IMPLEMENTASI ALGORITMA COSINE SIMILARITY UNTUK MENDETEKSI KEMIRIPAN TOPIK JUDUL

Heri Sutikno[1], Saniati[2]

Program Studi S1 Informatika
Fakultas Teknik Dan Ilmu Komputer, Unviersitas Teknokrat Indonesia

Abstrak - Perkembangan judul penelitian tugas akhir saat ini menyediakan cara yang mudah untuk mencari literatur akademis secara luas. Seseorang dapat mencari di seluruh bidang ilmu dan referensi. Dalam menentukan diterima atau tidaknya sebuah judul skripsi maka yang dilakukan adalah dengan membandingkan judul tersebut judul yang lain secara manual. Tentu saja hal ini akan membutuhkan waktu yang cukup lama. Oleh karena itu, diperlukan suatu sistem yang mampu mendeteksi persentase kemiripan judul secara cepat dan tepat.

Tujuan penelitian yaitu membentuk suatu system yang membantu dalam membandingkan beberapa tugas akhir. Dalam sistem ini judul di input kedalam sebuah sistem kemudian data akan melewati beberapa tahapan yaitu: *Text Minning (Tokenzing, Filtering, Stemming, Stopword Removal)*. Proses berikutnya yaitu pembobotan TF_IDF dan perhitungan *Cosine Similarity*. Hasil akhir dari proses tersebut adalah tingkat kemiripan antar judul yang di uji.

Berdasarkan tahap pengujian yang di lakukan maka dapat disimpulkan bahwa algoritma *cosine similarty* dapat melakukan perhitungan secara akurat terhadapat kata dengan proses *stemming*.Hasil pengujian dari empat skema yang di lakukan yaitu, jika membandingkan judul penelitian yang sama dengan judul penelitian yang sama tingkat kemiripanya bobot *cosine similarity* sebesar 100%. Dalam pengujian kombinasi kata dalam judul maka tingkat nilai *cosine similarity* yang di hasilkan adalah 100%. Dalam pegujian penghapusan tiga kata di judul yang sama maka tingkat kemiripan *cosine similarity* yang dihasilkan adalah 85%. Dalam pengujian persamaan kata (sinonim) maka bobot yang di hasilkan adalah 92%.

Kata Kunci: Judul, *Text Mining*, TF_IDF, *Cosine Similarity*, Hasil.

1. Pendahuluan

Perkembangan Judul penelitian tugas akhir saat ini menyediakan cara yang mudah untuk mencari literatur akademis secara luas. Seseorang dapat mencari di seluruh bidang ilmu dan referensi dari satu tempat: makalah peer-reviewed, thesis, buku, abstrak, dan artikel, dari penerbit akademis, komunitas profesional, pusat data pracetak, universitas, dan organisasi akademis lainnya. Judul Skripsi saat ini akan membantu seseorang mengidentifikasi penelitian paling relevan atau tidak relevanya suatu dari penelitian akademis.

Ketidak mampuan mahasiswa dalam memilih judul penelitian memang banyak sekali faktornya, salah satu diantaranya adalah karena tidak tahu atau bingung dalam menentukan tema. Ada beberapa mahasiswa yang sudah memiliki ide untuk judul skripsi mereka, akan tetapi belum dapat menentukan apakah judul tersebut akan diterima atau ditolak. Hal ini terjadi karena tidak adanya sistem yang menyediakan informasi mengenai daftar judul-judul yang telah diambil dan tidak adanya sistem yang mampu mengecek seberapa besar persentase kemiripan terhadap judul-judul yang telah telah ada.

Dalam menentukan diterima atau tidaknya sebuah judul skripsi yang sekarang ini dilakukan adalah dengan mengecek atau membandingkan judul tersebut dengan daftar judul-judul yang telah ada secara manual yaitu dengan melihat satu-persatu. Tentu saja hal ini akan membutuhkan waktu yang cukup lama belum lagi jika pengaju judul skripsi berjumlah banyak, tidak menutup kemungkinan waktu yang dibutuhkan juga akan semakin lama. Di samping itu, pengelolaan daftar judul-judul yang telah ada juga

masih manual sehingga tidak menutup kemungkinan data akan hilang dan sulit dalam mendistribusikan informasi tersebut kepada mahasiswa.

Oleh karena itu, diperlukan suatu sistem yang mampu mendeteksi persentase kemiripan judul secara cepat dan tepat dan mampu menyediakan informasi tersebut kepada seluruh mahasiswa. Dengan sistem ini, Koordinator Tugas Akhir atau Ketua Program Studi nantinya hanya memasukkan judul yang akan diajukan ke formulir yang telah disediakan, kemudian sistem akan mengecek secara otomatis dan menampilkan hasilnya. Hasil tersebut bisa dijadikan sebagai pertimbangan dalam menentukan apakah diterima atau ditolak judul tersebut.

Dalam proses penelusuran Judul Skripsi diperoleh informasi yang sangat banyak, akan tetapi sebagian besar diantaranya adalah informasi yang tidak dibutuhkan. Oleh karena itu, dari sudut pandang temu kembali informasi (*information retrieval*), maka di butuhkan Judul-judul yang relevan terhadap query dan mengurutkan Judul tersebut berdasarkan kesesuaiannya dengan query berdasarkan topik yang akan di teliti.

Penelitian tentang membandingkan dan mendeteksi kemiripan judul penelitian telah dilakukan dan dikembangkan oleh beberapa pengembang, yaitu berdasarkan penelitian yang telah dilakukan oleh (Sekarwati, 2015) yaitu tentang " Pengukuran Kemiripan Dokumen dengan Menggunakan *Tools Gensim*. Hasil pengujian kelompok dokumen similar menghasilkan prosentase kemiripan dengan Cosinus yang hampir semuanya mendekati 100% kecuali hasil pengukuran kemiripan Dice dan Jaccard.

Metode *cosine similarity* tidak memperhatikan makna semantic dari setiap kata. Berbeda dengan *distance based method*, *information content based method* dan *distance* and *information content based method* (Styawati, Yulita and Sarasvananda, 2020).

Penelitian tentang kemiripan dokumen juga pernah dilakukan oleh (Nurdiana 2016) tentang "Perbandingan Metode *Cosine similarity* dengan Metode *Jaccard Similarity* pada Aplikasi Pencarian Terjemah Al-qur'an Dalam Bahasa Indonesia". Hasil penelitian dari penelitian tersebut menyatakan bahwa metode cosine similarity dapat menunjukan nilai rata-rata 47% sedangkan metode jaccard similarity hanya menunjukan nilai rata-rata 21%. Dari hasil data tersebit maka dapat disimpulkan bahwa metode cosine similarity lebih efektif daripada metode jaccard similarity.

Berdasarkan urain di atas, maka penulis memilih untuk mengembangkan algoritma *Cosine Similarity* karena berbagai alasan, yaitu dari hasil perbandingan yang di dapat pada penelitian terdahulu menunjukan hasil bahwa algoritama *Cosine Similarity* lebih baik dari algoritma yang lain, sehingga penulis merasa tertarik untuk mengembangkan algoritma *Cosine Similarity* untuk dijadikan untuk sistem perbandingan Judul penelitian Skripsi.

1.1 Rumusan Masalah

Dari latar belakang yang ada dapat di rumuskan beberapa masalah:

1. Bagaiamana rancangan arsitektur sistem pendekteksi kesamaan judul penelitian tugas akhir
2. Apakah hasil pengukuran pendektsi kesamaan judul tugas akhir sesuai dengan yang di inginkan atau akurat?
3. Bagaiamana evaluasi hasil pengukuran?

1.2 Batasan Masalah

Dalam batasan masalah ini, penyusun membatasi permasalahan yang perlu, yaitu:

1. Hasil perbandingan judul penelitian tugas akhir relevan tidak nya judul di lihat dari nilai cosine tertinggi.
2. Judul Skripsi yang digunakan hanya berektensi (*extension*), doc, txt, dan pdf untuk penelitian perbandingan judul penelitian tugas akhir.
3. Judul Skripsi yang digunakan untuk penelitian adalah judul yang di ambil dari Google Cendekia*(Google Scholar*) berbahasa indonesia.
4. Pencarian judul menggunakan kata kunci suatu *query*.
5. Sistem tidak bisa mendeteksi persamaan kata dalam judul yang di teliti.

1.3 Tujuan Penelitian

Adapun tujuan dari perbandingan Judul penelitian tugas akhir adalah membuat sistem deteksi kemiripan Judul Skripsi menggunakan algoritama *Cosine Similarity*.

1.4 Manfaat Penelitian

Manfaat yang bisa di ambil dari penelitian ini adalah sebagai berikut:

Dengan adanya sistem ini diharapkan dapat memudahkan membandingkan judul skripsi dengan cara terkomputerisasi.

2. Tinjauan Pustaka

2.1 *Website*

Website adalah suatu halaman web yang saling berhubungan yang umumnya berada pada peladen yang sama berisikan kumpulan informasi yang disediakan secara perorangan, kelompok, atau organisasi. Sebuah situs web biasanya ditempatkan setidaknya pada sebuah *server web* yang dapat diakses melalui jaringan internet.

Website atau disingkat web, dapat diartikan sekumpulan halaman yang terdiri dari beberapa laman yang berisi informasi dalam bentuk data digital baik berupa text, gambar, video, audio, dan animasi lainnyayang disediakan melalui jalur internet. Lebih jelasnya, *website* merupakan halaman-halaman yang berisi informasi yang ditampilkan oleh *browser* seperti *Mozila Firefox, Google Chrome* atau yang lainnya." (Rohi Adulloh,2016). Website merupakan fasilitas internet yang menghubungkan dokumen dalam lingkup lokak maupun jarak jauh. Dokumen pada website disebut dengan web page dan link dalam website memungkinkan pengguna bisa bepindah dari satu halamanke-halaman yang lain, baik diantara halaman yang disimpan dalam server maupun server diseluruh dunia. (Hakim Lukmanul,2004).

2.4.1 *Software* Yang Dibutuhkan Dalam Pembuatan *Website* Dalam proses pembuatan *website*, ada beberapa macam software yang dibutuhkan diantaranya :

A. Text Editor

Text Editor digunakan untuk menuliskan skrip *HTML, CSS, PHP*, dan lainnya. Ada banyak editor yang dapat digunakan diantaranya Notepad, *Notepad++, Adobe Dreamweaver*, dan lainnya.

B. Paket *Apache dan Phpmyadmin*

Apache berfungsi sebagai *web server*, yaitu tempat menyimpan filefile php dan file-file lainnya yang diperlukan dalam *website*, sedangkan Phpmyadmin merupakan apikasi berbasis web yang digunakan untuk membuat database *MySQL* sebagai tempat untuk menyimpan data-data website. Keduanya, biasaya sudah disediakan dalam satu paket aplikasi seperti *Appserv* atau *Xampp*.

C. *Web Browser*

Web browser digunakan untuk menampilkan hasil *website* yang telah dibuat. Web browser yang paing sering digunakan diantaranya adalah *Mozilla Firefox, Google Chrome, dan Safari.*

2.4.2. Unsur-Unsur *Website*

Untuk menyediakan sebuah website, maka harus tersedia unsur-unsur

penunjangnya, adalah sebagai berikut:

A. Nama *Domain* (domain name) : Nama *domain* adalah alamat unik di dinia internet yang digunakan untuk mengidentifikasi sebuah *website*, atau dengan kata lain alamat yang digunakan untuk menemukan sebuah *website* pada dunia *internet*. Nama domain diperjualbelikan secara bebas di internet dengan status sewa tahunan.

B. Rumah tempat *website* (*web hosting*) : Web Hosting dapat diartikan sebagai ruangan yang terdapat menyimpan berbagai data, file-file, gambar, video, data *email*, statistik, *database*, dan lain sebagainya yang akan ditampilkan di website. Web Hosting juga diperoleh dengan menyewa. Pengguna akan memperoleh kontrol panel yang terproteksi dengan *username* dan *password* untuk administrasi websitenya.

C. Bahasa Program (*Scripts Program*) : Bahasa yang digunakan untuk menerjemahkan setiap perintah dalam *website* yang pada saat diakses. Jenis bahasa program sangat menentukan statis, dinamis, atau

interaktifnya sebuah *website*. Semakin banyak bahasa program yang digunakan maka akan terlihat *website* semakin dinamis dan interaktif. Jenis-jenis bahasa program yang banyak dipakai antara lain: HTML, ASP, PHP, JSP, Java Scripts, Java Applets, XML, Ajax dan sebagainya.

D. Desain *Website* : Setelah melakukan penyewaan domain name dan web *hosting* serta penguasaan bahasa program, unsur website yang penting dan utama adalah desain. Desain *web* menentukan kualitas dan keindahan sebuah *website*. Desain akan berpengaruh kepada penilaian pengunjung akan bagus tidaknya sebuah website.

E. Program *transfer* data ke pusat data : FTP (*File Transfer Protocol*) merupakan akses yang diberikan pada saat kita memesan web hosting, FTP berguna untuk memindahkan file-file *website* yang ada pada komputer kita ke pusat *web hosting* agar dapat terakses ke seluruh dunia.

2.2 Database

2.2.1 Pengertian Database

Berikut pengertian *database* menurut para ahli:

A. *Database* adalah struktur penyimpanan data. Untuk menambah, mengakses, dan memproses data yang disimpan dalam sebuah *database* komputer, diperlukan sistem manajemen *database* seperti MySql Server (Kustiyaningsih, 2011).

B. *Database* adalah sekumpulan tabel-tabel yang berisi data dan merupakan kumpulan dari *fieldI* atau kolom. Struktur file yang menyusun sebuah *database* adalah Data *Record* dan *Field* (Anhar, 2010).

C. *Database* adalah sebagai kumpulan data dari penempatan tenaga kerja yang saling terkait dan mempengaruhi sesuai dengan tingkat kepentinganya sehingga data tersebut terintegrasi dan *independence* (Martono, 2009).

Berdasarkan beberapa pendapat para ahli dikemukakan di atas dapat dtarik kseimpulan bahwa *databse* adalah sekelompok data yang mempuyai ciri-ciri khusu dan dapat dikelola sedemikian rupa sehingga bisa menghasilkan sebuah format data yang baru.

2.3 MySQL

Berikut pengertian MySQL menurut para ahli:

MySQL merupakan RDBMS atau *server database* yang mengelola *database* dengan cepat menampung dalam jumlah sangat besar dan dapat di akses oleh banyak user (Raharjo, 2011).MySQL adalah sebuah *software open source* yang digunakan untuk membuat sebuah *database* (Kadir, 2008).

Berdasarkan pendapat yang dikemukakan di atas dapat ditarik kesimpulan bahwa MySQL adalah suatu *software* atau program yang digunakan untuk membuat sebuah *database* yang bersifat *open sourc.* MySQL adalah sebuah implementasi dari sistem manajemen basis data relasional (RDBMS) yang didistribusikan secara gratis dibawah lisensi GPL (*General Public License*). Setiap pengguna dapat secara bebas menggunakan MySQL, namun dengan batasan perangkat lunak tersebut tidak boleh dijadikan produk turunan yang bersifat komersil.

MySQL sebenarnya merupakan turunan salah satu konsep utama dalam basis data yang telah ada sebelumnya SQL (*Structural Query Language).* SQL adalah sebuah konsep pengoperasian basis data, terutama untuk pemilihan atau seleksi dan pemasukan data, yang memungkinkan pengoperasian data dikerjakan dengan mudah secara otomatis.

Kehandalan suatu sistem basis data (DBMS) dapat diketahui dari cara kerja pengoptimasi-nya dalam melakukan proses peintah-perintah SQL yang dibuat oleh pengguna maupun program-program aplikasi yang memanfaatkanya. Sebagai penyedia basis data, MySQL mendukung operasi basis data transaksional maupun operasi basis data non-traksasional.

2.4 Sistem Temu Kembali Informasi

Sistem temu kembali informasi berasal dari kata *Information Retrieval System* (IRS). Temu kembali informasi adalah sebuah media layanan bagi pengguna untuk memperoleh informasi atau sumber informasi yang dibutuhkan oleh pengguna. Berikut ini beberapa definisi sistem temu kembali informasi menurut para ahli di bidang ilmu perpustakaan dan informasi, yaitu:

Hougthon (1977) menjelaskan bahwa sistem temu kembali informasi adalah penelusuran yang merupakan interaksi antara pemakai dan sistem dan pernyataan kebutuhan pengguna diekspresikan sebagai suatu istilah tertentu. Sedangkan menurut Salton (1983) secara sederhana menjelaskan bahwa temu kembali informasi merupakan suatu sistem yang menyimpan informasi dan menemukan kembali informasi tersebut.

Harter (1986) mengatakan bahwa Sistem temu kembali informasi (Information Retrieval System/IRS) adalah perangkat yang menghubungkan antara pemakai potensial dengan koleksi atau kumpulan informasi. Sedangkan menurut Sulistyo-Basuki (1991) mendefinisikan temu kembali informasi sebagai kegiatan yang bertujuan untuk menyediakan dan memasok informasi bagi pemakai sebagai jawaban atas permintaan atau berdasarkan kebutuhan pemakai.

Baeza-Bates dan Riberto-Neto (1999) mengatakan bahwa temu kembali informasi berkaitan dengan representasi, penyimpanan, dan akses terhadap dokumen representasi dokumen. Temu kembali informasi sebagai suatu proses pencarian dokumen dengan menggunakan istilah luas untuk mengidentifikasi dokumen yang berhubungan dengan subjek tertentu. Artinya dalamproses penemuan informasi perlu digunakan istilah-istilah tertentu (Lancaster dalam Hardi, 2006).

Sistem temu kembali informasi merupakan sistem yang mampu melakukan pencarian informasi pada kumpulan dokumen, pencarian dokumen itu sendiri, pencarian meta data untuk dokumen tersebut, atau pencarian teks, suara, gambar, atau data dalam basis data dan pengambilan dokumen yang relevan dari sebuah koleksi dokumen sesuai dengan *query* pengguna sistem. Input dari suatu sistem temu balik informasi adalah *query* dari pengguna dan koleksi dokumen,dan outputnya adalah dokumen yang dianggap relevan oleh sistem. Sistemtemu balik informasi ini digunakan untuk mengurangi informasi yang terlalu banyak sehingga sulit untuk dikelola. Dengan adanya sistem temu balik informasi maka diharapkan pencarian informasi dapat dilakukan dengan efektif dan memberikan hasil pencarian yang tepat (Ramadhany. 2008).

Sistem temu balik informasi terdiri dari komponen-komponen yang saling berkaitan satu sama lain. Menurut Chowdury 1999 dalam Zaenab, (2002), pada intinya dalam sistem temu balik informasi terdapat tiga komponen utama yang saling mempengaruhi, yaitu:

A. Kumpulan dokumen
B. Kebutuhan informasi pengguna
C. Proses pencocokan (*matching*) antara keduanya.

Adapun komponen-komponen sistem temu balik informasi menurut (Hasugian, 2008) antara lain:

A. Pengguna

Pengguna sistem temu kembali informasi adalah orang yang menggunakan atau memanfaatkan sistem dalam rangka kegiata pengelolaan dan pencarian informasi. Berdasarkan perannya, pengguna sistem dibedakan atas 2 kelompok, yaitu pengguna (user) dan pengguna akhir (*end user*).

B. Query

Query adalah format bahasa permintaan yang diinput oleh pengguna kedalam sistem temu balik informasi. Dalam interface sistem selalu disediakan kolom/ruas sebagai tempat bagi pengguna untuk mengetikkan querynya.

C. Dokumen

Dokumen adalah istilah yang digunakan untuk seluruh bahan pustaka, apakah itu artikel, buku, laporan penelitian dsb. Seluruh bahan pustakan dapat disebut sebagai dokumen.

D. Indeks Dokumen

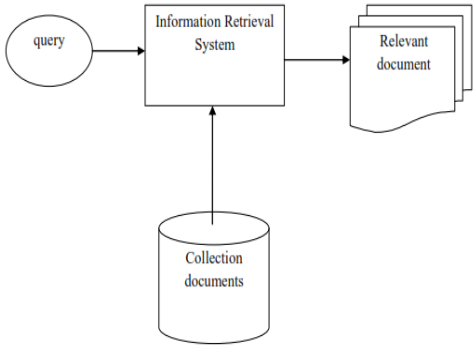
Indeks adalah daftar istilah atau kata (*listofterms*). Dokumen yang dimasukkan disimpan dalam database diwakili oleh indeks. Indeks ini disebut indeks dokumen

E. Pencocokan/*MatcherFungtion.*

Pencocokkan istilah *(query)* yang dimasukkan oleh pengguna dengan indeks dokumen yang tersimpan dalam database adalah dilakukan oleh mesin komputer. Komputerlah yang melakukan proses pencocokkan ini dalam waktu yang sangat singkat sesuai dengan kecepatan memory dan processing yang dimiliki oleh komputer.

2.7.1 Proses dalam Sistem Temu Kembali Informasi

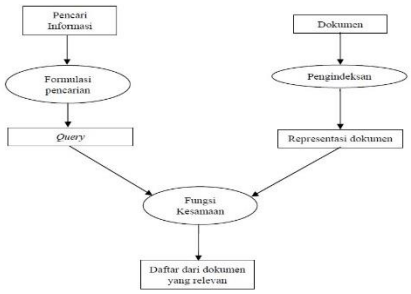
Proses dalam Sistem Temu Kembali Informasi dapat digambarkan sebagai sebuah proses untuk mendapatkan relevant documents dari collection documents yang melalui pencarian *query* yang diinputkan user dapat dilihat pada gambar 2.2.



Gambar 2.1 Proses sistem temu kembali informasi

Kerangka dari sistem temu kembali informasi sederhana terbagi menjadi dua bagian. Bagian yang pertama adalah bagian si pencari informasi atau pengguna dari sistem. Pengguna dari sistem temu-kembali informasi harus menerjemahkan informasi yang dicarinya agar dapat diproses oleh sistem dengan cara memasukan query. *Query* tersebut nanti diproses menjadi sebuah *query* yang dapat dimengert oleh komputer.

Bagian yang kedua adalah bagian dari dokumen. Pada bagian ini dokumen-dokumen direpresentasikan dalam bentuk indeks. Nanti *query* dari pengguna akan diproses melalui fungsi kesamaan untuk membandingkan *query* dengan indeks dari dokumen untuk mendapatkan dokumen yang relevan. Untuk lebih jelasnya mengenai kerangka sistem temu kembali informasi dapat dilihat pada Gambar 2.2

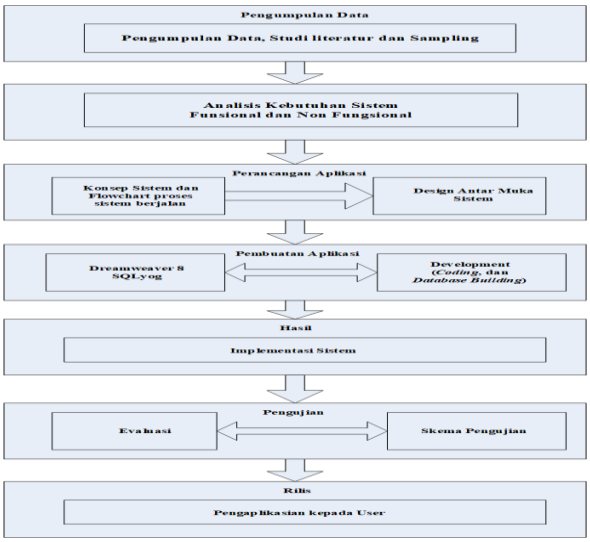


Gambar 2.2 Kerangka Dari Sistem Temu Kembali Informasi(Ingwersen, 1992)

3. Metodologi Penelitian

Kerangka Pemikiran

Kerangka pemikiran merupakan gambaran umum dari rangkaian proses yang dilakukan dalam suatu penelitian. Berikut merupakan kerangka pemikiran dari penelitian yang dilakukan.



Gambar 3.1 Kerangka Pemikiran
(Sumber : Olahan Penulis)

3.2 Jenis Penelitian

Teknik sampling adalah merupakan teknik pengambilan sampel (Sugiyono, 2001: 56). Pengertian teknik sampling menurut Margono (2004) adalah: Teknik sampling adalah cara untuk menentukan sampel yang jumlahnya sesuai dengan ukuran sampel yang akan dijadikan sumber data sebenarnya, dengan memperhatikan sifat-sifat dan penyebaran populasi agar diperoleh sampel yang representatif.

3.2.1 Data yang digunakan

Data yang digunakan dalam penelitian ini adalah judul skripsi yang berekstensi pdf. Judul skripsi tersebut akan menjadi bahan untuk membandingkan kemiripan antar judul penelitian.

3.2.2 Metode Pengumpulan Data

Metode pengumpulan data yang dilakukan pada penelitian ini adalah sebagai berikut :

A. Studi Literatur

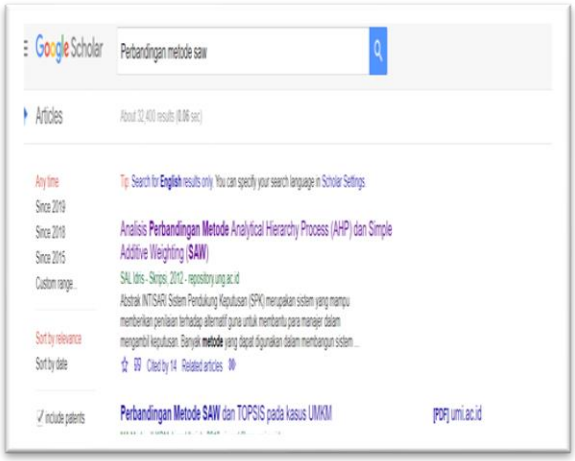
Pada tahap ini peneliti mengumpulkan informasi dan mempelajari materi serta sumber-sumber data yang diperlukan untuk membangun sistem perbandingan judul skripsi dengan algoritma *Cosine Similarity.*

Penelusuran Data *Online* / *Internet Searching* merupakan teknik pengumpulan data melalui bantuan

teknologi yang berupa alat / mesin pencari di internet dimana segala informasi dari berbagai era tersedia didalamnya. Internet searching sangat memudahkan dalam rangka membantu peneliti menemukan suatu file /data dimana kecepatan, kelengkapan dan ketersediaan data dari berbagai tahun tersedia. Mencari data di internet bisa dilakukan dengan cara searching, browsing, surfing ataupun downloading.

B. Sampling

Pada tahap ini peneliti membutuhkan suatu sampel data yang yang berekstensi pdf. Data yang akan dijadikan sebagai bahan uji coba sistem perbandingan judul penelitian tugas akhir. Sampel data yang digunaka judul artikel yang digunakan untuk perbandingan penelitan *system* yang di download dari https://scholar.google.com/. Contoh metode pengumpulan data artikel yang akan di teliti dapat di lihat di gambar di bawah ini:



Gambar 3.2 Pengumpulan Data Artikel

3.4 Analisis Kebutuhan

3.4.1 Analisis Kebutuhan Fungsional

Kebutuhan fungsional adalah kebutuhan yang berisi proses-proses apa saja atau layanan apa saja yang nantinya harus disediakan oleh sistem, mencakup bagaimana sistem harus bereaksi pada input tertentu dan bagaimana perilaku system pada situasi tertentu. Kebutuhan fungsional dapat dilihat pada Tabel 3.1 dan table.

3.4.2 Analisis Kebutuhan NonFungsional Sistem

Analisis perangkat keras adalah *device* yang digunakan untuk menunjang dalampembuatan sistem. Dalam pembuatan sistem ini perangkat keras yang digunakan yaitu komputer dengan spesifikasi :

A. Processor Core inside
B. RAM 4 GB
C. Intel HD
D. HDD 500 GB
E. Monitor 20”
F. Mouse
G. Keyboard

3.4.3 Analisis Kebuthan Perangkat Lunak

Perangkat lunak adalah program atau aplikasi yang digunakan untuk membangun sistem. Perangkat lunak yang dibutuhkan dalam pembuatan sistem ini adalah:

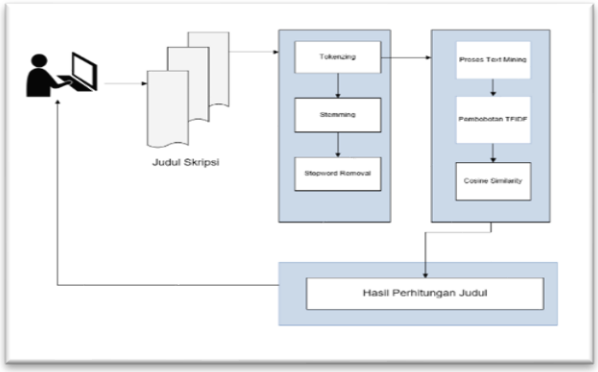
A. Windows 7
B. DreamWeaver CS8
C. Appserv
D. SQLyog Enterprise
E. Visio 2013
F. Astah

3.5 Perancangan Aplikasi

3.5.1 Konsep Sistem

System yang akan dibangun dalam penelitian ini berjudul “Implementasi *Algoritma Cosine Similarity* untuk Mendeteksi Kemiripan Topik Judul Skripsi”. Ide dasar dari *System* ini yaitu mengemas informasi-informasi umum mengenai ide bagaimana membandingkan judul skipsi dengan cara terkomputerisasi.

Pada bab ini akan dibahas mengenai perancangan sistem, yaitu formulasi penyelesaian masalah secara sederhana, perancangan antar muka pengguna, gambaran umum sistem yang berjalan dan diagram alir proses sistem. Gambar berikut merupakan diagram alir proses perbandingan judul skripsi.



Gambar 3.3 Gambaran umum sistem berjalan

Tabel 3.4 Cara kerja proses gambaran umum sistem yang berjalan

NO	User	*System*
1	Input Judul Skripsi	Proses text mining(*tokenizing, stemming,dan Stopword removal*)
2		Pembobotan TF_IDF
3		Perhitungan *Cosine Similarity*
4	user	Hasil

4. HASIL DAN PEMBAHASAN

Tahap implementasi ini dilakukan setalah tahap analisa dan perancangan selesai dilakukan.Dalam tahap implementasi ini akan dilakukan pengkodingan terhadap apa yang telah dilakukan pada tahap analisa dan perancangan.Diharapkan hasil dari implementasi ini sesuai dengan apa yang menjadi tujuan awal penelitian ini.Dalam tahapan ini proses yang dilakukan bertujuan untuk merancang sebuah aplikasi pendeteksi kemiripan topic judul penelitian menggunakan algoritma *cosine similarty* dengan proses bantuan pre – processing.Aplikasi yang dihasilkan dari proses implementasi pada pembuatan skripsi ini merupakan sebuah aplikasi *system*.

Dikatakan aplikasi system karena aplikasi yang dibangun merupakan jaringan kerja yang terdiri dari prosedur – prosedur yang saling berhubungan satu sama lain untuk melakukan suatu usaha atau sasaran tertentu, yaitu mendeteksi kemiripan topik judul penelitian. Pada umumnya di dalam sebuah *system* terdapat fungsi –fungsi teknologi seperti *input, process, output, storage* dan *communication*. Selain *hardware*, *software*, informasi dan prosedur –prosedur dalam bentuk program, factor brainware sebagai salah satu esensi *system* juga dibutuhkan. Sistem pedeteksi kemiripan yang dirancang dalam penulisan skripsi ini mengadopsi penjelasan tersebut dengan mengikut – sertakan pengguna sebagai user, operator dan *decision* maker terhadap bobot/nilai hasil keluaran *system* dalam menentukan pasangan judul apakah termasuk hasil tindakan yang pernah di ajukan sebelumnya. Hal ini menjelaskan bahwa bobot yang dihasilkan oleh *system* pendeteksi kemirpan topic judul penelitian ini merupakan nilai relatif.

4.1 Batasan Implementasi

Adapun batasan implementasi pada aplikasi perbandingan judul penelitian ini adalah sebagai berikut:

Input masukan judul penelitian yang akan di *stemming*, ini dimaksudkan untuk menemukan kata dasar dari sebuah kata. 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

4.2 Lingkungan Implementasi

Lingkungan implementasi adalah lingkungan dimana aplikasi ini dikembangkan. Lingkungan implementasi aplikasi perbandingan judul penelitian ini terdiriatas dua lingkungan yaitu lingkungan perangkat keras dan perangkat lunak.Berikut ini merupakan spesifikasi lingkungan tersebut:

1. Perangkat Keras

Berikut spesifikasi dari perangkat keras yang digunakan:

- Processor :Intel® CPU N3350 @1.10GHz
- Memori (RAM) : 4,00 GB
- Harddisk : 320 GB
- 2. Perangkat Lunak
- Sistem Operasi : Windows 10
- Bahasa Program: PHP (Hypertext Prepocessor)
- Tools Perancangan : Notepad++, Google Chrome

4.3 Pengujian *System*

Pengujian sistem akan dilakukan untuk mengetahui apakah sistem yang dibangun sudah sesuai dengan analisa dan tujuan dari dibangunnya sistem ini.Untuk mengetahui hasil dari sistem ini apakah sudah sesuai dengan yang diharapkan, dilakukan penghitungan *similarity* dengan inputan judul yang diuji.

4.5.1 Rencana Pengujian

Dalam tahapan rencana pengujian akan dilakukan sebagai berikut:

A. Untuk tahapan pengujian sistem akan dilakukan menggunakan 30 judul yang akan dibandingan dengan 500 judul yang tersimpan pada *Database* jurnal yang dugunakan jurnal ISSN yang di dapat dari *google scholar*.
B. Dalam tahapan pengujian pertama dilakukan dengan cara membandingkan judul penelitian yang sama dengan judul peneltian yang tersimpan dalam *database* maka terlihat tingkat kemiripanya bobot *cosine similarity* berapa persen.
C. Dalam tahapan pengujian kedua inputan awal dengan menggunakan judul yang sama dengan konsep mengubah susunan kata atau kombinasi acak kata dalam judul penelitian, maka tingat nilai *cosine similarity* hasil akan dibandingkan dengan judul tanpa mengubah susunan kata apakah bobot nilai cosine similarity akan berubah apa akan tetap sama.
D. Dalam pegujian ketiga dilakukan dengan cara penghapusan tiga kata di judul yang sama maka akan dilihat nilai tingkat kemiripan *cosine similarity* berapa persen apakah nilai *cosine similarity* apakah akan turun bobot kemiripannya.

Dalam pengujian keempat dilakukan dengan cara pengujian persamaan kata (sinonim) pada judul yang di teliti Konsep kata yang di gunakan dalam judul yang di pilih adalah kata pertama dalam judul yang di teliti, apakah bobot *cosine similarity* akan tetap sama atau berubah.

Berdasarkan dari empat pengujian diatas data yang digunakan akan terlihat hasil bobot perbandingan judul penelitian, maka hasil perbandingan judul yang di teliti berapakah bobot nilai *cosine similarity* dapat di lihat diringkasan tabel di bawah ini.

4.6 Kesimpulan Hasil Pengujian

Berdasarkan tahap pengujian yang di lakukan maka dapat kita simpulkan bahwa algoritma *cosine similarty* dapat melakukan perhitungan secara akurat terhadapat kata dengan proses *stemming* kata siap judul yang akan di inputkan, berikut ini adalah hasil yang di peroleh dari hasil yang di teliti:

A. Dalam tahapan pengujian kesatu dilakukan dengan cara membandingkan judul penelitian yang sama dengan judul peneltian yang tersimpan dalam *database* maka terlihat tingkat kemiripanya bobot *cosine similarity* sebesar 100%.
B. Dalam tahapan pengujian inputan kedua dengan menggunakan judul yang sama dengan konsep mengubah susunan kata atau kombinasi acak kata dalam judul penelitian, maka tingat nilai *cosine similarity* yang di hasilkan adalah 100%.
C. Dalam pegujian ketiga dilakukan dengan cara penghapusan tiga kata di judul yang sama maka akan dilihat nilai tingkat kemiripan *cosine similarity* bobot yang dihasilkan adalah 85%.
D. Dalam pengujian empat dilakukan dengan cara pengujian persamaan kata (sinonim) pada judul yang di teliti. Konsep kata yang di gunakan dalam judul yang di pilih adalah kata pertama dalam judul yang di teliti, apakah bobot *cosine similarity* maka bobot yang di hasilkan adalah 92%.

5. Kesimpulan dan Saran

Kesimpulan

Berdasarkan hasil analisis dan pembahasan data, penulis memperoleh kesimpulan yang dapat diambil dari penelitian mengenai analisis faktor-faktor yang mempengaruhi tingkat kemirpan hasil perbandingan mengunakan algoritma *cosine similarity* yang teah di uji coba adalah sebagai berikut:

A. Rancangan arsiteksur program perbandingan judul penelitian tugas akhir terdiri dari text mining *tokonzing*, *filtering*, *stemming*, dan *stopword removal* untuk mendekteksi kemiripan judul yang di teliti.

B. Atribut untuk judul yang sama dan judul yang mengunakan susunan berbeda (isi tetap sama) maka tingkat kemiripan yang di hasilkan 100%. Dari pengujian penghapusan kata dalam judul maka diperoleh bobot *cosine similarity* 85%, namun untuk mendekteksi kemirpan judul menggunkan persamaan kata (sinonim) maka tidak dapat medekteksi kemiripan judul.

C. Evaluasi hasil pengujian dari 4 skema yang di lakukan maka di peroleh hasil hasil akhir berikut ini:

1. Dalam tahapan pengujian kesatu dilakukan dengan cara membandingkan judul penelitian yang sama dengan judul peneltian yang tersimpan dalam *database* maka terlihat tingkat kemiripanya bobot *cosine similarity* sebesar 100%.
2. Dalam tahapan pengujian inputan kedua dengan menggunakan judul yang sama dengan konsep mengubah susunan kata atau kombinasi acak kata dalam judul penelitian, maka tingat nilai *cosine similarity* yang di hasilkan adalah 100%.
3. Dalam pegujian ketiga dilakukan dengan cara penghapusan tiga kata di judul yang sama maka akan dilihat nilai tingkat kemiripan *cosine similarity* bobot yang dihasilkan adalah 85%.
4. Dalam pengujian empat dilakukan dengan cara pengujian persamaan kata (sinonim) pada judul yang di teliti. Konsep kata yang di gunakan dalam judul yang di pilih adalah kata pertama dalam judul yang di teliti, apakah bobot *cosine similarity* maka bobot yang di hasilkan adalah 92%.

5.2 Saran

Adapun manfaat yang diharapkan dari hasil penelitian ini bagi pembaca, hasil penelitian ini diharapkan dapat menambah wawasan pengetahuan terkait dengan perbandingan untuk mendeteksi kemiripan topik judul skripsi menggunakan algoritma *cosine similarity*. Berdasarkan hasil dari pengujian sinonim (persamaan kata) *system* perbandingan judul penelitian tugas akhir tidak bisa mendeteksi persamaan kata, untuk itu untuk perkembangan *system* lebih lanjut di perlukan algoritma yang mampu mendeteksi sinonim (persamaan kata) untuk penelitian selanjutnya.

DAFTAR PUSTAKA

A. D. Tahitoe, D. Purwitasari. 2010**. "***Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia dengan Metode Corpus Based Stemming*". Institut Teknologi Sepuluh Nopember (ITS) – Surabaya, 60111, Indonesia.

Andriani, M., & Nazief, B.(1996). "*Approach to stemming algorithm for bahasa indonesia"*. Internal Publication, Falkultas Ilmu Komputer, Universitas Indonesia, Depok, Jakarta.

Anhar. 2010, "*Panduan Menguasai PHP & MySQL Secara Otodidak"*. Mediakita. Jakarta.

Dewa, A. Et al., 2016*, "Pengukuran Kemiripan Dokumen Teks Bahasa Indonesia Menggunakan Metode Cosine Similarity".* E-Journal Teknik Informatika Volume 9, No 1 (2016).

Library,"*library stemming sastrawi*". https://github.com/sastrawi/sastrawi (diakses pada tanggal 11 september 2019).

Library,"*library stopword removal*". https://github.com/gsarwohadi/PySastrawi (diakses pada tanggal 13 september 2019).

Licency," *stopword removal mengunakan kamus kata dasar yang diberikan oleh Kateglo dengan lisensi CC-BY-NC-SA 3.0*". https://creativecommons.org/licenses/by-nc-sa/3.0/ (diakses tanggal 1 september 2019).

Maryaeni, Masnur Muslich.,2009,"*Bagaimana menulis Skripsi*". Jakarta: Bumi Aksara.

Novanta, Audy. (2009)."*Pendeteksian Plagiarisme Dokumen Teks dengan Menggunakan Algoritma Smith-Waterman"*. (http://repository.usu.ac.id, diakses 24 october 2019).

Nurdin, dan Munthoha, A. (2017). "*Sistem pendeteksi kemiripan judul skripsi menggunakan algoritma winnowing".* Jurnal Nasional Informatika Dan Teknologi Jaringan.Vol 2, No 1, September 2017.

Pembobotan."*Term Weighting TF-IDF".* https://informatikalogi.com/term-weighting-tf-idf/. (dikses pada tanggal 17 september 2018).

Tokenzing,"*library tokenzing*". https://github.com/sastrawi/tokenizer (diakses pada tanggal 16 september 2019).

Raharjo. B., 2011, "*Membuat Database Menggunakan MySql".* Informatika. Bandung.

Mulyanto, Agus. (2017). Penyelesaian Kata Ambigu Pada Proses POS Tagging Menggunakan Algoritma Hidden Markov Model (HMM).

Styawati, S., Yulita, W. and Sarasvananda, S. (2020). SURVEY UKURAN KESAMAAN SEMANTIC ANTAR KATA. Jurnal Data Mining dan Sistem Informasi, 1(1), p.32.

Styawati, S. and Mustofa, K. (2019). A Support Vector Machine-Firefly Algorithm for Movie Opinion Data Classification. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 13(3), p.219.

Isnain, A.R., Marga, N.S. and Alita, D. (2021). Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 15(1), p.55.

Alita, D., Fernando, Y. and Sulistiani, H. (2020). IMPLEMENTASI ALGORITMA MULTICLASS SVM PADA OPINI PUBLIK BERBAHASA INDONESIA DI TWITTER. Jurnal Tekno Kompak, 14(2), p.86.