PERANCANGAN APLIKASI PENCARIAN KATA YANG BERKAITAN SECARA SEMANTIK MENGGUNAKAN TEORI *MUTUAL INFORMATION*

Mira Ziveria[1)]
[1]Program Studi Sistem Informasi, Fakultas Ilmu Komputer dan Ilmu Komunikasi,
Institut Teknologi dan Bisnis Kalbis
Jl. Pulomas Selatan Kav. 22, Jakarta Timur 13210
HP (penulis utama): +62 81 314 013 416
E-mail : ziveria_mr@yahoo.com

Abstrak

Penelitian ini membahas tentang bagaimana merancang aplikasi yang mengenali kata yang berkaitan secara semantik dalam Bahasa Indonesia dan bagaimana teori statistik mutual information dapat digunakan untuk menentukan seberapa besar keterkaitan dua buah kata. Aplikasi ini dirancang agar bisa menerima masukan sejumlah artikel berbahasa Indonesia dan mampu menghitung nilai mutual information setiap pasangan kata yang berbeda dari artikel tersebut sehingga menghasilkan basis data yang berisi pasangan kata beserta nilai mutual informationnya. Tahap analisis dan perancangan menggunakan metode analisis dan desain terstruktur yaitu berorientasi aliran data yang menghasilkan diagram aliran data. Hasil dari penelitian ini adalah bahwa teori mutual information dapat digunakan untuk mengestimasi keterkaitan kata secara semantik, yaitu apabila dua buah kata sering muncul bersamaan dalam artikel maka nilai mutual informationnya akan semakin tinggi. Rancangan aplikasi yang telah dibuat dapat membangun basis data yang berisi kata-kata yang berkaitan dalam bahasa Indonesia secara otomatis dari kumpulan artikel.

Kata kunci: *aplikasi, semantik, mutual information.*

This research discusses how to design applications that identify semantic related words in Indonesian languages and how the statistical theory of mutual information can be used to determine the association of two words. This application is designed to receive a number of articles in Indonesian language as input and to have ability to calculate the value of the mutual information of each pair of different words in the articles so as to produce a database that contains the word pairs with the value of mutual information. Phase of analysis and design using structured analysis and design methods that are oriented in data flo, so that generates the data flow diagram. Results from this research is the mutual information theory can be used to estimate the semantic word relevance, i.e if two words frequently appear simultaneously in the value of the mutual informationnya article will be higher. The design of applications that have been made to build databases that contain related words in Indonesian language automatically from a collection of articles.

Keywords*: application, semantics, mutual information.*

1. PENDAHULUAN

Bagian ini berisi latar belakang, rumusan masalah, tujuan, batasan masalah, dan metode penelitian.

1.1 Latar Belakang

Seiring perkembangan teknologi informatika, produktifitas bahasa yang dikomputerisasi terus meningkat, diantaranya asosiasi kata. Keterkaitan semantik antara satu kata dengan kata lainnya yang disusun dengan aturan tertentu lebih dikenal dengan *thesaurus. Thesaurus* biasanya digunakan pada mesin penterjemah, sistem temu kembali informasi, sistem pengolah kata, dan sebagainya. Aplikasi yang mengenali kata-kata yang berkaitan dengan kata tertentu baru dimiliki oleh pengolah kata tertentu, contohnya *MS-Word* dilengkapi dengan aplikasi yang bernama *thesaurus,* tetapi hanya untuk kata dalam bahasa Inggris, Perancis, dan Spanyol. Sementara untuk bahasa Indonesia, fasilitas tersebut belum dapat digunakan. Untuk membangun aplikasi *thesaurus* bahasa Indonesia secara manual akan membutuhkan biaya dan usaha yang tinggi. Oleh karena itu penelitian ini mencoba

untuk merancang pembangunan aplikasi secara otomatis yang mengenali kata-kata yang berkaitan secara semantik dari kumpulan artikel menggunakan teori *mutual information*.

1.2 Rumusan Masalah

Permasalahan yang dibahas pada penelitian ini adalah apakah teori *mutual information* dapat mengestimasi dengan tepat keterkaitan kata secara semantik dan bagaimana merancang aplikasi yang secara otomatis dapat mengenali kata-kata yang berkaitan secara semantik dalam bahasa Indonesia.

1.3 Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Mempelajari konsep *mutual information* dan menggunakan konsep tersebut untuk mencari keterkaitan semantik antara satu kata dengan yang lainnya.
2. Membuat perancangan perangkat lunak yang dapat membentuk suatu aplikasi yang mengenali kata-kata yang berkaitan secara otomatis dari kumpulan artikel dengan menggunakan metode *mutual information*.

1.4 Batasan Masalah

Untuk memperjelas apa yang akan dibahas pada penelitian ini, maka diperlukan batasan sebagai berikut:

1. Perancangan aplikasi yang dihasilkan tidak berbentuk hirarki, tetapi hanya mengenal kata-kata yang berkaitan secara semantik dari suatu kata, seperti presiden, tugas, negara, kekuasaan, dan lain-lain.
2. Rancangan aplikasi hanya untuk bahasa Indonesia saja.

1.5 Metode Penelitian

Penelitian dilakukan dengan cara:

1. Melakukan studi literatur tentang semantik dalam bahasa Indonesia, *thesaurus*, dan *mutual information.*
2. Melakukan analisis dan perancangan perangkat lunak menggunakan metode System Development Life Cycle (SDLC). Analisis, yaitu mendefinisikan spesifikasi kebutuhan, fungsi, arsitektur sistem, masukan dan keluaran, serta model fungsional yang dimulai dari diagram konteks sampai DFD level 3. Perancangan, yaitu batasan rancangan, perancangan struktur data, basis data, struktur menu, dan layar tampilan dari perangkat lunak.

2. TINJAUAN PUSTAKA

Bagian ini menguraikan konsep semantik dalam bahasa Indonesia, *thesaurus*, teori *mutual information* untuk menentukan keterkaitan kata secara semantik dalam bahasa Indonesia.

2.1 Semantik dalam Bahasa Indonesia

Semantik adalah bidang studi dalam linguistik yang mempelajari makna atau arti dalam bahasa. Dalam setiap bahasa, seringkali ditemui adanya hubungan makna atau relasi semantik antara sebuah kata dengan kata lainnya. Relasi makna ini mungkin menyangkut kesamaan makna atau sinonim, kebalikan makna atau antonim, kegandaan makna atau polisemi, ketercakupan makna atau hiponim, kelainan makna atau homonim, dan sebagainya. [5]

2.2 *Thesaurus*

Thesaurus adalah sekumpulan pernyataan yang disusun dengan aturan tertentu yang saling berkaitan secara hierarki, asosiasi, atau hubungan kesepadanan. *Thesaurus* biasanya digunakan sebagai salah satu komponen untuk membangun mesin penterjemah, sistem pengolah kata, dan sebagainya. Konsep pembangunan *thesaurus* berdasarkan pada hubungan hierarki, hubungan kesamaan, dan bisa juga hubungan asosiatif kata. [3]

2.*3 Mutual Information*

Sejumlah tool dan teknik yang telah diterapkan untuk pemrosesan bahasa antara lain: entropy, perplexity, mutual information, traditional statistics, connectionism, genetic algorithms, formal language theory, Markov Modelling dan non-linear dinamical modelling. Diantara teknik-teknik tersebut terdapat hubungan secara matematika. [4]

Mutual information adalah apabila dua buah pengamatan *x* dan *y* yang saling bebas memiliki peluang *P(x)* dan *P(y),* sedangkan *P(x,y)* adalah peluang pengamatan *x* dan *y* secara bersama-sama, maka *mutual information I(x,y)*:

$$I(x,y) = \log \frac{P(x,y)}{P(x).P(y)}$$

Secara formal, *mutual information* membandingkan peluang pengamatan *x* dan *y* secara bersama dengan peluang pengamatan *x* dan *y* secara bebas. Jika terdapat hubungan yang kuat antara pengamatan *x* dan *y*, maka *P(x,y)* akan lebih besar dari *P(x).P(y)* dan sebagai akibatnya *I(x,y) >> 0*. Apabila tidak ada hubungan keterikatan antara pengamatan *x* dan *y* maka $P(x,y) \approx P(x).P(y)$ dan $I(x,y) \approx 0$. [2]

Dalam aplikasinya, peluang kata P(x) dan P(y) diperkirakan dengan menghitung jumlah kata x dan jumlah kata y yang terdapat pada suatu artikel, yaitu f(x) dan f(y), lalu dinormalkan dengan N, yaitu jumlah seluruh artikel. Peluang bersama P(x,y) diperkirakan dengan menghitung berapa kali kata x yang diikuti oleh kata y yang dinotasikan dengan $f_w(x,y)$ dalam suatu artikel dengan jumlah w kata-kata, lalu dinomalkan dengan N.

$$I(x,y) = \log \frac{f_w(x,y)/N}{(f(x)/N)(f(y)/N)}$$

$f_w(x,y)$ melambangkan jumlah kata x berada sebelum kata y dalam suatu artikel dengan w kata-kata, bukanlah jumlah berapa kali dua buah kata berbeda berada dalam sembarang urutan. Jadi I(x,y) tidak harus sama dengan I(y,x). Parameter ukuran artikel mengakibatkan adanya perbedaan skala. Artikel yang lebih kecil mengakibatkan sulitnya mengambil kesimpulan yang benar. Ukuran artikel yang lebih besar akan memperjelas konsep semantik dan kesimpulan yang dihasilkan akan lebih baik. [2]

Sebagai contoh pengukuran *mutual information*, misalkan terdapat sejumlah 10 juta artikel (N = 10.000.000) akan ditentukan kuatnya keterikatan kata dokter dengan kata perawat yaitu dengan menghitung nilai mutual information antara kata dokter dengan kata perawat atau I(dokter,perawat). Dari 10 juta artikel tersebut, terdapat 25 buah artikel yang memuat kata dokter dengan kata perawat secara bersama-sama (f(dokter,perawat) = 25) dan terdapat 5 buah artikel yang hanya memuat kata dokter saja tanpa kata perawat (f(dokter) = 5) dan 12 buah artikel yang hanya memuat kata perawat saja tanpa kata dokter (f(perawat) = 12), maka

$$I(dokter, perawat) = \log \frac{f(dokter, perawat)/N}{(f(dokter)/N).(f(perawat)/N)}$$

$$= \log \frac{25/10000000}{(5/10000000).(12/10000000)}$$

$$= 6{,}62$$

Jadi nilai *mutual information* kata dokter dengan kata perawat adalah 6,62.

3. ANALISA KEBUTUHAN PERANGKAT LUNAK

Pada bagian ini akan dijelaskan mengenai analisis kebutuhan perangkat lunak untuk mencari kata-kata yang berkaitan dalam bahasa Indonesia secara otomatis. Perangkat lunak yang dirancang diberi nama Thesindo (Thesaurus Indonesia Otomatis). Aspek analisis ini meliputi spesifikasi kebutuhan, spesifikasi fungsi, arsitektur sistem, serta model fungsional perangkat lunak.

3.1 Spesifikasi Kebutuhan Perangkat Lunak

Rancangan aplikasi Thesindo bertujuan untuk menyediakan pengolahan bahasa Indonesia khususnya dalam pencarian kata-kata yang berkaitan secara semantik dengan otomatis. Paket aplikasi tersebut mampu melakukan penghitungan nilai *mutual information* sebagai petunjuk dalam menentukan kata yang berkaitan dalam bahasa Indonesia dan mudah untuk digunakan.

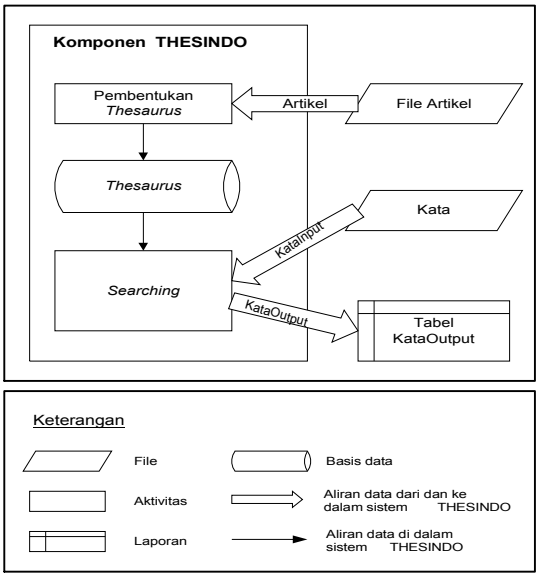
3.2 Spesifikasi Fungsi Perangkat Lunak

Aplikasi Thesindo dirancang agar memiliki kemampuan sebagai berikut:

1. Dapat menerima masukan berupa artikel berbahasa Indonesia yang disimpan dalam file teks.
2. Dapat menerima masukan berupa teks yaitu sebuah kata dari papan kunci.
3. Dapat melakukan proses perhitungan jumlah kemunculan kata dan *mutual information* terhadap kata-kata dalam artikel yang telah disimpan dalam file.
4. Dapat menampilkan keluaran berupa kata-kata yang berkaitan dengan kata yang dimasukkan oleh pengguna.

3.3 Arsitektur Sistem

Arsitektur sebuah perangkat lunak memberikan gambaran mengenai komponen-komponen yang terdapat pada sebuah perangkat lunak. Secara global arsitektur perangkat lunak dapat dilihat pada Gambar berikut:



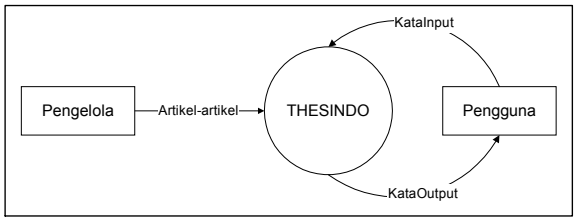
Gambar 1. Arsitektur Sistem

Aplikasi Thesindo terdiri dari dua komponen utama, yaitu komponen pembentukan thesaurus dan komponen searching. Komponen pembentukan thesaurus merupakan komponen yang berfungsi untuk mempersiapkan artikel yang akan menjadi bahan baku bagi pengolahan Thesindo. Komponen ini berhubungan dengan eksternal entitas melalui file yang dimasukkan oleh pengelola. File eksternal yang diterima, yaitu file artikel yang akan diolah oleh komponen pembentukan thesaurus sehingga menghasilkan thesaurus untuk selanjutnya dipergunakan oleh komponen lainnya. Komponen pembentukan thesaurus akan membentuk sebuah basis data yang bernama Thesaurus. Basis data ini menampung data yang siap diolah oleh komponen searching. Pada komponen ini diakukan pengolahan terhadap artikel sampai kepada penghitungan mutual information.

Komponen searching yaitu komponen yang melakukan proses pencarian kata yang dimasukkan oleh pengguna. Proses pencarian kata pada komponen ini sangat bergantung kepada KataInput yaitu kata yang dimasukkan oleh pengguna dari pengolahan Thesindo. Pengguna memasukkan KataInput setelah thesaurus dibentuk sebelumnya oleh komponen pembentukan thesaurus. Kata yang dimasukkan pengguna dicari pada basis data Thesaurus. Komponen searching ini memberikan laporan kepada pengguna yaitu keluaran berupa KataOutput. Laporan yang dihasilkan oleh komponen ini adalah laporan berupa tabel kata-kata yang berkaitan secara semantik dengan KataInput yang dimasukkan oleh pengguna.

3.4 Model Fungsional

Model fungsional Thesindo digambarkan dalam diagram aliran data (*Data Flow Diagran atau DFD*). DFD dimulai dengan gambaran global dari Thesindo dalam bentuk diagram konteks (DFD level 0). Sedangkan DFD level 1 dan selanjutnya menerangkan proses mulai dari yang bersifat global sampai pada proses yang lebih rinci.

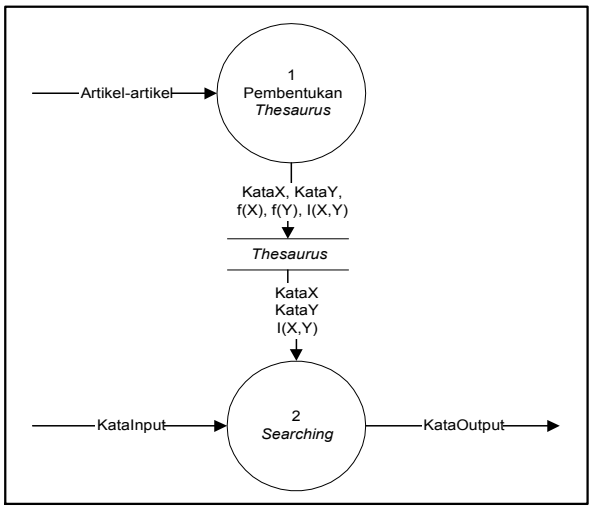


Gambar 2. Diagram Konteks

3.4.1 Diagram Konteks

Thesindo menerima masukan berupa sejumlah artikel dari pengelola dan sebuah kata bahasa Indonesia dari pengguna. Pengguna memasukkan sebuah kata setelah terbentuk thesaurus yang dibentuk berdasarkan artikel-artikel yang dimasukkan oleh pengelola. Sejumlah artikel tersebut terdapat dalam sebuah file. Sedangkan kata dimasukkan oleh pengguna melalui suatu form dialog pada saat eksekusi. Masukan-masukan yang ada akan diproses oleh sistem THESINDO untuk kemudian memberikan keluaran berupa laporan kata kepada pengguna.

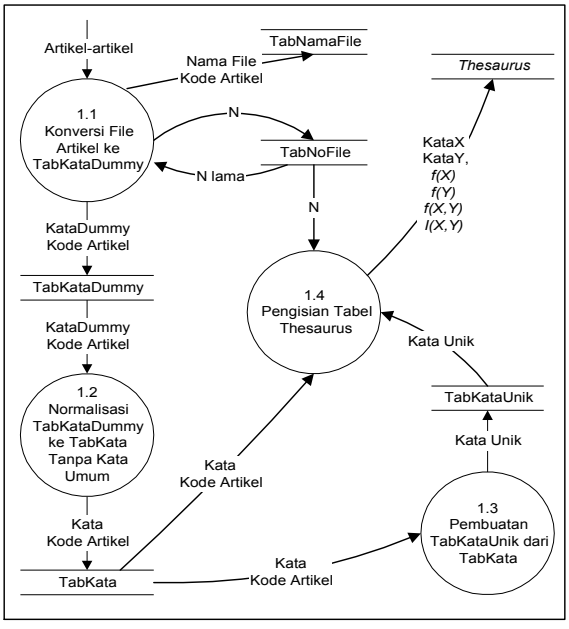
3.4.2 DFD Level 1



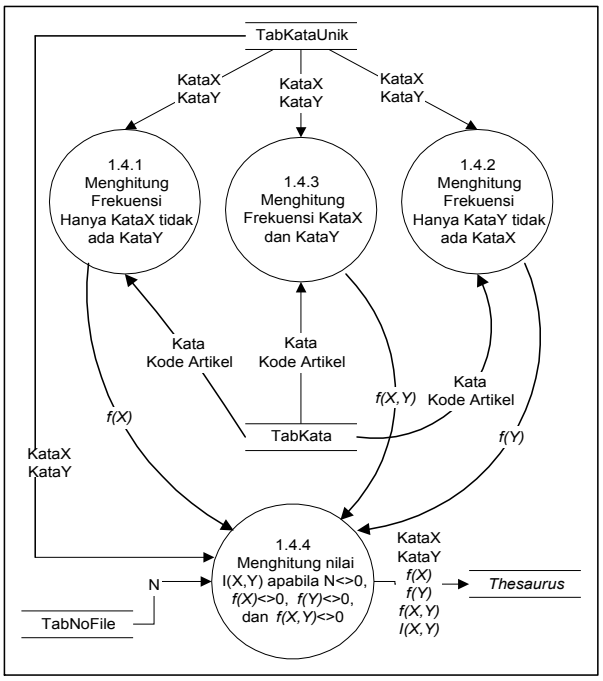
Gambar 3. DFD Level 1

Proses 1 yaitu Pembentukan Thesaurus melakukan proses pengolahan terhadap masukan yang berupa File Artikel dan menghasilkan thesaurus. Pada proses ini dilakukan pengolahan artikel dan melakukan perhitungan mutual information dan hasil dari proses ini akan disimpan pada basis data Thesaurus yang akan dimanfaatkan oleh proses searching. Proses 2 yaitu Searching membaca kata yang dimasukkan oleh pengguna dan menghasilkan kata-kata yang berkaitan secara semantik dengan kata tersebut. Proses pencarian kata-kata yang berkaitan secara semantik ini dilakukan setelah terbentuknya thesaurus yang dihasilkan oleh proses lainnya.

3.4.3 DFD Level 2

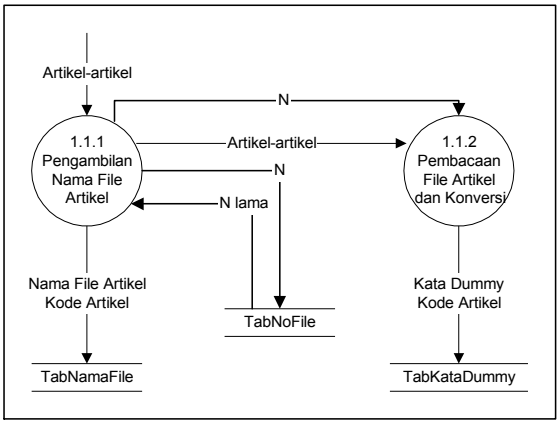


Gambar 4. DFD Level 2 Proses Pembentukan Thesaurus

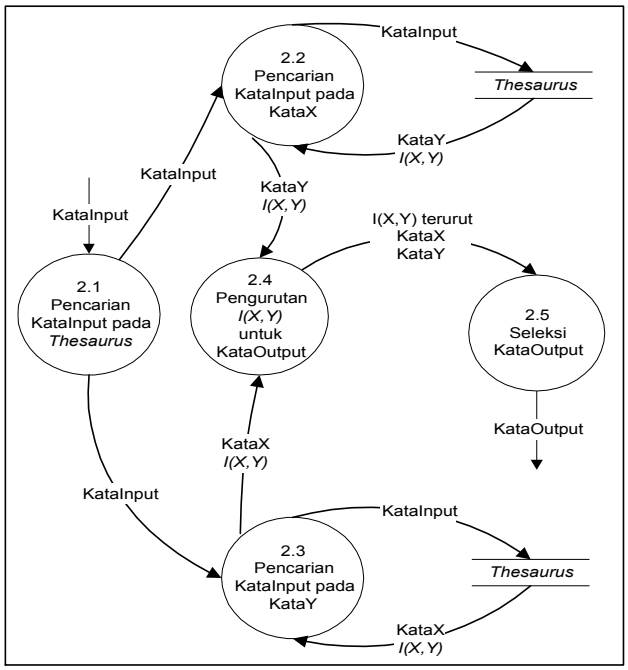


Gambar 5. DFD Level 2 Proses Proses Searching

3.4.4 DFD Level 3



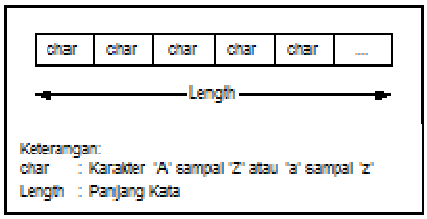
Gambar 6. DFD Level 3 Proses Konversi File Artikel ke TabKataDummy



Gambar 7. DFD Level 3 Proses Pengisian Tabel Thesaurus

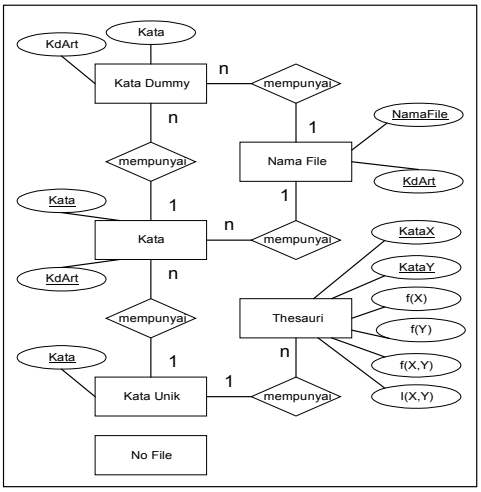
3.5 Perancangan Struktur Data

3.5.1 Perancangan Struktur Logika Data

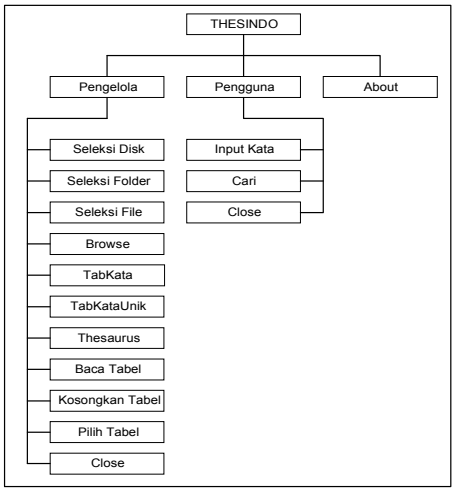


Gambar 8. Struktur lojik data

3.5.2 Perancangan Basis Data dan Struktur Menu



Gambar 9. Diagram ER

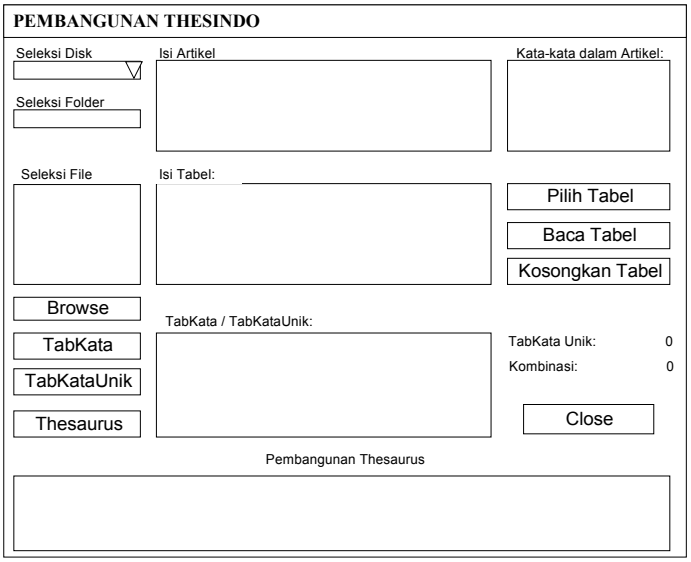


Gambar 10. Struktur Menu

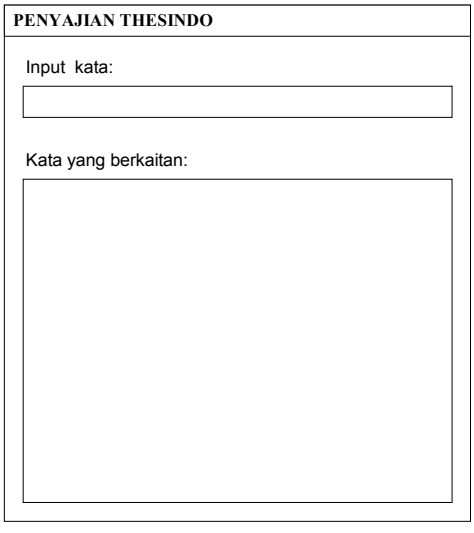
Susunan menu Thesindo terdiri dari menu utama dan beberapa sub menu. Menu utama tersebut antara lain:

1. Pengelola
   1.1. Seleksi Disk, yaitu menu untuk memilih sebuah disk (drive) yang menyimpan satu atau beberapa folder yang menyimpan file artikel yang akan digunakan.
   1.2. Seleksi Folder, yaitu menu untuk membuka dan menampilkan folder yang menyimpan satu atau beberapa file artikel yang akan digunakan.
   1.3. Seleksi File, menu ini menampilkan nama-nama file artikel, untuk dipilih satu atau beberapa nama file artikel yang akan digunakan.
   1.4. Browse, menu ini untuk melakukan pembacaan untuk satu atau beberapa artikel yang akan dikonversi menjadi kata-kata, yang hasil dari proses ini disimpan pada TabKataDummy. Nama file artikel disimpan pada TabNamaFile, dan jumlah artikel disimpan pada TabNoFile.
   1.5. TabKata, menu ini melakukan proses peringkasan dari TabKataDummy ke TabKata, dimana apabila pada TabKataDummy terdapat kata-kata yang sama dengan Kode Artikel yang sama, maka pada TabKata hanya akan disimpan jadi satu kata saja beserta Kode Artikelnya. Disini juga dilakukan penyaringan kata pada TabKataDummy dengan Kata Umum. Jadi apabila pada TabKataDummy mengandung Kata Umum maka pada TabKata tidak disimpan.
   1.6. TabKataUnik, menu ini melakukan proses peringkasan dari TabKata ke TabKataUnik, dimana apabila pada TabKata terdapat kata-kata yang sama, maka pada TabKataUnik hanya akan disimpan jadi satu kata saja.
   1.7. Thesaurus, menu ini melakukan pengisian tabel Thesaurus yang berisi KataX, KataY, f(X), f(Y), f(X,Y), dan I(X,Y).
   1.8. Baca Tabel, yaitu menu untuk membaca atau menampilkan tabel setelah sebelumnya ditentukan nama tabel yang akan dibaca atau ditampilkan pada menu Pilih Tabel.
   1.9. Kosongkan Tabel, menu untuk menghapus atau mengosongkan tabel setelah sebelumnya ditentukan nama tabel yang akan dihapus atau dikosongkan pada menu Pilih Tabel.
   1.10. Pilih Tabel, menu ini dibuat untuk melakukan pemilihan tabel mana yang akan dibaca atau dikosongkan. Jenis tabel yang dipilih yaitu: TabKataDummy, TabKata, TabKataUnik, Thesaurus, TabNoFile.
   1.11. Close, menu ini berfungsi keluar dari menu Pembangunan Thesindo.
2. Pengguna
   2.1. Input Kata, Berfungsi untuk memasukkan kata dalam bahasa Indonesia kedalam sistem THESINDO.
   2.2. Cari, berfungsi untuk memproses kata yang dimasukkan oleh pengguna, sehingga diperoleh hasil berupa tabel kata-kata yang berkaitan secara semantik dengan kata yang dimasukkan oleh pengguna.
   2.3. Close, berfungsi keluar dari menu Pengguna perangkat lunak THESINDO.
3. About, menu ini merupakan tampilan identitas dari pengembang sistem perangkat lunak THESINDO.

3.6 Perancangan Layar untuk Pembangunan Thesindo (Pengelola) dan Layar Penyajian



Gambar 11. Layar Pengelola



Gambar 12. Layar Penyajian

4. SIMPULAN DAN SARAN

Bagian ini berisi kesimpulan yang diperoleh dari penelitian dan juga saran untuk pengembangan penelitian selanjutnya.

4.1 Simpulan

1. Teori *mutual information* dapat digunakan untuk mengestimasi keterkaitan kata secara semantik.
2. Dengan metode SDLC dapat dibuat perancangan aplikasi yang mengenali kata-kata bahasa Indonesia yang berkaitan secara semantik secara otomatis dari kumpulan artikel menggunakan teori *mutual information*.

4.2 Saran

1. Rancangan aplikasi yang telah dibuat sebaiknya diimplementasikan dan diuji dalam lingkungan operasi tertentu.
2. Artikel yang dipakai sebaiknya mencakup seluruh bidang agar hasil yang didapatkan mencakup semua kata bahasa Indonesia.

5. DAFTAR RUJUKAN

[1] Emmert, Frank and Dehmer, Matthias. 2009. *Information Theory and Statistical Learning*. Springer. Seattle USA.

[2] Nordstrom, Bengt and Ranta, Aarne. 2008. *Advances in Natural Language Processing: 6th International Conference, GoTAL*. Springer. Gothenburg, Sweden.

[3] Roe, Sandra K and Thomas, Alan R. 2010. *The Thesaurus: Review, Renaissance, and Revision*. The Haworth Information Press. Binghamton, New York.

[4] Wei, Li and Gruyte, Mouton De. 2011. *Applied Linguistics Review 20112, Volume 2*. Walter de Gruyter GmbH & Co. KG. New York.

[5] Wijana, I Dewa Putu dan Rohmadi, Muhammad. 2011. *Semantik: Teori dan Analisis*. Yuma Pustaka. Surakarta, Indonesia.