batasan Masalah

Penelitian ini hanya membahas implementasi algoritma *stemming* Nazief-Andriani, operator pencarian (AND, OR, *quote pharse*, dan *tilde*), dan *stop word* guna memperbaiki model *search engine* yang sudah dibangun sebelumnya. *Search engine* yang dibangun untuk pencarian artikel. Lingkup penelitian untuk operator *tilde* hanya sebatas guna *stemming single word*.

Variabel penelitian dibatasi pada dua parameter metode evaluasi Paice, yaitu *understemming index* dan *overstemming index*. Untuk proses pengujian interpretasi *query* digunakan tiga kelas kata, yaitu kata tunggal atau jamak, kata gabung, dan kata kerja dengan banyak akhiran.

Dataset crawling yang digunakan adalah hasil *crawling* pada situs www.umn.ac.id dan tv.umn.ac.id. Pencarian tidak merepresentasikan *ontology*/ ilmu pemaknaan kata.

UNIVERSITAS MULTIMEDIA NUSANTARA

1.5 Tujuan Penelitian

Berdasarkan rumusan masalah, maka penelitian ini bertujuan untuk mengimplementasikan algoritma *stemming* Nazief-Andriani, operator pencarian (AND, OR, *quote pharse*, dan *tilde*), dan *stop words* serta menyajikan hasil pencarian dalam sebuah model *web search engine*.

1.6 Manfaat Penelitian

Adapun penelitian ini memiliki manfaat sebagai berikut.

a. Model *web search engine* dapat dijadikan referensi bagi situs UMN dalam hal peningkatan kinerja *search engine*-nya agar dapat menelusuri informasi yang relevan,
b. Menginterpretasi *query* yang pengguna masukkan sehingga menjembatani kondisi dari pengguna yang tidak selalu merumuskan *query* pencarian menggunakan dan mengkombinasikan kata yang tepat, dan
c. Meningkatkan kinerja model *web search engine* sehingga tidak hanya menyajikan hasil pencarian berdasarkan peringkat (*PageRank*) tetapi dengan adanya preproses *query* pencarian.

UNIVERSITAS
MULTIMEDIA
NUSANTARA