PERANCANGAN SISTEM REKOMENDASI DOKUMEN DENGAN PENDEKATAN CONTENT-BASED FILTERING

Wayan Gede Suka Parwita[1], Made HanindiaPrami Swari[2], Welda[3]
[123]*Program Studi Teknik Informatika, STMIK STIKOM Indonesia*
Jln. TukadPakerisan, Denpasar-Bali, Indonesia
gede.suka@gmail.com[1], hanindia@stiki-indonesia.ac.id[2], welda@yahoo.com[3]

***Abstrak* — Penentuan dosen Pembimbing Tugas Akhir merupakan hal yang riskan dalam menunjang keberhasilan dan kelancaran mahasiswa dalam menempuh Tugas Akhir. Di STMIK STIKOM Indonesia, mahasiswa mengajukan dosen pembimbing hanya berdasarkan tingkat kedekatan tanpa mempertimbangkan relevansi topik tugas akhir yang diangkatnya dengan bidang keahlian dosen yang diajukan. Berbagai batasan yang muncul dalam menentukan dosen pembimbing bag masing-masing mahasiswa menyebabkan kegiatan ini menjadi permasalahan tersendiri di STMIK STIKOM Indonesia. Permasalahan penentuan dosen pembimbing ini dapat diselesaikan dengan memabangun sebuah sistem rekomendasi yang membandingkan kedekatan topik penelitian mahasiswa dengan bidang keilmuan seluruh dosen yang tersedia melalui dokumen penelitian-penelitian yang pernah dilakukan oleh masing-masing dosen. Penggunaan sistem rekomendasi ini akan menghasilkan rekomendasi dosen pembimbing tugas akhir yang memiliki bidang minat dan keahlian yang mendekati topik penelitian tugas akhir yang diajukan oleh mahasiswa.**

***Kata kunci* — Sistem Rekomendasi Dokumen, Content-based Filtering.**

I. PENDAHULUAN

Penentuan dosen pembimbing di STMIK STIKOM Indonesia saat ini menggunakan dokumen Usulan Proposal Penelitian (UPP) sebagai dasar pertimbangan untuk menentukan dosen pembimbing. Koordinator KP dan TA STMIK STIKOM Indonesia harus menyesuaikan usulan calon dosen pembimbing dari mahasiswa dengan bidang ilmu calon dosen pembimbing sebelum menetapkan dosen pembimbing tugas akhir mahasiswa. Dokumen yang digunakan sebagai pertimbangan dalam penentuan dosen pembimbing adalah usulan proposal penelitian (UPP) yang diajukan oleh masing-masing mahasiswa yang akan menempuh tugas akhir. Jumlah mahasiswa yang mengajukan usulan proposal penelitian dalam satu semester mencapai kurang lebih 200 mahasiswa, sehingga akan sangat menyulitkan bagi koordinator KP dan TA dalam menentukan kesesuaiaan usulan proposal penelitian mahasiswa dengan bidang ilmu yang relevan dari calon dosen pembimbingnya masing-masing. Kesulitan lain yang dialami dalam penentuan dosen pembimbing tugas akhir adalah kurangnya dokumen serta informasi tentang penelitian-penelitian yang dilakukan oleh calon dosen pembimbing. Permasalahan ini diperparah dengan jumlah dosen yang menguasai suatu bidang ilmu tertentu jumlahnya terbatas, sehingga seringkali satu dosen pembimbing membimbing mahasiswa dengan topik yang tidak sesuai dengan bidang keahliannya, sedangkan ada dosen pembimbing lain yang sebenarnya menguasai bidang ilmu sesuai topik yang diajukan oleh mahasiswa tersebut. Hal ini tentunya menjadi permasalahan tersendiri bagi dosen pembimbing tersebut.

Penentuan pembimbing tugas akhir merupakan suatu pengambilan keputusan yang tidak mudah dan riskan. Hal ini disebabkan karena pembimbing merupakan salah satu faktor yang berpengaruh pada pengerjaan tugas akhir mahasiswa. Saat ini, mahasiswa STMIK STIKOM Indonesia cenderung hanya mempertimbangkan faktor kedekatan untuk menentukan calon dosen pembimbing tugas akhir. Pengguanaan formulir pengajuan tugas akhir yang mencantumkan usulan dosen pembimbing tugas akhir juga dirasa kurang tepat guna karena mahasiswa kurang memperhatikan bidang ilmu dari calon dosen pembimbing.

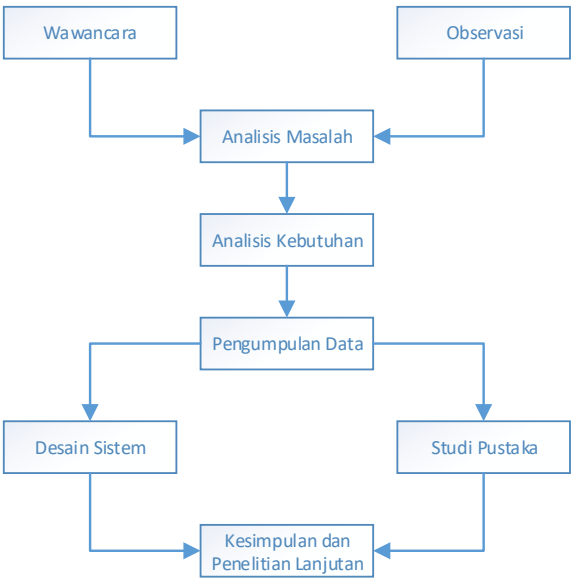
Permasalahan penentuan dosen pembimbing seperti yang terjadi di STMIK STIKOM Indonesia dapat diatasi dengan merancang sebuah sistem rekomendasi. Sistem rekomendasi mulai diperhatikan sejak kemunculan penelitian tentang collaborative filtering pada pertengahan 90'an [4], [12]. Sistem ini memerlukan model rekomendasi yang tepat agar apa yang direkomendasikan sesuai dengan keinginan pengguna, serta mempermudah pengguna mengambil keputusan yang tepat [10]. Sistem rekomendasi dapat di integrasikan ke dalam STIMIK STIKOM Indonesia telah memiliki Syncronized Student's Final Project Management System (Sintesys) untuk memaksimalkan fungsi sistem. Sintesys yaitu sistem informasi yang mengelola kegiatan TA dan KP serta membantu proses yang terlibat didalamnya.

Sistem rekomendasi dibangun dengan memanfaatkan fungsi untuk mengukur relevansi antara UPP mahasiswa dengan bidang ilmu pembimbing melalui dokumen penelitian dosen. Dokumen (dengan format pdf) usulan proposal penelitian dari masing-masing mahasiswa yang mengajukan tugas akhir diupload ke sistem dan akan dibandingkan kemiripannya dengan dokumen penelitian dari seluruh dosen pembimbing yang ada. Penentuan rekomendasi dosen pembimbing dapat mempertimbangkan isi dari UPP yang dapat menggambarkan secara umum penelitian yang akan dilakukan oleh mahasiswa. Dokumen UPP lalu dibandingkan dengan dokumen-dokumen penelitian calon dosen pembimbing yang jumlahnya banyak untuk melihat nilai kedekatan dari dokumen tersebut. Perbandingan kedekatan dapat dilakukan dengan terlebih dahulu melakukan ektraksi isi dari setiap dokumen. Sehingga selanjutnya penilaian kedekatan dapat menggunakan pendekatan cosine similarity. Sistem yang dibangun berupa sistem rekomendasi, dimana peringkat dari rekomendasi ditentukan dari nilai perhitungan *cosine similarity*.

Melalui sistem rekomendasi yang dibuat, rekomendasi pembimbing untuk masing-masing mahasiswa yang mengambil tugas akhir tidak dilakukan dengan mambaca satu-persatu UPP yang diajukan mahasiswa. Penentuan dosen pembimbing akan didasarkan pada nilai tertinggi antara UPP yang diajukan dengan dokumen penelitan dari masing-masing dosen. Manfaat lain yang diharapkan dari sistem ini adalah dosen pembimbing yang ditentukan untuk masing-masing mahasiswa memiliki relevansi bidang ilmu yang mendekati dengan topik penelitian yang diajukan sehingga diharapkan mahasiswa dapat mengerjakan tugas akhir dengan baik dan berdampak pada menurunnya rata-rata waktu mahasiswa menempuh TA.

II. METODOLOGI PENELITIAN

Pengumpulan data dilakukan dengan metode wawancara dan observasi. Wawancara dan Observasi diperlukan untuk pencarian data awal mengenai mekanisme yang ada pada sistem pembagian dosen pemabimbing dan sistem informasi yang berjalan saat ini. Data awal akan digunakan untuk menentukan masalah dan kebutuhan yang diharapkan dari sistem yang akan diimplementasikan.



Gbr. 1 Alur penelitian

Pengumpulan data dan dokumen penelitian dosen akan dilakukan pada tahap pengumpulan data. Sistem yang akandibagun disesuaikan dengan struktur data dan dokumen penelitian dosen. Gambar 1 menunjukkan alur penelitian yang diusulkan.

Secara garis besar, langkah-langkah penelitian yang dilakukan pada penelitian ini adalah sebagai berikut :

a. Analisis Masalah
b. Analisis Kebutuhan
c. Pengumpulan Data
d. Perancangan Sistem

III. KAJIAN PUSTAKA

A. *Ekstraksi Keyword*

Dalam dokumen ilmiah, keyword adalah kata pokok yang merepresentasikan masalah yang diteliti atau istilah-istilah yang merupakan dasar pemikiran dan dapat berupa kata tunggal atau gabungan kata. Similaritas keyword dokumen dapat digunakan untuk menentukan relevansi dokumen terhadap dokumen lain [15]. Automatic keyword extraction system memiliki tugas untuk mengidentifikasi kumpulan kata, frase kunci, keyword, atau segmen kunci dari sebuah dokumen yang dapat menggambarkan arti dari dokumen [6]. Tujuan dari ekstraksi otomatis adalah menekan kelemahan pada ekstraksi manual yang dilakukan manusia yaitu pada kecepatan, ketahanan, cakupan, dan juga biaya yang dikeluarkan. Salah satu pendekatan yang dapat digunakan dalam automatic keyword extraction yaitu pendekatan tata bahasa. Pendekatan ini menggunakan fitur tata bahasa dari kata-kata, kalimat, dan dokumen. Metode ini memperhatikan fitur tata bahasa seperti bagian kalimat, struktur sintaksis, dan makna yang dapat menambah bobot. Fitur tata bahasa tersebut dapat digunakan sebagai penyaring untuk keyword yang buruk. Dalam ekstraksi keyword dengan pendekatan

tata bahasa berbasis struktur sintaksis, ada beberapa tahap yang dilakukan yaitu tokenisasi, stopword removal, stemming, dan pembobotan kata [11].

B. Tokenisasi

Teks elektronik adalah urutan linear simbol (karakter, kata-kata atau frase). Sebelum dilakukan pengolahan, teks perlu disegmentasi ke dalam unit-unit linguistik seperti kata-kata, tanda baca, angka, alpha-numeric, dan lain-lain Proses ini disebut tokenisasi. Tokenisasi sederhana (white space tokenization) merupakan tokenisasi yang memisahkan kata berdasarkan karakter spasi, tab, dan baris baru [15]. Namun, tidak setiap bahasa melakukan hal ini (misalnya bahasa Cina, Jepang, Thailand). Dalam bahasa Indonesia, selain tokenisasi sederhana diperlukan juga tokenisasi yang memisahkan kata-kata berdasarkan karakter lain seperti "/" dan "-".

C. Stopword Removal

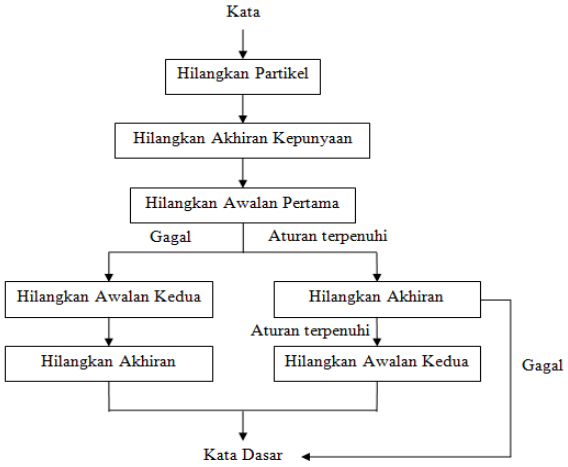
Stopword removal adalah pendekatan mendasar dalam preprocessing yang menghilangkan kata-kata yang sering muncul (stopword). Fungsi utamanya adalah untuk mencegah hasil proses selanjutnya terpengaruh oleh stopword tersebut. Banyak diantara stopword tersebut tidak berguna dalam Information Retrival (IR) dan text mining karena kata-kata tersebut tidak membawa informasi (seperti ke, dari, dan, atau). Cara biasa untuk menentukan apa yang dianggap sebagai stopword adalah menggunakan stoplist. Stoplist merupakan kumpulan kata atau kamus yang berisi daftar stopword. Langkah penghilangan stopword ini adalah langkah yang sangat penting dan berguna [13].

D. Stemming

Algoritma stemming adalah proses yang melakukan pemetaan varian morfologi yang berbeda dari kata-kata ke dalam kata dasar/kata umum (stem). Stemming berguna pada banyak bidang komputasi linguistik dan information retrieval [8]. Dalam kasus bahasa Indonesia, sejauh ini hanya ada dua algoritma untuk melakukan proses stemming yaitu algoritma yang dikembangkan oleh Nazief dan Adriani serta algoritma yang dikembangkan oleh Tala. AlgoritmaNazief dan Adriani dikembangkan dengan menggunakan pendekatan confix stripping dengan disertai pemindaian pada kamus. Sedangkan stemming yang dikembangkan Tala menggunakan pendekatan yang berbasis aturan (rule-based).

Pengembangan Algoritma Tala didasarkan pada kenyataan bahwa sumber daya seperti kamus besar digital untuk bahasa mahal karena kurangnya penelitian komputasi di bidang linguistik. Maka, ada kebutuhan untuk algoritma stemming tanpa keterlibatan kamus. Algoritma Tala sendiri dikembangkan dari algoritma Porter stemmer yang dimodifikasi untuk bahasa Indonesia. Algoritma Tala menghasilkan banyak kata yang tidak dipahami. Ini disebabkan oleh ambiguitas dalam aturan morfologi Bahasa Indonesia. Dalam beberapa kasus kesalahan tidak mempegaruhi kinerja, tetapi dalam kasus lain menurunkan kinerja [14].

E. Tala Stemmer



Gbr. 2 Skema Tala Stemmer [14]

Algoritma Tala memproses awalan, akhiran, dan kombinasi keduanya dalam kata turunan. Walaupun dalam bahasa Indonesia terdapat sisipan, jumlah kata yang diturunkan menggunakan sisipan sangat sedikit. Karena hal tersebut dan juga demi penyederhanaan, sisipan akan diabaikan.

Algoritma Porter stemmer dibangun berdasarkan ide tentang akhiran pada bahasa Inggris yaitu kebanyakan merupakan kombinasi dari akhiran yang lebih sederhana dan lebih kecil. Beberapa perubahan dilakukan pada algoritma Porter stemmer agar sesuai dengan Bahasa Indonesia. Perubahan dilakukan pada bagian kumpulan aturan dan penilaian kondisi. Karena algoritma Porter stemmer hanya dapat menangani akhiran, maka perlu penambahan agar dapat menangani awalan, akhiran, dan juga penyesuaian penulisan dalam kasus dimana terjadi perubahan karakter pertama kata dasar. Gambar II menunjukkan langkah-langkah proses pada algoritma Tala.

Dalam Bahasa Indonesia, unit terkecil dari suatu kata adalah suku kata. Suku kata paling sedikit terdiri dari satu huruf vokal. Desain implementasi algoritma Tala belum dapat mengenali seluruh suku kata. Ini disebabkan karena adanya dua huruf vokal yang dianggap satu suku kata yaitu ai, au, dan oi. Kombinasi dua huruf vokal (terutama ai, oi) tersebut dapat menjadi masalah, apalagi jika berada pada akhir sebuah kata. Ini disebabkan oleh sulitnya membedakannya dengan kata yang mengandung akhiran –i. Hal ini menyebabkan kombinasi huruf vokal ai/oi akan diperlakukan seperti kata turunan. Huruf terakhir (-i) akandihapus pada hasil proses stemming. Kebanyakan kata dasar terdiri dari minimal dua suku kata. Inilah alasan kenapa kata yang akan diproses memiliki minimal dua suku kata.

F. Pembobotan

Tahapan ini dilakukan dengan tujuan untuk memberikan suatu bobot pada term yang terdapat pada suatu dokumen. Term adalah satu kata atau lebih yang dipilih langsung dari corpus dokumen asli dengan menggunakan metode term-extraction. Fitur tingkat term, hanya terdiri dari kata-kata tertentu dan ekspresi yang ditemukan dalam dokumen asli [3].

Dalam pengkategorian teks dan aplikasi lain di information retrieval maupun machine learning, pembobotan term biasanya ditangani melalui metode yang diambil dari metode pencarian teks, yaitu yang tidak melibatkan tahap belajar [2]. Ada tiga asumsi monoton yang muncul di hampir semua metode pembobotan dapat dalam satu atau bentuk lain yaitu [16]:

a. Term yang langka tidak kalah penting daripada term yang sering muncul (asumsi IDF).
b. Kemunculan berkali-kali dari term pada dokumen tidak kalah penting daripada kemunculan tunggal (asumsi TF).
c. Untuk pencocokan term dengan jumlah pencocokan yang sama, dokumen panjang tidak lebih penting daripada dokumen pendek (asumsi normalisasi).

Bobot diperlukan untuk menentukan apakah term tersebut penting atau tidak. Bobot yang diberikan terhadap sebuah term bergantung kepada metode yang digunakan untuk membobotinya.

G. Cosine Similarity

Pendekatan cosine similarity sering digunakan untuk mengetahui kedekatan antara dokumen teks. Perhitungan cosine similarity dimulai dengan menghitung dot product. Dot product merupakan perhitungan sederhana untuk setiap komponen dari kedua vektor. Vektor merupakan representasi dari masing-masing dokumen dengan jumlah term pada masing-masing dokumen sebagai dimensi dari vektor [9]. Vektor ditunjukkan oleh notasi (2.1) dan (2.2). Hasil dot product bukan berupa vektor tetapi berupa skalar. Persamaan (III) merupakan perhitungan dot product dimana n merupakan dimensi dari vector [1].

$$\vec{a} = (a_1, a_2, a_3, \dots, a_n)$$

$$\vec{b} = (b_1, b_2, b_3, \dots, b_n)$$

$\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$

a_n dan b_n merupakan komponen dari vektor (bobot term masing-masing dokumen) dan n merupakan dimensi dari vektor. Cosine similarity merupakan perhitungan yang mengukur nilai cosine dari sudut antara dua vektor (atau dua dokumen dalam vector space). Cosine similarity dapat dilihat sebagai perbandingan antara dokumen karena tidak hanya mempertimbangkan besarnya masing-masing jumlah kata (bobot) dari setiap dokumen, tetapi sudut antara dokumen. Persamaan (IV) dan (V) adalah notasi dari metode cosine similarity dimana $\|\vec{a}\|$ merupakan Euclidean norm dari vektor a dan $\|\vec{b}\|$ merupakan Euclidean norm vektor b [5].

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

Dari notasi (2.5) dapat dibentuk persamaan matematika yang ditunjukkan oleh persamaan (2.6) [7].

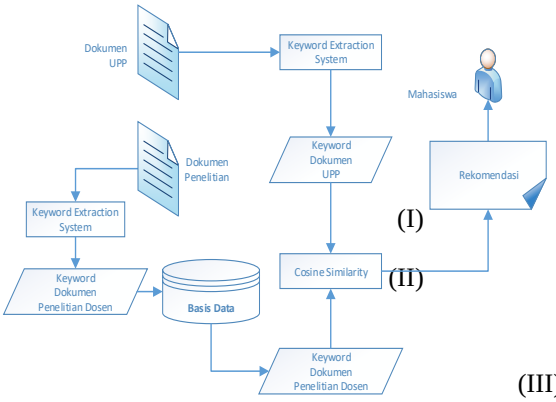
$$Similarity(x, y) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2 \cdot \sum_{i=1}^{n} b_i^2}}$$

Dimana:
ai : term ke-i yang terdapat pada dokumen a.
bi : term ke-i yang terdapat pada dokumen b.

IV. HASIL DAN PEMBAHASAN

A. Gambaran Umum Sistem

Pembangunan sistem akan disesuaikan dengan fasilitas yang ada pada STMIK STIKOM Indonesia. Dokumen baik dokumen UPP maupun penelitian dosen diektraksi untuk menemukan keyword yang ada pada setiap dokumen. Proses ekstraksi melalui beberapa tahap yaitu tokenisasi, stopword removal, steamming dan pembobotan. Proses stopword removal menggunakan stopword yang pada penelitian Parwita dan proses steammingakan memanfaatkan algoritma yang dikembangkan oleh Tala.



Gbr. 3 Gambaran umum sistem

Keyword dari dokumen penelitian dosen akan disimpan dalam basis data agar ekstraksi tidak dilakukan berulang-ulang. Pembandingan kedekatan antara keyword dihitung dengan memanfaatkan pendekatan cosine similariry dengan menghitung

pembobotan yang telah diberikan pada proses ekstraksi keyword. Secara umum, sistem yang akan dibangun ditunjukkan oleh Gambar 3.

Pengembangan sistem akan dibagi menjadi 3 tahap, yaitu tahap pembangunan sistem ekstraksi keyword dengan menyesuaikan bentuk sistem ekstraksi dengan dokumen yang ada, pengembangan sistem rekomendasi dengan memanfaatkan hasil ekstraksi keyword, dan terakhir integrasi sistem rekomendasi ke dalam Sintesys.

B. Analisisa Sistem

Berdasarkan wawancara dan observasi yang dilakukan, diperlukan sebuah sistem rekomendasi pembimbing tugas akhir untuk mahasiswa yang mengajukan tugas akhir. Permasalahan yang terjadi adalah dosen memiliki banyak dokumen penelitian dan pengajuan UPP sangat banyak sehingga tidak dimungkinkan untuk melakukan pemeriksaan bidang yang sesuai dengan pengajuan mahasiswa bedasarkan dokumen penelitian dan UPP.

Pencarian rekomendasi yang dilakukan oleh sistem yang diusulkan membutuhkan dokumen penelitian setiap dosen yang digunakan sebagai pembanding terhadap usulan proposal penelitian (UPP) yang diajukan mahasiswa. Dokumen penelitian dosen akan dicari kemiripannya berdasarkan term-term yang ada pada dokumen tersebut. Term dokumen akan diekstraksi melalui beberapa tahapan seperti tokenisasi, stopword removal, dan stemming. Setelah itu term akan diberikan bobot berdasarkan jumlah kemunculan setiap term. Hal yang sama akan diterapkan pada dokumen UPP mahasiswa. Hasik ekstraksi dan bobot setiap term akan dihitung keiripannya dengan pendekatan cosine similarity.

Selain dokumen, sistem juga memerlukan administrator yang mengelola pengaturan dasar dalam sistem rekomendasi seperti pengelolaan treshold (ambang batas) dan stopword. Administrator sistem merupakan bagian yang terlibat dalam pengelolaan tugas akhir mahasiswa yaitu Koordinator KPTA. Untuk treshold dalam sistem yaitu besaran similaritas dokumen yang direkomendasi dan jumlah rekomendasi yang dimunculkan.

C. Perancangan Sistem

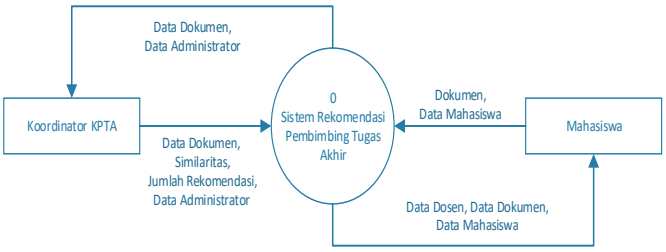
Perancangan yang dibuat pada penelitian ini meliputi perancangan Data Flow Diagram dan Perancangan Struktur Tabel.

1) Data Flow Diagram

Dalam perancangan sistem digunakan model perancangan terstruktur. Dalam memodelkan sistem yang akan dibangun, perancangan yang digunakan adalah data flow diagram (DFD).

a. DFD Level 0

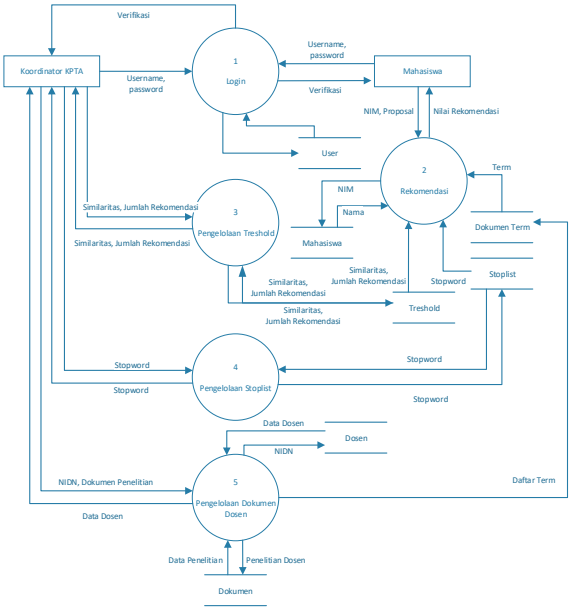
Pengguna sistem adalah koordinator KPTA dan mahasiswa. Koordinator KPTA memiliki hak akses sebagai administrator yang mengelola data yang diperlukan sistem. Sedangkan mahasiswa merupakan entitas yang menggunakan spesifik hanya pada bagian sistem rekomendasi. Gambar IV merupakan DFD Level 0 Sistem Rekomendasi.



Gbr. 4 DFD Level 0 Sistem rekomendasi pembimbing

b. DFD Level 1

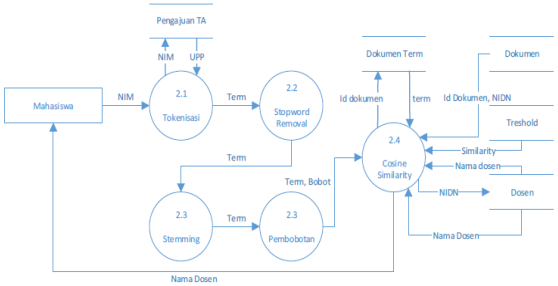
Dari penurunan DFD Level 0 akan dijabarkan lebih lanjut proses-proses yang dimiliki oleh sistem dalam DFD Level 1. Adapun proses yang dimiliki oleh sistem diantaranya login, rekomendasi dosen pembimbing, pengelolaan dokumen penelitian dosen, pengelolaan treshold (ambang batas), dan pengelolaan stoplist yang berisi daftar stopword. Gambar V merupakan DFD Level 1 dari sistem rekomendasi.



Gbr. 5 DFD Level 1 Sistem rekomendasi pembimbing

c. DFD Level 2 Proses Rekomendasi

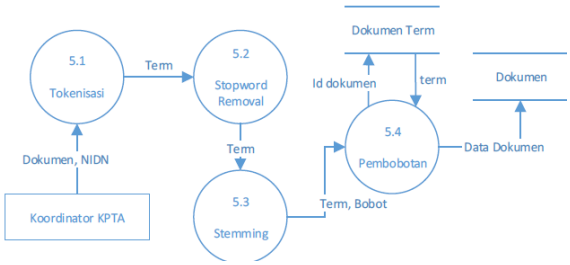
UPP mahasiswa diambil dari *data store* pengajuan TA yang dimiliki oleh Sintesys. Ektraksi dokumen UPP melalui proses tokenisasi, *stopword removal*, *stemming*, dan pembobotan. Hasil ekstraksi berupa term yang memiliki bobot akan dibandingkan dengan dokumen penelitian dosen menggunkan pendekatan cosine similarity. Berbeda dengan dokumen penelitian dosen, hasil ekstraksi UPP tidak dimasukkan ke dalam basis data. Keluaran dari proses ini merupakan rekomendasi dosen pembimbing. Gambar VI merupakan DFD Level 2 dari proses rekomendasi.



Gbr. 6 DFD Level 2 proses rekomendasi

d. DFD Level 2 Proses Ekstraksi

Proses ekstraksi dokumen penelitian dosen pembimbing seperti ditunjukkan pada gambar VII memiliki proses yang sedikit berbeda dibandingkan proses rekomendasi. Pada proses ekstraksi dokumen penelitian dosen tidak dilakukan proses perhitungan similaritas karena pada ektraksi dokumen belum memerlukan perbandingan dokumen.



Gbr. 7 DFD Level 2 proses ekstraksi

2) Struktur Tabel

Integrasi ke dalam Sintesys dilakukan pada basis data yang digunakan. Adapun tabel yang berbeda hanya pada tabel dokumen, dokumen term, treshold, dan stoplist. Dalam penjabaran setiap tabel berikut, beberapa field pada tabel yang saat ini digunakan dalam sintesys tidak dicantumkan untuk menjaga kerahasiaan struktur tabel.

a. Tabel User

Tabel user seperti yang ditunjukkan pada Tabel I merupakan tabel yang digunakan untu menyimpan autentifikasi user berupa username dan password. User hanya terbatas untuk koordinator KPTA.

TABEL I
STRUKTUR TABEL USER

Field	Type
id	int(11) NOT NULL
username	varchar(50) NOT NULL
password	varchar(100) NOT NULL

b. Tabel Dosen

Tabel dosen merupakan tabel yang diambil dari salah satu tabel yang digunakan pada Sintesys. Pada sistem rekomendasi yang dibagun hanya menggunakan nidn dan nama dosen saja. Tabel II merupakan struktur tabel dosen.

TABEL II
STRUKTUR TABEL DOSEN

Field	Type
id	mediumint(9) NOT NULL
dsnid	varchar(20) NULL
nidn	varchar(10) NOT NULL
nmdosen	varchar(50) NOT NULL
id_jurusan	varchar(4) NULL
keahlian1	varchar(50) NULL
keahlian2	varchar(50) NULL
keahlian3	varchar(50) NULL
no_telp	varchar(15) NULL
flag_delete	int(2) NOT NULL

c. Tabel Mahasiswa

Dalam basis data yang digunakan Sintesys, tabel mahasiswa diberi nama du. Tabel du merupakan tabel yang diambil dari sistem informasi akedemik yang dimiliki oleh STMIK STIKOM Indonesia. Tabel III merupakan struktur tabel mahasiswa.

TABEL III
STRUKTUR TABEL MAHASISWA

Field	Type
nim	varchar(10) NOT NULL
kdjur	varchar(10) NULL
nama	varchar(100) NULL
jkel	tinyint(1) NULL
tmplahir	varchar(35) NULL
tgllahir	date NULL
status	varchar(10) NULL
agama	varchar(15) NULL
wargangr	varchar(15) NULL
noid	varchar(30) NULL
jnsid	varchar(10) NULL
Al_asal	varchar(100) NULL
kota	varchar(15) NULL
kdpos	varchar(10) NULL
telp	varchar(15) NULL
Al_tingl	varchar(100) NULL
telp2	varchar(15) NULL
hp	varchar(15) NULL
Al_kant	varchar(80) NULL
telp3	varchar(15) NULL

d. Tabel Dokumen

Pada tabel dokumen disimpan detail dokumen penelitian yang dimiliki oleh dosen STMIK STIKOM Indonesia. Tabel IV merupakan struktur tabel dokumen.

TABEL IV
STRUKTUR TABEL DOKUMEN

Field	Type
id	int(11) NOT NULL
nama	varchar(100) NOT NULL
judul	varchar(200) NOT NULL
abstrak	text NOT NULL
path	varchar(100) NOT NULL
nidn	varchar(11) NOT NULL

e. Tabel Dokumen term

Untuk efisiensi waktu pencarian rekomendasi, maka term beserta term frequency akan disimpan dalam tabel dokumen term. Efisiensi dapat dilakukan karena saat perbandingan similaritas keyword, tidak dibutuhkan lagi ekstraksi term 24 dari dokumen yang dicari similaritasnya. Sistem hanya akan mengambil term serta term frequency tersebut dari tabel dokumen term. Term masing-masing dokumen yang disimpan dalam tabel ini adalah kumpulan term-term yang diekstraksi dari dokumen. Jadi satu record memuat kumpulan term dan term frequency yang merepresentasikan satu dokumen. Rancagan tabel ini ditunjukkan oleh Tabel V.

TABEL V
STRUKTUR TABEL DOKUMEN TERM

Field	Type
id	int(11) NOT NULL
term	text NOT NULL

f. Tabel Stoplist

Tabel stoplist berisi daftar stopword yang digunakan oleh proses keyword extraction system pada tahap stopword removal. Stopword menentukan jumlah term yang dihasilkan oleh keyword extraction system karena stopword digunakan pada proses stopword removal yaitu penghilangan kata yang tidak membawa informasi. Rancangan tabel ini ditunjukkan oleh Tabel VI.

TABEL VI
STRUKTUR TABEL DOKUMEN TERM

Field	Type
stopword	varchar(20) NOT NULL

g. Tabel Treshold

Tabel ambang batas yang rancangannya ditunjukkan oleh Tabel VII, merupakan tabel yang menyimpan ambang batas yang digunakan oleh sistem. Ambang batas ini berupa minimum similarity yang digunakan untuk perbandingan kedekatan keyword antara dokumen serta jumlah rekomendasi untuk menentukan jumlah rekomendasi yang dihasilkan oleh sistem.

TABEL VII
STRUKTUR TABEL TRESHOLD

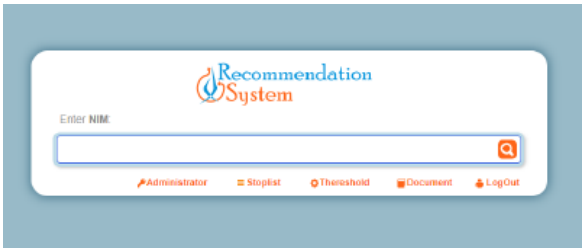
Field	Type
similarity	int(3) NOT NULL
jumlah	int(5) NOT NULL

D. Implementasi Sistem

Pada bagian implementasi, dilakukan proses penerapan hasil rancangan ke dalam perangkat lunak. Berikut adalah hasil implementasi yang telah dilakukan.Antarmuka yang dibutuhkan antara lain antarmuka beranda, rekomendasi dosen pembimbing, administrator, stoplist, pengaturan ambang batas, daftar dokumen penelitian dosen, dan halaman upload dokumen.

1) Beranda Sistem

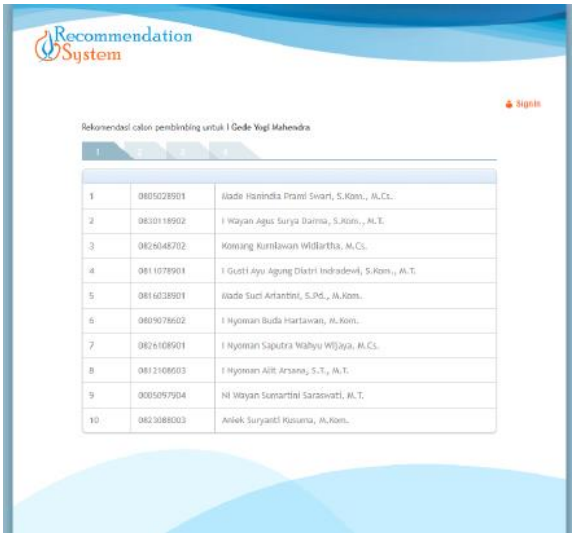
Beranda sistem merupakan halaman pertama kali yang muncul pada saat membuka website sistem rekomendasi. Pada halaman ini mahasiswa hanya perlu memasukkan NIM untuk dapat menelusuri dosen yang memiliki ketertarikan pada bidang penelitian mahasiswa sesuai dengan similaritas UPP yang diajukan. Gambar VIII merupakan halaman beranda sistem.



Gbr 8. Halaman Beranda Sistem

2) Rekomendasi Dosen Pembimbing

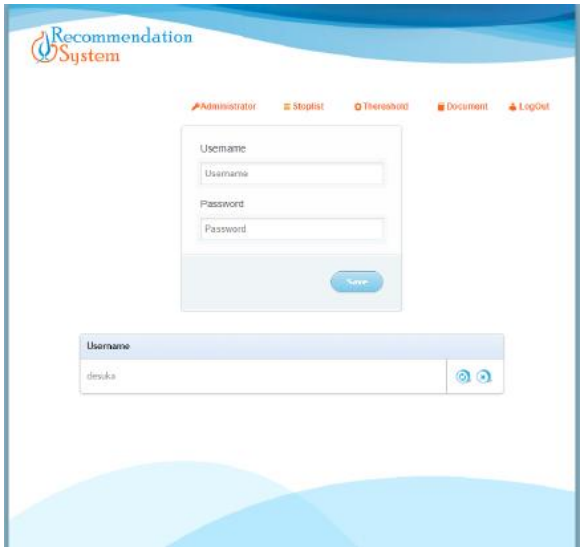
Halaman rekomendasi memiliki daftar rekomendasi dosen yang memiliki nilai kemiripan dokumen penelitian terhadap UPP yang diurutkan berdasarkan nilai similaritas tertinggi ke terendah. Jumlah rekomendasi ditentukan oleh nilai ambang batas yang diberikan. Gambar IX merupakan halaman rekomendasi pembimbing.



Gbr. 9 Halaman rekomendasi dosen pembimbing

3) Administrator

Halaman antarmuka untuk penambahan atau peruahan administrator sistem rekomendasi tidak dipisahkan antara daftar adinistrator dan form administrator. Hal ini dimungkinkan karena administrator sistem hanya koordinator KPTA, sehingga tidak perlu pencarian pada daftar administrator.



Gbr. 10 Halaman administrator

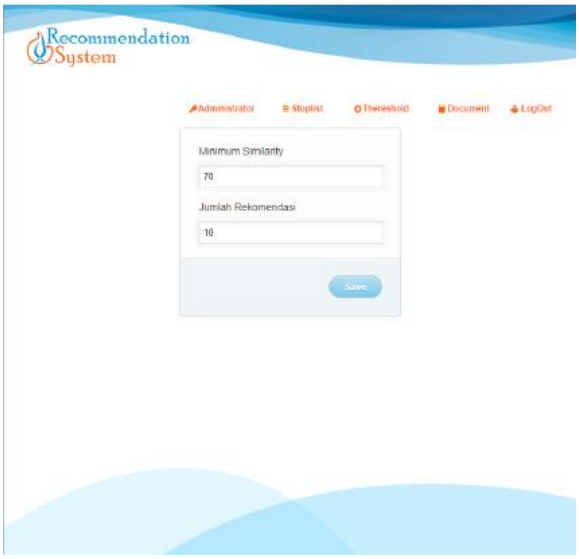
4) Stoplist

Antarmuka stoplist diimplementasikan dalam format tab untuk setiap huruf awalan stopword. Bentuk halaman stoplist dibuat sesederhana mungkin untuk memudahkan dalam penambahan maupun penghapusan stopword. Gambar XI merupakan implementasi halaman stoplist.

Gbr. 11 Halaman stoplist

5) Ambang Batas

Halaman ambang batas hanya berisi 2 field yaitu untuk setting minimum similarity dan jumlah rekomendasi. Gambar XII merupakan implementasi halaman ambang batas.



Gbr. 12 Halaman penentuan ambang batas

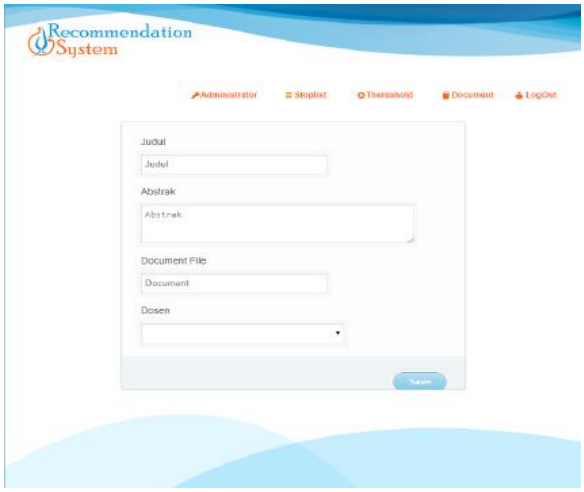
6) Dokumen Penelitian Dosen

Halaman dokumen penelitian hanya berisi daftar dokumen yang telah diunggah ke dalam sistem. Halaman ini juga dilengkapi dengan tombol tambah dokumen dan pencarian dokumen penelitian. Gambar XIII merupakan implementasi halaman dokumen penelitian dosen.

Gbr. 13 Halaman unggah dokumen penelitian dosen

7) Upload Dokumen

Halaman upload dokumen akan muncul apabila tombol upload dokumen pada halaman dokumen penelitian ditekan. Halaman ini berisi form untuk memasukkan detail dari dokumen penelitian yang diunggah. Dokumen penelitian yang diunggah harus dalam format PDF. Proses ekstraksi dilakukan setelah tombol save ditekan. Gambar XIV merupakan implementasi dari halaman upload dokumen.



Gbr. 14 Halaman upload dokumen

E. Pengujian Sistem

Pengujian dilakukan dengan pendekatan black box. Salah satu cara pengujian dalam metode black box yaitu pengujian dilakukan berdasarkan skenario yang telah ditentukan. Pengujian untuk sistem rekomendasi yang dibagun dibagi menjadi beberapa bagian sesuai dengan antarmuka yang diimplementasikan. Tabel VIII merupakan skenario beserta hasil pengujian sesuai dengan skenario.

TABEL VIII
PENGUJIAN SISTEM

Bagian	Skenario	Hasil yang Diharapkan	Hasil
Login	Memasukkan *username* dan *password* benar	Sistem masuk ke beranda *admin*	Berhasil
	Memasukkan username atau password salah	Sistem menampilkan pesan username dan password salah	Berhasil
Administrator	Memasukkan admin baru	Sistem menambahkan admin baru	Berhasil
	Menghapus admin yang sudah ada	Sistem menghapus data admin	Berhasil
	Mengubah data admin	Sistem memperbaharui data berdasarkan data yang dimasukkan	Berhasil
Stoplist	Menambahkan Stopword "apabila"	Sistem menambahkan stopword yang dimasukkan oleh admin	Berhasil
	Menghapus stopword "apabila"	Sistem menghapus data *stopword*	Berhasil
Treshold	Mengisi data similarity atau jumlah rekomendasi	Sistem memperbaharui *minimum similarty* dan jumlah rekomendasi	Berhasil
Dokumen	Menampilkan data dokumen penelitian	Sistem membuat list perhalaman untuk semua dokumen penelitian dengan kelengkapan nama dosen dan judul dokumen	Berhasil
	Meghapus data dokumen penelitian	Sistem mengupdate data penelitian dosen	Berhasil
	Mengubah detail data penelitian dosen	Masuk ke halaman upload dokumen penelitian	Berhasil
Upload Dokumen	Mengisi detail dokumen dan menekan tombol proses	Sistem memperbaharui atau menambahkan data penelitian dosen dan melakukan ekstraksi dokumen PDF yang dimasukkan	Berhasil

V. PENUTUP

A. Kesimpulan

1. Sistem dibangun dengan membandingkan hasil ekstraksi dari UPP dan dokumen penelitian dosen. Proses yang dilakukan dalam ektraksi teks yaitu tokenisasi, stopword removal, stemming, dan pembobotan. Hasil ektraksi lalu dibandingkan dengan menggunakan pendekatan *cosine similarity*. Semakin besar nilai cosine-similarity yang dihasilkan, maka semakin mirip kedua dokumen tersebut, sehingga rekomendasi pembimbing akan didasarkan pada nilai cosine-similarity terkecil antara ekstraksi dokumen UPP dan penelitian dosen.
2. Sistem yang dibangun telah memenuhi kebutuhan fungsionalitas yang menjawab hasil dari analisis permasalahan yang ditentukan pada awal penelitian. Hal tersebut sesuai dengan hasil pengujian fungsionalitas sistem yang dilakukan menggunakan blackbox testing. Sistem telah sesuai dengan perancangan

B. Saran

Penelitian yang telah diselesaikan ini membuka beberapa penyempurnaan untuk menjadikan sistem rekomendasi ini menjadi lebih baik lagi. Penelitian lanjutan yang dapat dilakukan diantaranya :

1. Sistem rekomendasi ini belum diuji dari sisi akurasi rekomendasi yang dihasilkan, sehingga selanjutnya dapat dilakukan pengujian akurasi dan performa Sistem.
2. Berdasarkan hasil implementasi sistem, ditemukan bahwa stemming memakan waktu yang banyak, hal ini tentunya menurukan efisiensi kinerja sistem. Untuk itu diperlukan penelitian untuk melakukan perbandingan sistem dengan menggunakan stemming dan tidak menggunakan stemming.
3. Sistem rekomendasi yang dibuat saat ini masih berupa stand-alone system, dimana data Usulan Proposal Penelitian didapatkan dari sistem lain yaitu Sintesys. Kedepannya akan sangat optimal jika sistem ini dapat terintegrasi dengan Sintesys.

REFERENSI

[1] Axler, S., 1997, *Linear Algebra Done Right*, Springer, New York.
[2] Debole, F. dan Sebastiani, F., 2003, Supervised Term Weighting for Automated Text Categorization, *18th ACM Symposium on Applied Computing*, 784–788.
[3] Feldman, R. dan Sanger, J., 2007, *The Text Mining Handbook*, Cambridge University Press, Cambridge.
[4] Goldberg, D., Nichols, D., Oki, B. M., dan Terry, D., 1992, Using collaborative filtering to weave an information tapestry, Commun. ACM 35, 12, 61-70.D. Sudoku, G. T. Jawasand G.K. Tambau ,"Implementation of a Direct Access Files",*IEEE Transactions on Information*, vol. 4, no 104 , pp.70-77, 2008.
[5] Han, J., Kamber, M., dan Pei, J.,2011, *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publisher, San Francisco.
[6] Hulth, A., 2003, Improved Automatic Keyword Extraction Given More Linguistic Knowledge, *Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing*, Sapporo, Japan, 216-223.
[7] Lops, P., Gemmis, M. d., dan Semeraro, G., 2011, Content-based Recommender Systems: State of the Art and Trends,

Recommender Systems Handbook, Springer, New York, 73-105.

[8] Lovins, J. B., 1968, Development of a Stemming Algorithm, *Mechanical Translation and Computational Linguistics*, 11, Massachusetts Institute of Technology, Cambridge, Massachusetts, Maret dan Juni 1968, 22-31.

[9] Manning, C. D., Raghavan, P., dan Schütze, H., 2009, *An Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England.

[10] McGinty, L. dan Smyth, B., 2006, Adaptive Selection : An Analysis of Critiquing and Preference-Based Feedback in Conversational Recommender Systems, *International Journal of Electronic Commerce*, 11, 35–57.

[11] Oelze, I., 2009, Automatic Keyword Extraction for Database Search, *Ph.D. Thesis*, University of Hannover, Hannover.

[12] Resnick, P., Iakovou, N., Sushak, M., Bergstrom, P., and Riedl J., 1994, GroupLens: An Open Architecture for Collaborative Filtering of Netnews, Proc. 1994 Computer Supported Cooperative Work Conf.N. *Mohan* and T. M. Undeland,Power Electronics,2 ed., New York: John Wiley & Sons, 2015, p. 11, 2005.

[13] Srividhya, V. dan Anitha, R., 2010, Evaluating Preprocessing Techniques in Text Categorization, *International Journal of Computer Science and Application*,Issue 2010, 49-51.

[14] Tala, F.Z. 2003. *A Study of Stemming Efects on Information Retrieval in Bahasa Indonesia*. Thesis. Institute for Logic Language and Computation Universiteit van Amsterdam TheNetherlands.

[15] Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F. 2005. *Text Mining: Predictive Methods for Analyzing Unstructured Information.* New York: Springer.

[16] *Zobel, J. dan Moffat, A., 1998, Exploring the similarity space, ACM SIGIR Forum, 32, 18-34*