Perancangan Modul Bahasa Indonesia untuk Microsoft FAST ESP Search Engine

Alfian Akbar Gozali
Universitas Telkom
alfian@tass.telkomuniversity.ac.id

Abstrak

Fungsi utama dari search engine atau mesin pencari adalah mempermudah dan mempercepat pencarian. Latar belakang dibuatnya mesin pencari ini adalah banyaknya jumlah dokumen di internet yang mempersulit proses pencarian user.Selain digunakan di internet, mesin pencari juga banyak diterapkan pada perusahaan.Mesin pencari untuk perusahaan sudah ada banyak, salah satunya adalah FAST ESP Search Engine.FAST ESP adalah sebuah software terintegrasi yang menyediakan sebuah platform untuk layanan pencarian dan filtering. Sistem ini dibuat terdistribusi sehingga dapat menyediakan temu balik informasi dari beberapa jenis informasi (dan dataset).FAST ESP secara otomatis mengenali 81 jenis bahasa termasuk Indonesia namun jika ingin mengaplikasikan FAST ESP untuk dokumen berbahasa Indonesia, penambahan modul custom harus dilakukan.

Proses perancangan modul bahasa Indonesia untuk FAST ESP adalah analisis proses Microsoft FAST ESP, analisis strategi awal perancangan modul bahasa Indonesia, perancangan model linguistik bahasa Indonesia, dan studi prioritas proses pada model lingustik bahasa Indonesia. Pada akhirnya environment dan proses dalam FAST ESP sangat menentukan posisi dari modul bahasa Indonesia yang akan dimasukkan. Berdasarkan hasil analisis didapatkan bahwa tahap pipeline adalah satu-satunya tahap yang dapat digunakan. Proses penyisipan modul bahasa Indonesia adalah dengan mengganti atau menambahkan stage pada pipeline. Stage yang akan disisipi atau bahkan diubah dapat dibagi berdasarkan kebergantungannya kepada bahasa.
Kata kunci:FAST ESP, search engine, search enginebahasa Indonesia

I. Pendahuluan

1.1. Latar Belakang

Search engine pertama kali dibuat oleh McBrian dan diberi nama World Wide Web Worm(WWW) [1]. McBrian membuat WWWW dengan tujuan untuk ‘menjinakkan’ world wide web alias internet.Hal ini karena masalah yang muncul setelah internet muncul adalah jumlah website yang demikian banyaknya, sehingga diperlukan alat untuk mengindeks website yang ada di internet untuk mempermudah pencarian.Dari sinilah muncul sebuah rumpun penelitian yang bernama sistem temu balik informasi (information retrieval). WWWW ini jugalah yang menginspirasi Sergey Brin and Lawrence (Larry) Page untuk membuat Google[2], mesin pencari yang saat ini masih menjadi pemimpin di world wide web dengan algoritma Google Page Rank-nya.

Karena kehandalan dalam mememukan sebuah informasi dalam bongkahan data yang sangat besar, kini mesin pencari bukan hanya milik world wide web. Saat ini telah banyak perusahaan yang memasukkan mesin pencari dalam proses bisnis atau sistemnya. Salah satu contohnya adalah Facebook yang menggandeng Microsoft Bing untuk pencarian non-graf [3] walaupun pada akhirnya Facebook mengembangkan mesin pencarinya sendiri [4]. Contoh yang lain adalah open source search engine Solr, yang mempunyai core dari Apache Lucene, mempunyai puluhan pengguna yang terdiri dari public website hingga private company [5]. Bahkan perusahaan sebesar Microsoft yang mempunyai search engine, Bing [6]pun pada bulan April 2008 melakukan akuisisi pada FAST ESP search engine [7]. FAST ESP search engine adalah sebuah mesin pencari yang dikhususkan untuk enterprise (perusahaan).

FAST ESP search engine merupakan search engine yang cukup lengkap dan tinggi skalabilitasnya. Namun FAST ESP mempunyai kekurangan jika ingin diterapkan di perusahaan Indonesia karena hingga saat ini, FAST ESP belum mendukung bahasa Indonesiawalaupun sudah dapat mendeteksi adanya kata dalam bahasa Indonesia [8]. Oleh karena itu diperlukan tambahan modul khusus jika ingin menerapkan FAST ESP search engine ini untuk dokumen berbahasa Indonesia.

1.2. Rumusan Masalah

Dari latar belakang tersebut dapat ditarik beberapa permasalahan yaitu tidak adanya dukungan FAST ESP Engine untuk bahasa Indonesia. FAST ESP dapat mendeteksi adanya kata dalam bahasa Indonesia namun tidak mendukung untuk proses lingustik selanjutnya. Oleh karena itu permasalahan selanjutnya adalah bagaimana merancang modul bahasa Indonesia untuk FAST ESP. Dalam rangka perancangan ini pun harus memperhatikan proses pengindeksan yang terjadi pada FAST ESP. Dengan demikian, masalah terakhir adalah apa saja proses yang harus ada dalam rangka membangun modul bahasa Indonesia pada FAST ESP. Hal ini juga mencakupskala prioritas dari proses-proses tersebut.

1.3. Tujuan

Tujuan dari penelitian ini diturunkan dari rumusan permasalahan sebelumnya.Adapun tujuan utama dari penelitian ini adalah melakukan perancangan modul bahasa Indonesia pada FAST ESP search engine. Dari tujuan utama tersebut dapat di-break down menjadi beberapa sub-tujuan, yaitu:

- Menganalisis environmentdan proses dalam FAST ESP,
- Membuat strategi awal dalam melakukan perancangan modul bahasa Indonesia,
- Melakukan studi prioritas untuk sub-modul bahasa Indonesia, dan
- Melakukan perancangan modul bahasa Indonesia untuk FAST ESP

II. Tinjauan Pustaka

2.1. Search Engine

Fungsi utama dari search engine atau mesin pencari adalah mempermudah dan mempercepat pencarian. Mesin pencari pertama kali dibuat oleh McBrian dengan nama World Wide Web Worm (WWWW) pada tahun 1993[1].Latar belakang dibuatnya WWWW ini adalah banyaknya jumlah website di internet yang mempersulit proses pencarian user.Mesin pencari ini mempunyai database berupa 300.000 objek multimedia. Mesin pencari pertama ini bahkan support Perl regex, sesuatu yang tidak ada di mesin pencari saat ini.

Setelah kemunculan WWWW, dua tahun selanjutnya muncullah Google oleh Lawrence “Larry” Page dan Sergey Brin[2]. Walaupun Google terinspirasi dari WWWW, algoritma pencarian yang digunakan Google berbeda. Mereka menggunakan Page Rank [9] sebagai core utama mesin pencari tersebut. Berbeda dengan pendahulunya, Google menggunakan algoritma yang dinamis dan berorientasi kepada pengguna. Google Page Rank menggunakan feedback dari pengguna untuk memperbaiki hasil pencariannya, terutama urutan/ranking dari hasil pencarian.

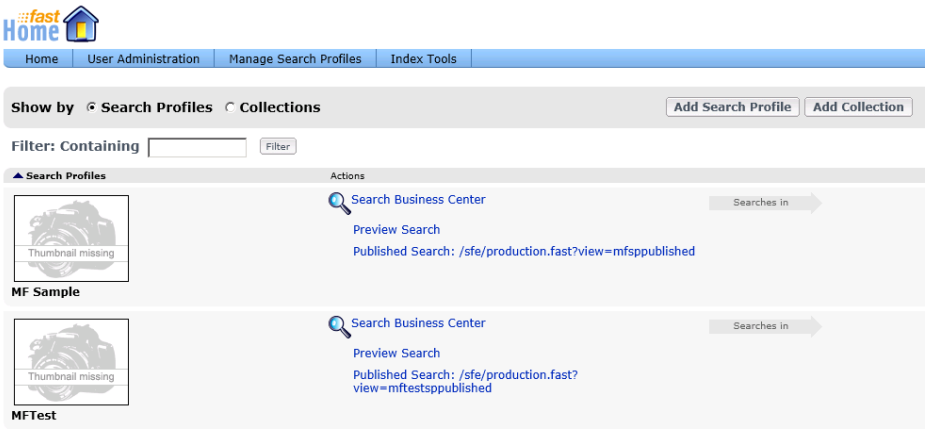
Selain digunakan di internet, mesin pencari juga banyak diterapkan pada perusahaan. Menurut Hawking [10], mesin pencari pada perusahaan (enterprise search engine) termasuk:

- setiap organisasi dengan konten teks dalam bentuk elektronik;
- pencarian pada website eksternal organisasi;
- pencarian pada website internal (intranet) organisasi; dan
- pencarian pada teks elektronik lain yang dimiliki oleh organisasi dalam bentuk email, record basis data, dokumen yang di-share, dan sebagainya.

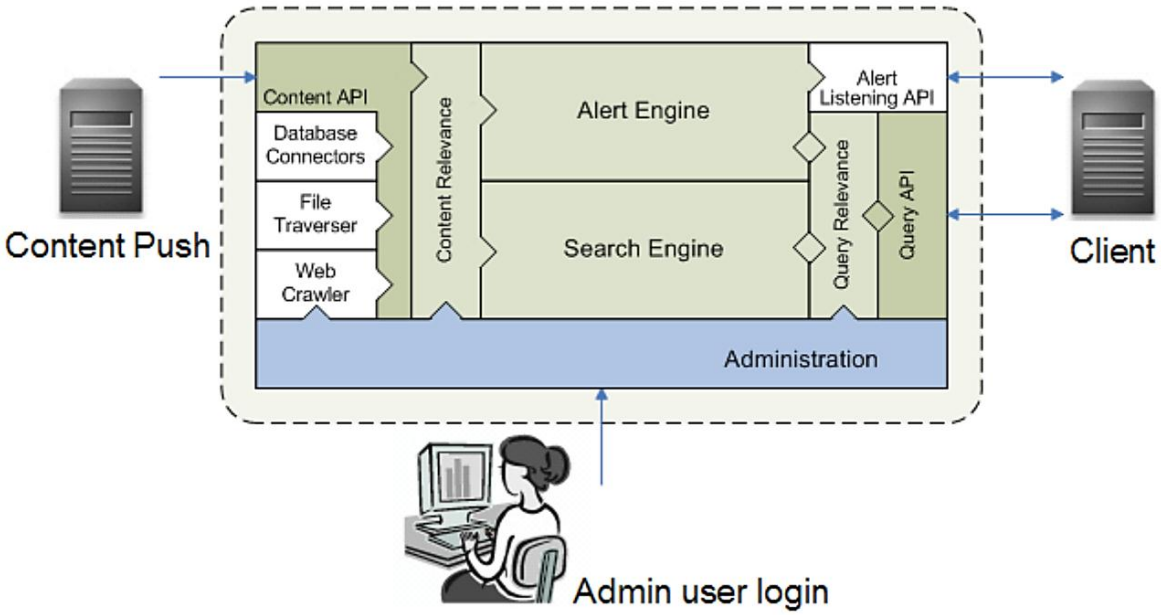
Mesin pencari untuk perusahaan sudah ada banyak, dari yang open source seperti Apache Solr [5] hingga propetiary sepertiFAST ESP Search Engine yang sekarang telah menjadi satu dengan Microsoft Share Point [11]. Daftar tools pencarian untuk perusahaan sendiri dapat dilihat di Wikipedia [12]. Sedangkan pada paper ini akan dibahas mesin pencari dari Microsoft yang bernama FAST ESP Search Engine.

2.2. Microsoft FAST ESP Search Engine

FAST ESP adalah sebuah software terintegrasi yang menyediakan sebuah platform untuk layanan pencarian dan filtering.Sistem ini dibuat terdistribusi sehingga dapat menyediakan temu balik informasi dari beberapa jenis informasi (dan dataset). ESP mengkombinasikan pencarian secara real-time, linguistik lanjut, dan pilihan akses konten yang bervariasi pada produk yang modular dan scalable[13].Gambar 1 menunjukkan tampilan halaman home FAST ESP sedangkanarsitektur sistem dari FAST ESP ditunjukkan pada Gambar 2.



Gambar 1. Tampilan halaman home FAST ESP



Gambar 2. Arsitektur Sistem FAST ESP [13]

FAST ESP secara otomatis mengenali 81 jenis bahasa dengan encoding umum seperti ISO 639. Bahasa yang dapat dideteksioleh FAST ESP ditunjukkan pada Gambar 3. Namun bahasa yang didukung secara penuh oleh FAST ESP secara default hanya ada 9: Dutch, Inggris, Perancis, Jerman, Itali, Korea, Norwegia, Portugis, dan Spanyol.Oleh karena itu jika ingin mengaplikasikan FAST ESP untuk dokumen berbahasa Indonesia, penambahan modul custom harus dilakukan.

Language	ISO 639 Code	Language	ISO 639 Code
Afrikaans	af	Japanese	ja
Albanian	sq	Kazakh	kk
Arabic	ar	Kirghiz	ky
Armenian	hy	Korean	ko
Azeri	az	Kurdish	ku
Bangla	bn	Latin	la
Basque	eu	Latvian	lv
Bosnian	bs	Letzeburgish	lb
Breton	br	Lithuanian	lt
Bulgarian	bg	Macedonian	mk
Byelorussian	be	Malay	ms
Catalan	ca	Maltese	mt
Chinese (simplified)	zh-simplified	Maori	mi
Chinese (traditional)	zh-traditional	Mongolian	mn
Croatian	hr	Norwegian (Bokmaal)	nb
Czech	cs	Norwegian (Nynorsk)	nn
Danish	da	Pashto	ps
Dutch	nl	Polish	pl
English	en	Portuguese	pt
Esperanto	eo	Rhaeto-Romance	rm
Estonian	et	Romanian	ro

Language	ISO 639 Code	Language	ISO 639 Code
Faeroese	fo	Russian	ru
Farsi	fa	Sami (Northern)	se
Filipino	tl	Serbian	sr
Finnish	fi	Slovak	sk
French	fr	Slovenian	sl
Frisian	fy	Sorbian	wen[1]
Galician	gl	Spanish	es
Georgian	ka	Swahili	sw
German	de	Swedish	sv
Greek	el	Tamil	ta
Greenlandic	kl	Thai	th
Hausa	ha	Turkish	tr
Hebrew	he	Ukrainian	uk
Hindi	hi	Urdu	ur
Hungarian	hu	Uzbek	uz
Icelandic	is	Vietnamese	vi
Indonesian	id	Welsh	cy
Irish (Gaelic)	ga	Yiddish	yi
Italian	it	Zulu	zu

Gambar 3. Bahasa yang dapat dideteksi secara otomatis oleh FAST ESP [8]

2.3. Bahasa Indonesia

Pengembangan modul custom bahasa Indonesia tidak lepas dari prinsip linguistik bahasa Indonesia sendiri. Oleh karena itu sebelum mengembangkan modul bahasa Indonesia, harus dipahami akar rumpun ilmu bahasa Indonesia. Dengan kita mempelajari akar rumpun bahasa Indonesia, diharapkan proses perancangan modul bahasa Indonesia menjadi lebih terstruktur dan mempunyai akurasi yang tinggi.

Proses perancangan model linguistik bahasa Indonesia harus memperhatikan berbagai konsep, seperti gaya bahasa (majas), tata bahasa (gramatikal), kosa kata, dan sebagainya. Menurut Keraf [14], “gaya bahasa dapat dibatasi sebagai cara mengungkapkan pikiran melalui bahasa secara khas yang memperlihatkan jiwa dan kepribadian penulis (pemakai bahasa)”. Gaya bahasa adalah konsepyang paling dinamis dalam pembangunan model linguistik. Sedangkan tata bahasa adalahilmu yang mempelajari kaidah-kaidah yang mengatur penggunaan bahasa. Ilmu ini merupakan bagian dari

bidang ilmu yang mempelajari bahasa yaitu linguistik. Gambaran umum dari tata bahasa antara lain [15]:

- Pembentukan kata dilihat dari afikasi (pengimbuhan) dan reduplikasi (pengulangan)
- Sarana-sarana dari tingkat leksikal mau pun di tingkat gramatikal dapat digunakan untuk menyatakan arti
- Satuan sintaksis bersifat senyawa
- Jalinan tingkat gramatikal dan leksikal yang perlu diperhatikan

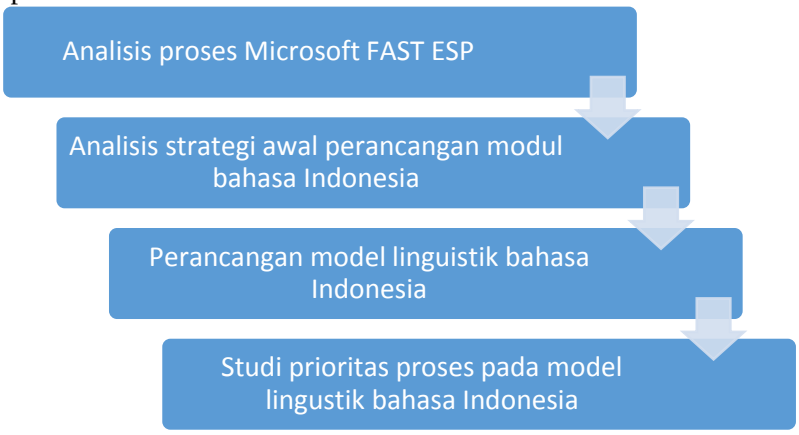
Tata bahasa memilili tingkat struktur gramatikal dari bawah ke atas [15]:

- Ciri-ciri morfem: meliputi morfem akar, morfem bebas dan terikat, dan penjenisan morfem secara gramatikal
- Ciri-ciri kata:meliputi kata penuh dan kata tugas, kata akar, dan kata turunan
- Hubungan kata dalam setiap segmentasi kalimat
- Tipe-tipe struktur dalam segmentasi kalimat

Selain gaya dan tata bahasa, kosa kata juga memegang peranan penting.Menurut kamus besar bahasa Indonesia edisi 2013, bahasa Indonesia memiliki sekitar 92.000 kata [16]. Jumlah kosa kata dalam bahasa Indonesiatersebutmasih belum termasuk kata populer dan tidak baku.

Metode Penelitian

Dalam melakukan perancangan modul custom bahasa Indonesia, dibutuhkan metode yang tepat agar didapatkan akurasi yang optimal. Karena penelitian ini berfokus pada proses perancangan maka output dari penelitian ini adalah diagram perancangan seperti blok diagram dan laporan usul perancangan. Gambar 4menujukkanmetode yang akan dilakukan dalam rangka merancang modul bahasa Indonesia pada FAST ESP:



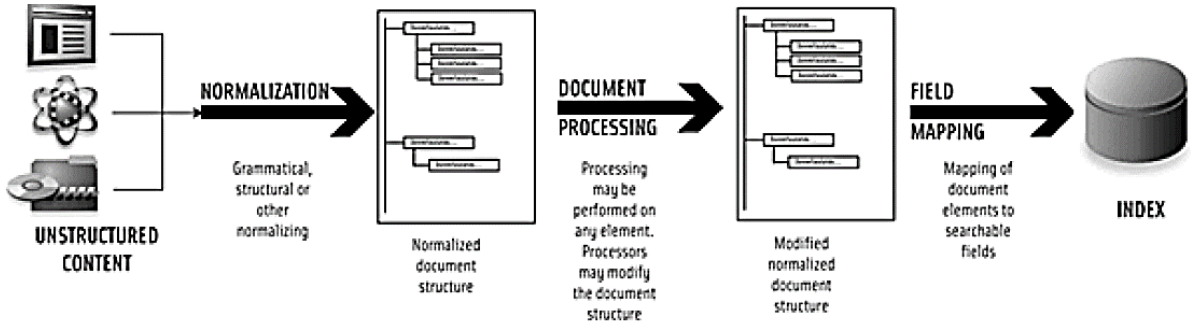
Gambar 4. Metode penelitian perancangan modul bahasa Indonesia untuk FAST ESP

2.4. Analisis proses Microsoft FAST ESP

Tahap pertama adalah analisa proses yang terjadi pada FAST ESP, terutama pada fase pengindeksan. Dengan mempelajari alur proses dan perubahan data yang mengalir, dapat dilakukan analisis pendekatan proses-proses mana saja yang ada dalam FAST ESP yang harus diganti dengan proses custom modul bahasa Indonesia.

Konten dan aliran proses

Gambar 5 menunjukkan konten dan aliran proses yang terjadi pada proses pengindeksan dokumen pada FAST ESP. Pada FAST ESP, data yang belum dimasukkan ke FAST ESP disebut content (konten) sedangkan konten yang sudah dapat dicari (searchable) disebut document (dokumen). Contoh dari konten adalah file Microsoft Word, halaman HTML, atau entri basis data.



Gambar 5. Konten dan aliran proses indexing FAST ESP[13]

Perlu dicatat bahwa konten yang telah masuk dan mengalir dalam FAST ESP mengalami perlakukan dan tahap yang berbeda, meliputi normalisasi, pemrosesan dokumen, dan pengindeksan sebelum konten tersebut dapat dicari. Pada FAST ESP, tahap tersebut dapat kita spesifikasikan sendiri atau memakai default dari FAST ESP.

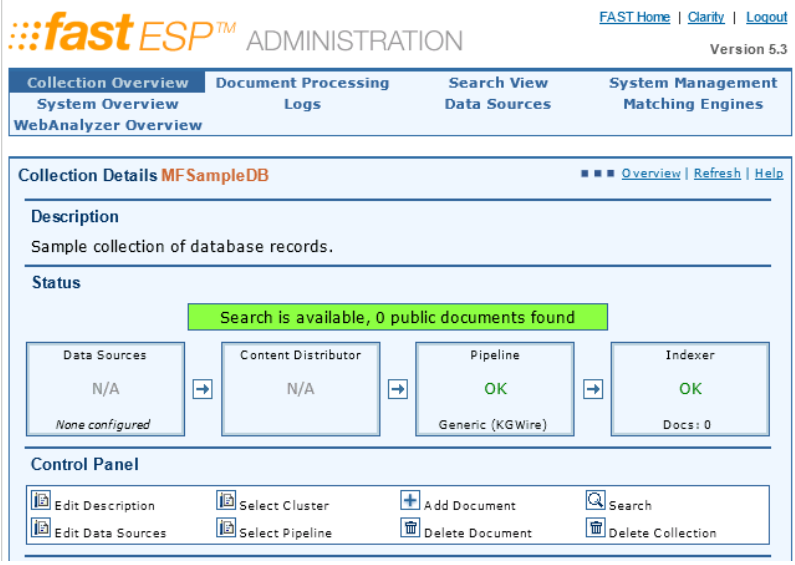
Collections

Konten yang telah masuk ke FAST ESP akan diproses, dibuat agar bisa dicari, dan dikelompokkan ke dalam collection pada ESP. Collection juga memungkinkan untuk membagi konten menjadi kelompok (group) konten yang berbeda. Masing-masing collection dispesifikasikan cara dokumen tersebut diproses dan diindeks.

Pengelompokan konten ke dalam collection biasanya berdasarkan kriteria sebagai berikut:

- Perbedaan cara pandang konten dilihat dari aplikasi end-user sepertiproduk data, halaman website, dan berita
- Kepemilikan konten seperti konten intranet atau extranet
- Aturan pemrosesan khusus seperti penanganan metadata

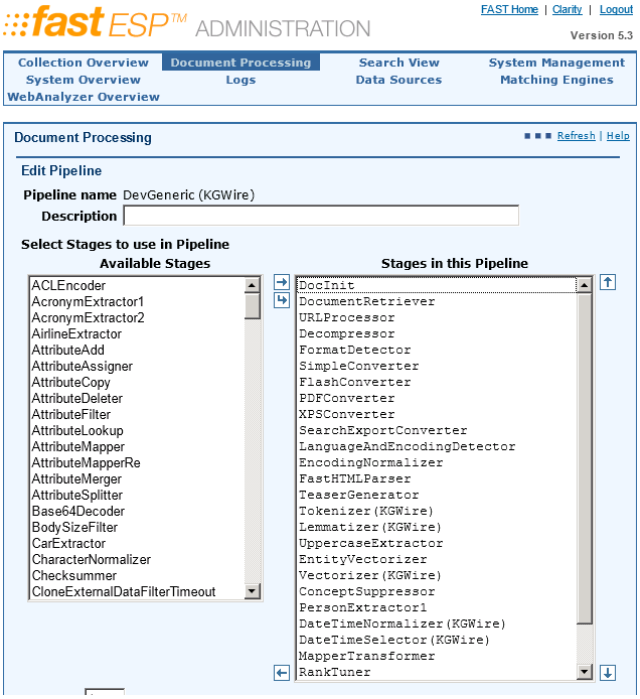
Pengelompokan konten memungkinkan end-user atau aplikasi query eksternal untuk dapat mengkerucutkan cakupan pencarian ke suatu tipe dokumen. Sebagai tambahan, collection memungkinkan juga untuk mengkhususkan urutan dari tipe konten yang akan diproses selama pemrosesan dokumen dengan memprioritaskan collection individu.Gambar 6 adalah contoh tampilan manajemen collection pada FAST ESP. Collection yang dikonfigurasi bernama MF SampleDB.



Gambar 6. Tampilan konfigurasi collection FAST ESP

Berdasarkan arsitektur dan proses dari FAST ESP tersebut, dapat dilakukan analisis kemungkinan pos untuk meletakkan modul bahasa Indonesia. Dari segi konten dan collection sudah tidak bisa disisipi modul tambahan.Kesempatan satu-satunya untuk menyisipkan modul bahasa

Indonesia hanya dapat dilakukan pada proses FAST ESP, yaitu pada pipeline (lihat Gambar 6). Jika kita masuk lebih dalam lagi pada konfigurasi pipeline maka akan didapatkan tampilan seperti padaGambar 7.



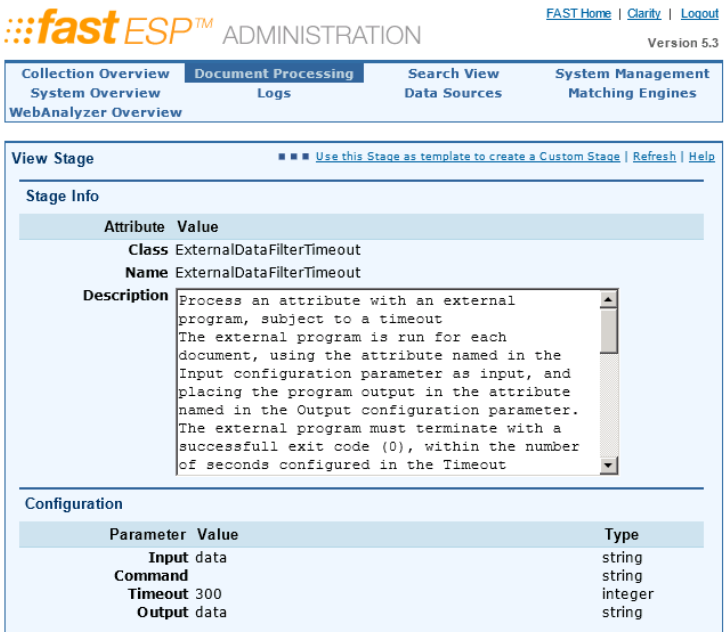
Gambar 7. Tampilan konfigurasi pipeline FAST ESP

Pada Gambar 7 dapat dilihat bahwa pipeline terdiri dari beberapa stage yang berurutan (list sebelah kanan). Stage adalah sub-proses yang akan dilakukan saat sebuah konten masuk ke FAST ESP untuk diindeks. Oleh karena itu, penyisipan modul bahasa Indonesia akan dilakukan di bagian ini.

2.5. Analisis strategi awal perancangan modul bahasa Indonesia

Setelah dilakukan analisis pada environment dan proses pada FAST ESP, perlu dibuat strategi awal dalam rangka perancangan modul bahasa Indoensia. Pada fase sebelumnya telah didapatkan kesimpulan bahwa hanya pada pipeline yang dapat dilakukan penyisipan modul bahasa Indonesia. Pertanyaan pada tahap ini adalah: “Bagaimana cara melakukan penyisipan pada FAST ESP?”

Pada FAST ESP, penyisipan atau penambahan modul (stage) dapat dilakukan dengan dua cara: (1) membuat stage custom dari python dan (2) menggunakan stage bernama *externaldatafiltertimeout* dengan menyediakan engine eksternal sebagai pengolah data.Gambar 8 menunjukkan stage *externaldatafiltertimeout.*Stage tersebut menggunakan command script (seperti pada windows command prompt atau linux terminal) untuk melakukan invoking modul eksternal.



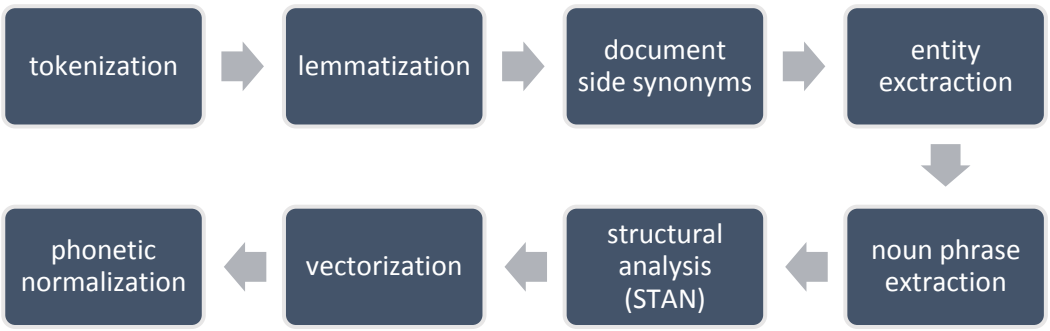
Gambar 8. Tampilan externaldatafiltertimeout stage

2.6. Perancangan model linguistik bahasa Indonesia

Setelah cara menyisipkan atau menambahkan modul bahasa Indonesia telah ditemukan maka langkah selanjutnya adalah menentukan stage apa sajakah yang akan diganti dengan external stage. Atau dengan kata lain, proses linguistik apa sajakah yang akan dibuat versi bahasa Indonesia-nya? Pertanyaan ini dapat dijawab dengan tiga langkah, yaitu: (1) menganalisa modul linguistik FAST ESP, (2) menentukan proses mana saja pada modul linguistik yang bergantung bahasa dan mana yang tidak bergantung bahasa, dan (3) analisis satu per satu proses pada modul linguistik tersebut.

Modul Linguistik FAST ESP

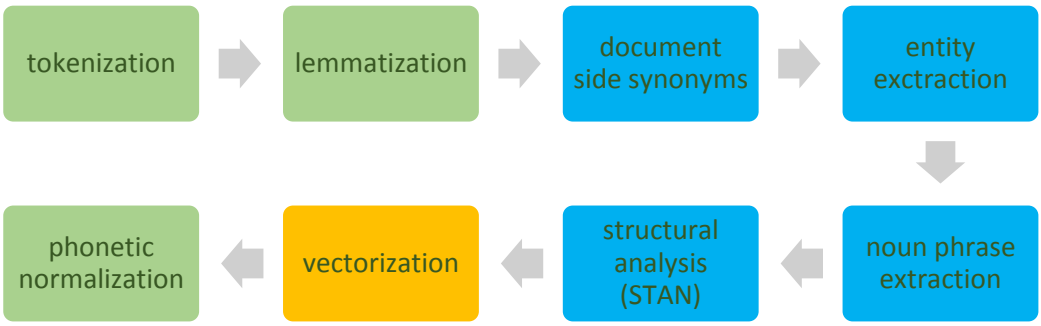
Sebelum menentukan proses yang akan dibuat model linguistik bahasa Indonesia-nya, perlu dianalisis terlebih dahulu gambaran umum dari model linguistik FAST ESP. FAST ERP Linguistic Modul memiliki gambaran umum sistem sebagaimana pada Gambar 9.



Gambar 9. Gambaran umum model linguistik FAST ESP

2.7. Ketergantungan Modul Linguistik pada Bahasa

Semua proses pada modul linguistil Gambar 9 harus dilakukan analisis terhadap kebergantungan kepada bahasa. Hal ini dilakukan untuk mengetahui proses mana saja yang harus dibuat modul bahasa Indonesia dan mana yang dapat dibiarkan saja. Oleh karena itu Gambar 9 harus diidentifikasi ulang menjadi diagram yang lebih menunjukkan kebergantungan pada bahasa. Hasilnya ditunjukkan pada



Gambar 10. Gambaran umum dan kebergantungan pada bahasa model linguistik FAST ESP

Keterangan:

Bagian yang berwarna hijau adalah proses yang menurut manual FAST ESP sangat sensitif terhadap jenis bahasa yang digunakan. Sebenarnya ada satu proses lagi yang sensitif terhadap jenis bahasa yaitu spell checking.

Sedangkan bagian yang berwarna biru adalah proses yang sangat bergantung kepada dictionary yang ada. Namun khusus untuk structural analysis (STAN), proses ini menggunakan format dokumen, ukuran font, lokasi kata, dan beberapa kriteria lain untuk mengenali sebuah artikel.

Adapun satu proses yang berwarna jingga dapat dikatakan independen terhadap perubahan jenis bahasa dikarenakan proses ini hanya merupakan proses penghitungan nilai kemiripan (similarity) antar dokumen.

2.8. Analisis Modul Bahasa Indonesia

Adapun analisis modul bahasa Indonesia untuk Gambar 10 di atas adalah sebagai beikut:

Tokenization

Proses tokenization pada FAST ESP merupakan proses yang bergantung kepada jenis bahasa. Namun demikian karena alfabet yang digunakan pada bahasa Indonesia sama dengan bahasa Inggris, secara umum teknik tokenisasi bahasa Indonesia dan bahasa Inggris sama. Oleh karena itu pembuatan modul tidak menanganani tokenisasi secara khusus.

Lemmatization

Proses lemmatization adalah proses mencari kata dasar dari sebuah kata bentukan. Oleh karena itu proses ini sangat sensitif terhadap bahasa. Dikarenakan proses ini adalah salah satu proses kritikal dan terletak pada hulu sistem linguistik pada FAST ESP maka pengembangan akan dimulai dari proses ini.

Document Side Synonyms

Proses document side synonims adalah proses look-up pada kamus sinonim yang sudah tersedia. Kamus sinonim pada FAST ESP tidak mendukung bahasa Indonesia. Oleh karena itu pembangunan kamus sinonim juga merupakan fokus dari pembangunan modul linguistik ini.

Entity Extraction

Proses entity extraction merupakan proses ekstraksi entitas pada artikel seperti nama orang, jalan, kota, waktu, dan lain sebagainya. Proses entity extraction merupakan salah satu fitur pada FAST ESP yang sangat powerful. Fitur ini juga merupakan fitur andalan dari FAST ESP. Ekstraksi entitas untuk bahasa Indonesia tidak didukung oleh FAST ESP. Dikarenakan pentingnya fitur entity extraction maka pengembangan modul juga difokuskan untuk fitur ini.

Noun Phrase Extraction

Proses noun phrase extraction merupakan proses ekstraksi frasa pada kalimat. Proses ini adalah proses yang sangat tergantung pada kamus. Modul pengembangan FAST ESP juga mencakup pengembangan kamus frasa bahasa Indonesia.

Structural Analysis (STAN)

Proses structural analysis (STAN) hampir sama seperti proses ekstraksi entitas namun bedanya proses STAN menggunakan format dokumen, ukuran font, lokasi kata, dan beberapa kriteria lain untuk mengenali sebuah artikel.

Proses ini dapat dikatakan tidak terlalu vital pada pengembangan modul FAST ESP bahasa Indonesia dikarenakan format dokumen perusahaan umumnya homogen, yaitu artikel berita atau laporan. Oleh karena itu proses STAN akan menjadi suplemen di akhir perancangan.

Vectorization

Sebagaimana telah dijelaskan sebelumnya, Proses vectorization pada FAST ESP merupakan dapat dikatakan independen terhadap perubahan jenis bahasa dikarenakan proses ini hanya merupakan proses penghitungan nilai kemiripan (similarity) antar dokumen. Oleh karena itu, vectorization bukan merupakan bagian dari pengembangan modul ini.

Phonetic Normalization

Proses phonetic normalization adalah proses normalisasiterhadap kata yang dianggap ‘tidak normal’. Biasanya kata tersebutadalah kata yang tidak formal dan tidak ada dalam kamus besar bahasa Indonesia (KBBI).

Pada pertemuan terakhir telah dinyatakan bahwa by assumption artikel pada perusahaan secara umum telah melalui proses editing yang ketat. Oleh karena itu dapat dipastikan bahwa dokumen yang dimiliki perusahaan telah mengikuti kaidah ejaan yang disempurnakan (EYD) dan menggunakan kata bakubahasa Indonesia.

Berdasarkan hal tersebut, pengembangan modul bahasa Indonesia tidak terlalu fokus pada normalisasi.

Proses Tambahan

Dikarenakan proses pengembangan modul bahasa Indonesia tidak lepas dari pembangunan skema dan model linguistik maka dibutuhkan beberapa proses tambahan yang vital untuk dapat menghasilkan akurasi tinggi. Proses tersebut antara lain adalah proses ekstraksi Part of Speech (PoS) atau biasa disebut dengan PoS Tagging. Proses ini akan mengekstrak jenis dari sebuah kata misalnya kata benda, kata sifat, maupun kata kerja.

Studi prioritas proses pada model lingustik bahasa Indonesia

Setelah dilakukan analisis terhadap modul bahasa Indonesia, perlu dilakukan studi prioritas pada tiap proses dalam model linguistik. Hasil dari tahap ini adalah skala prioritas masing-masing proses. Dengan mempertimbangkan analisis sebelumnya, maka dapat dibuat skala prioritas seperti pada Tabel 1. Skala Prioritas Proses dalam Model Linguistik FAST:

Tabel 1. Skala Prioritas Proses dalam Model Linguistik FAST

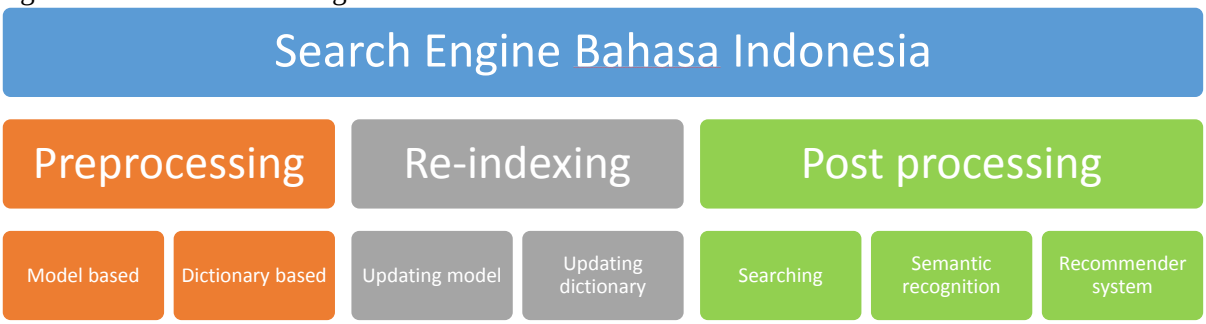
Prioritas	**Proses**
Penting	PoS Tagging
	Lemmatization
	Document Synonyms
	Entity extraction
	Noun phrase extraction
Sedang	STAN
	Phonetic normalization
Suplemen	Tokenization
	Vectorization

III. Pembahasan

Pada bab ini akan dibahas mengenai perancangan modul bahasa Indonesia pada FAST ESP berdasarkan analisis pada bab sebelumnya. Perancangan modul bahasa Indonesia mencakup: metode yang digunakan dan kebutuhan sistem.

Metode yang Digunakan

Gambaran umum sistem pembangunan serch engine bahasa Indonesia untuk FAST ESP dapat digambarkan dalam blok diagram Gambar 11:



Gambar 11. Gambaran Umum Perancangan Modul Bahasa Indoensia

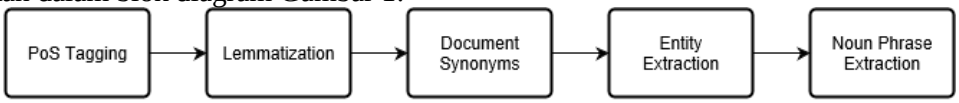
Metode yang dapat digunakan dalam melakukan pengembangan modul bahasa Indonesia adalah metode machine learning berbasis statistik stokastik. Secara spesifik, metode yang dapat digunakan adalah Hidden Markov Method (HMM), sebuah algoritma machine learning yang dapat mengenai pola untuk memprediksi atau merekognisi sesuatu. HMM yang akan digunakan berjalan secara supervised atau membutuhkan data latih.

Model HMM yang dapat digunakan adalah metode yang sama pada paper berjudul Analysis of Hidden Markov Model Method Implementation in Documents Topic Sentence Extraction for Information Retrieval [17]. Untuk menambah efisiensi dan akurasi model juga akan diterapkan algoritma Baum-Welch.

Untuk proses lemmatization, dapat menggunakan teknik awal seperti yang digunakan oleh Suhartono, dkk.[18] dan stemmer dari Apache Lucene [19]. Engine ini hanya digunakan untuk awal pengenalan lemma Indonesia. Setelah diukur akurasinya, akan dikembangkan sendiri algoritma lemmatisasi yang akan menggabungkan teknik stemming, lemmatisasi deterministik, dan pendekatan statistik.

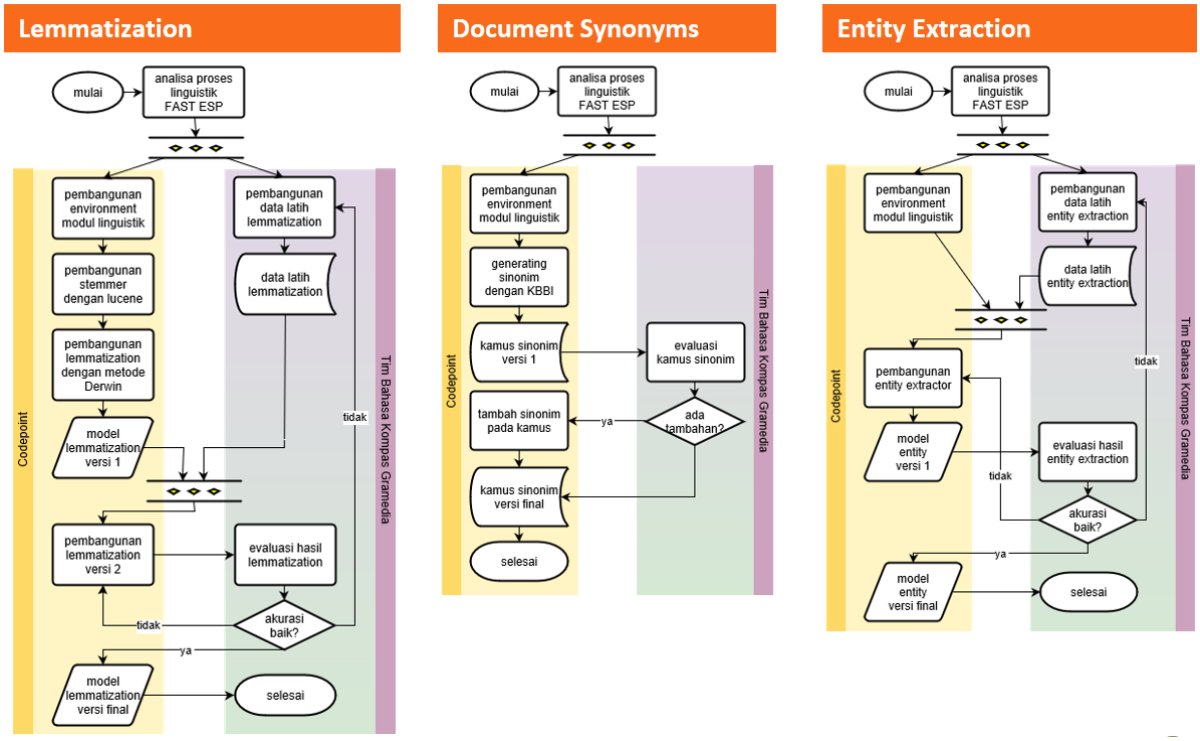
Metode Riset

Gambaran umum pembangunan search engine bahasa Indonesia untuk FAST ESP dapat digambarkan dalam blok diagram Gambar 1:

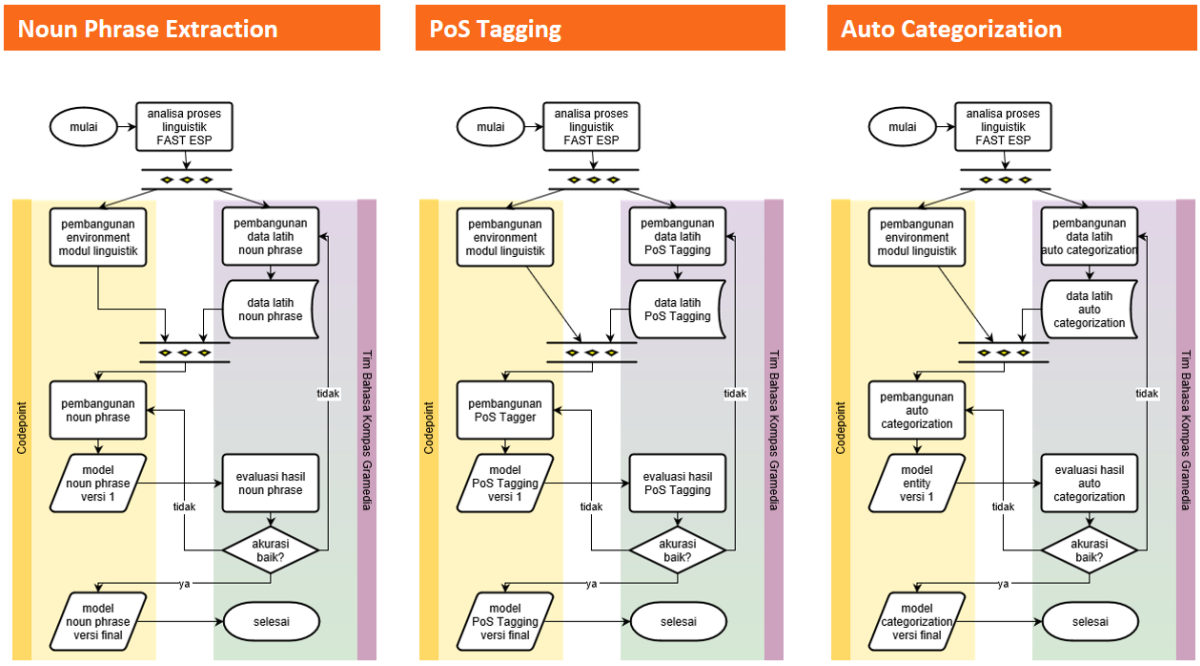


Gambar 12. Blok Diagram Metode Riset

Penjelasan lebih detail blok diagram ada padaGambar 13 dan Gambar 14:



Gambar 13. Blok Diagram Lemmatization, Document Synonyms, dan Entity Extraction



Gambar 14. Blok Diagram Noun Phrase Extraction, PoS Tagging, dan Auto Categorization

Kebutuhan Sistem

3.1. Kebutuhan Hardware dan Dataset

Kebutuhan hardware dan data untuk membangun modul ini:

Super Komputer

- Digunakan untuk melakukan training model dan generating dataset.

- Memory: min 16 GB RAM
- Processor: min Quadcore CPUs
- Hardisk: min 500 GB

Tim Ahli Bahasa

Dibutuhkan untuk membuat data latih machine learning dan evaluasi sistem.

Data Latih

Berupa artikel berita dari perusahaan. Untuk data latih awal dapat dimulai dari 500 artikel yang dipilih oleh ahli bahasa.

3.2. Kebutuhan Software

Kebutuhan software untuk membangun modul ini:

Java Development Kit terbaru

Digunakan dalam pembuatan engine linguistik. Hal ini karena kebanyakan engine third party yang akan dipakai memakai Java.

Full Access ke FAST ESP

Termasuk akses ke direktori (FTP) dimana FAST ESP ditanam (diinstall).

Third Party Engine Linguistik

Untuk mengefisienkan waktu dalam pembangunan modul linguistik bahasa Indonesia maka diperlukan third party engine sebagai berikut:

- Alfian's HMM Linguistic Engine
- Apache Lucene Library
- Stanford Linguistic Library
- Derwin's Indonesia Stemmer Model
- Softcopy KBBI terbaru

IV. Kesimpulan

Kesimpulan yang dapat diambil dari hasil perancangan modul bahasa Indonesia pada FAST ESP adalah environment dan proses dalam FAST ESP sangat menentukan posisi dari modul bahasa Indonesia yang akan dimasukkan. Berdasarkan hasil analisis didapatkan bahwa tahap pipeline adalah satu-satunya tahap yang dapat digunakan. Proses penyisipan modul bahasa Indonesia adalah dengan mengganti atau menambahkan stage pada pipeline. Stage tersebut dapat berupa script python maupun stage external data filter timeout.Stage yang akan disisipi atau bahkan diubah dapat dibagi menjadi 3 prioritas berdasarkan kebergantungannya kepada bahasa: Penting (PoS tagging, lemmatization, document synonyms, entity extraction, dan noun phrase extraction), Sedang (STAN dan phonetic normalization), dan Suplemen (tokenization danvectorization). Kebutuhan sistem seperti hardware, dataset, dan software juga perlu diperhatikan.

Daftar Pustaka

[1] O. A. McBryan, "GENVL and WWWW: Tools for Taming the Web," dalam *First International Conference on the World Wide Web*, Geneva, 1994.

[2] S. Brin dan L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Computer Networks*, vol. 58, no. 18, p. 3825–3833, 2012.

[3] C. Taylor, "If It's Not In Graph Search, Facebook Hands Your Query Off To Bing," TechCrunch, 15 Januari 2013. [Online]. Available: http://techcrunch.com/2013/01/15/if-its-not-in-graph-search-facebook-hands-your-query-off-to-bing/. [Diakses 20 November 2015].

[4] J. Crook, "Facebook Dumps Bing, Will Introduce Its Own Search Tool," TechCrunch, 13 Desember 2014. [Online]. Available: http://techcrunch.com/2014/12/13/facebook-dumps-bing-will-introduce-its-own-search-tool/. [Diakses 20 November 2015].

[5] C. Marquardt, "Public Websites using Solr," Apache Solr, 16 September 2015. [Online]. Available: https://wiki.apache.org/solr/PublicServers. [Diakses 20 November 2015].

[6] MIcrosoft, "Microsoft Bing," Microsoft, 20 November 2015. [Online]. Available: http://www.bing.com/. [Diakses 20 November 2015].

[7] "Microsoft Completes FAST Purchase," IDG News Service, PCWorld, 24 April 2008.

[Online]. Available: http://www.pcworld.com/article/145117/article.html. [Diakses 20 November 2015].
[8] F. ESP, "FAST Enterprise Search Platform - Advanced Linguistics Guide," Fast ESP, Needham, 2008.
[9] L. Page, S. Brin, R. Motwani dan T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford InfoLab, California, 1999.
[10] D. Hawking, "Challenges in enterprise search," dalam *Proceedings of the 15th Australasian database conference*, Darlinghurst, 2004.
[11] Microsoft, "SharePoint - Perangkat Lunak Alat Kolaborasi Tim," Microsoft, 2015. [Online]. Available: https://products.office.com/id-ID/sharepoint. [Diakses 21 November 2015].
[12] anonymous, "List of enterprise search vendors," Wikipedia, 17 November 2015. [Online]. Available: https://en.wikipedia.org/wiki/List_of_enterprise_search_vendors. [Diakses 21 Desember 2015].
[13] F. ESP, "FAST Enterprise Search Platform - Product Overview Guide," 3 April 2008. [Online]. Available: http://download.microsoft.com/download/1/4/8/1483939B-15B8-4DD3-B06D-204D03EC8A1E/Fast_ESP_Prod_Guide.pdf. [Diakses 21 November 2015].
[14] G. Keraf, Diksi dan gaya bahasa : komposisi lanjutan I, Jakarta: Gramedia, 2004.
[15] N. Alieva, Bahasa Indonesia Deskripsi dan Teori, Yogyakarta: Kanisius, 1991.
[16] D. P. Nasional, Kamus Besar Bahasa Indonesia, Jakarta: Balai Pustaka, 2013.
[17] A. A. Gozali dan I. Atastina, "Analysis of Hidden Markov Model Method Implementation in Documents Topic Sentence Extraction for Information Retrieval," dalam *International Conference of Telecomunication (ICTel)*, Bandung, 2010.
[18] D. Suhartono, D. Christiandy dan Rolando, "Lemmatization Technique in Bahasa: Indonesian Language," *Journal of Software*, vol. 9, no. 5, pp. 1202-1209, 2014.
[19] T. A. S. Foundation, "Apache Lucene - Apache Lucene Core," The Apache Software Foundation, 2011. [Online]. Available: https://lucene.apache.org/core/. [Diakses 21 November 2015].