**PROPOSAL**

**TUGAS AKHIR INFORMATION RETRIEVEL**

**RANCANG BANGUN FITUR DEKTEKSI PERBEDAAN GAYA TULIS DALAM BEBERAPA DOKUMEN DENGAN BOOLEAN RETRIVAL MODEL**

Diajukan untuk memenuhi syarat mata kuliah Sistem Temu Kembali Informasi



Disusun oleh :

**RICKY TRISNA A NIM. A11.2017.10340**

**MOH IQBAL . NIM. A11.2017.10343**

**A11.4716**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS DIAN NUSWANTORO**

**SEMARANG**

**2019**

**DAFTAR ISI**

|  |  |  |  |
| --- | --- | --- | --- |
| [DAFTAR ISI](#page2) | | [...........................................................................................................](#page2) | [ii](#page2) |
| [BAB I ......................................................................................................................](#page3) | |  | [1](#page3) |
| [PENDAHULUAN...................................................................................................](#page3) | | | [1](#page3) |
| [1.1](#page3) | [Latar Belakang..........................................................................................](#page3) | | [1](#page3) |
| [1.2](#page4) | [Rumusan Masalah ....................................................................................](#page4) | | [2](#page4) |
| [1.3](#page4) | [Batasan Masalah Penelitian ......................................................................](#page4) | | [2](#page4) |
| [1.4](#page4) | [Tujuan Penelitian ......................................................................................](#page4) | | [2](#page4) |
| [1.5](#page5) | [Manfaat Penelitian ....................................................................................](#page5) | | [3](#page5) |
| [BAB II .....................................................................................................................](#page6) | |  | [4](#page6) |
| [LANDASAN TEORI ..............................................................................................](#page6) | | | [4](#page6) |
| [2.1](#page6) | [Tinjauan Studi ..........................................................................................](#page6) | | [4](#page6) |
| [2.2](#page11) | [Tinjauan Pustaka ......................................................................................](#page11) | | [9](#page11) |
| [2.3](#page13) | [Kerangka Pemikiran ...............................................................................](#page13) | | [11](#page13) |
| [BAB III..................................................................................................................](#page15) | |  | [13](#page15) |
| [METODE PENELITIAN ......................................................................................](#page15) | | | [13](#page15) |
| [3.1](#page15) | [Instrumen Penelitian ...............................................................................](#page15) | | [13](#page15) |
| [3.1.1](#page15) | | [Kebutuhan Perangkat Lunak (Software) .........................................](#page15) | [13](#page15) |
| [3.1.2](#page15) | | [Kebutuhan Perangkat Lunak (Software) .........................................](#page15) | [13](#page15) |
| [3.2](#page15) | [Prosedur Pengambilan dan Pengumpulan Data ......................................](#page15) | | [13](#page15) |
| [3.3](#page16) | [Teknik Analisis Data ..............................................................................](#page16) | | [14](#page16) |
| [3.4](#page16) | [Model/Metode yang Diusulkan ..............................................................](#page16) | | [14](#page16) |

ii

**BAB I**

**PENDAHULUAN**

**1.1 Latar Belakang**

Dewasa ini dimana internet marak digunakan mendapatkan informasi sangat mudah didapatkan, marak pula plagiarism dimana tulisan orang lain di pamerkan seakan tulisan sendiri, yang sedihnya adalah kita tidak tau itu memang tulisan hasil plagiat atau tidak, kadang ide mempunyai dasar sama dengan yang ada di internet tetapi bukan hasil plagiat, kadang juga hanya mengambil sebagian dari tulisan orang lain.

Dalam sebuah sistem diamana seorang penulis diharuskan menulis beberapa tulisan maka kadang penlis tersebut mengambil jalan pintas dengan meniru beberapa kata dalam sebuah karya milik orang lain, kadang itu bukan merupakan hal yang salah tetapi tulisan yang ditiru belum tentu menggunakan gaya bahasa yang sama.

Sebagai pengawas penulis hal tersebut merupakan hal yang perlu diperhatikan, maka pengawas harus dapat memonitor penulis agar tidak menulis yang bukan merupakan hasil pikiran sendiri, tetapi kesulitan muncul saat jumlah penulis yang diawasi oleh seorang pengawas melebihi kemampuan pengawas.

1

2

Solusi yang diusulkan adalah menerapkan Sistem Temu Kembali Informasi dimana setiap dokumen yang ditulis seorang penulis dibandingkan tulisan lama penulis dan mendeteksi dokumen yang kemungkinan bukan merupakan karya seorang penulis.

*sistem temu kembali informasi* merupakan menemukan kembali informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Penelitian ini hendak mengimplementasikan teknik Sistem Temu Kembali Informasi dengan metode Boolean Retrival Model (BRM). Dalam Boolean Retrival Model setiap kata akan dibuat incidence matrix dan dibandingkan dengan 0 dan 1 (false dan true).

**1.2 Rumusan Masalah**

Berdasarkan uraian latar belakang, penulis dapat merumuskan masalah sebaga berikut :

1. Bagaimana melakukan deteksi cara tulis seorang penulis yang kemungkinan tidak ditulis penulis tersebut?
2. Bagaimana Cara cek agar dapat dilakukan secara otomatis?

**1.3 Batasan Masalah Penelitian**

Penulis membangun fitur pada sistem dengan beberapa batasan masalah, seperti diantaranya :

1. Program yang ditulis belum sempurna masih banyak yang belum ditambah, rule dasar deteksi belum banyak.
2. Fitur deteksi membedakan gaya tulis dengan yang lain sedangkan manusia menulis sesuai dengan suasana hati, maka jika dalam tulisan 1 penulis menulis cabai tetapi tulisan 2 penulis menulis cabe.
3. Algorithma mendeteksi dokumen mana yang berbeda dengan yang lain jadi bukan pendeteksi plagiarisme tetapi algoritme bisa digunakan untuk deteksi plagiarisme.

**1.4 Tujuan Penelitian**

Tujuan dari pembuatan fitur deteksi Perbedaan cara tulis seorang penulis dengan beberapa rule (Penanda) yang sudah ditentukan dari tulisan penulis sebelumnya sehingga pengawas dapat menentukan dari dokumen yang dicek apakah ada dokumen yang dicurigai merupakan hasil plagiat.

3

**1.5 Manfaat Penelitian**

Manfaat dari adanya penelitian ini, antara lain :

1. Bagi penulis
   1. Sebagai wadah dalam penerapan ilmu yang telah diperoleh selama proses perkuliahan
   2. Sebagai bekal untuk menempuh dunia kerja
2. Bagi Instansi terkait
   1. Terciptanya fitur deteksi perbedaan karya tulis dalam sistem yang mengharuskan seorang penulis menulis beberapa tulisan.
   2. Mempermudah pengecekan beberapa hasil karya dari beberapa penulis.

**BAB II**

**LANDASAN TEORI**

**2.1 Tinjauan Studi**

Dalam menulis tulisan ini penulis menggunakan beberapa tulisan yang bisa digunakan sebagai acuan perbandingan untuk meningkatkan ketepatan hasil program, penulis mengambil referensi dari jurnal yang berkaitan dengan Sistem temu kembali informasi.

**Tabel 2.1** *State of the art*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Judul** | **Tahun** | **Metode** | **Hasil** |
| 1 | The Performance of Boolean Retrieval and Vector Space Model in Textual Information Retrieval | 2017 | Pembandingan Boolean Retrival model dan Vektor Space Model dengan corpus Reuters | By using corpus Reuters (2610 docs, 2 MB), we find there is no significant time difference for creating inverted index for both method. The differences come from querying index. Number of generated indexes, generated file size, duration of reading and searching index, and results found are growing inline with corpus file number and file size |
| 2 | Efficiency of Boolean Search strings for Information Retrieval | 2017 | By taking over the control of which document are relevant, the search for the articles is more efficient and time economical | This is carefully analysed by the two results (free text query and the Boolean search strings) and it shows that the Boolean search strings return more relevant documents (in less number) than the free text query which returned large number of documents with lots of irrelevant ones. |
| 3 | Boolean Retrieval Model for Indexed Documents | 2018 | based on boolean retrieval which retrieves the results according to the according to the Boolean operation specified within the terms of the search query. Also the proposed model is capable to store large indexes. | Boolean retrieval model used for information fetch is more accurate as compared to other retrieval models. The model creates the inverted indexes of terms and docs, on which boolean operation can be applied easily and show accurate result. |
| 4 | Arabic Book Retrieval using Class and Book Index Based Term Weighting | 2017 | class-based indexing method called inverse class frequency (ICF) and book-based indexing method inverse book frequency (IBF) for this Arabic information retrieval | TF.IDF.ICF.IBF term weighting method can be applied to the retrieval of Arabic documents that have a hierarchy of books with many pages. The experiment results showed that this method has the highest precision, recall and F-Measure value compared with other term weighting methods including TF.IDF, TF.IDF.ICF, and TF.IDF.IBF |
| 5 | Fuzzy Rule Based Profiling Approach For Enterprise Information Seeking and Retrieval | 2017 | Development of fuzzy logic based profiling approach for an enterprise system using enterprise IS&R was presented. The approach provides a new mechanism for developing and integrating task, user and document profiles into a unified index through the use of relevance feedback and a fuzzy rule based summarisation technique | From 35 users on 20 predefined simulated enterprise IS&R research carried out as part of this research clearly f ound significant co-relation between the implicit and explicit relevance feedback. It was found that there was a linear relatio nship between the implicit parameters (i.e. time on page, mouse movements, and mouse clicks) and the explicit document relevance |

5

**2.2 Tinjauan Pustaka**

**2.2.1. Karya Tulis**

Pengertian karya Tulis adalah sebuah tulisan dari hasil buah pikiran penulis dari apa yang ingin disampaikan dari sebuah penelitian atau sebuah pola pikir penulis.

Karya Tulis dari sebuah penelitian biasanya dihasilkan dari pemikiran penulis terhadap apa yang ingin dibuktikan dari pengamatan dan kesimpulan, karya tulis penelitian bisanya berupa jurnal dan tulisan akademik lainya, sedangkan karya tulis dari pola pikir biasanya diahasilkan dari apa yang pendapat dan imajinasi seorang penulis, karya tulis pola pikir bisanya berupa novel, cerita pendek, puisi, sair dan karya fiksi lainya.

Dalam sebuah karya tulis terdapat gaya tulisan penulis dan kebiasaan penulis bisa dikatakan bahwa karya tulis merupakan bagian dari penulis, dari pemilihan kata sampai kecerobohan penulis tersimpan dalam karya tulis, tidak hanya dalam karya tulis pola pikir tetapi dalam karya tulis penelitian juga terdapat pemilihan kata penulis.

Masalah yang dialami saat mengasumsikan dengan gaya tulis adalah kadang penulis menulis dengan sesuai suasana hati, seperti jika karya tulis bersifat formal maka biasanya penulis yang menggunkan kata “aku” diganti dengan “saya” agar terlihat lebih sopan, dan masalah lain yang dialami adalah jika penulis menggunkan rujukan maka penulis mengganti kata seperti rujukan misal biasanya “cabai” diganti dengan “cabe”.

6

**2.2.2. Boolean Retrival Model**

Boolean Retrival Model (BRM) model proses Pencarian informasi dari query, yang menggunakan ekspresi boolean yang berupa 1 dan 0 (true dan false), ekspresi boolean dapat berupa operator logika and,or and not.

Pada BRM yang ada hanya dokumen relevan atau tidak sama sekali, tidak ada pertimbangan dokumen yang ‘mirip’, dalam pengerjaan operator boolean (and, not, or) ada urutan pengerjaanya (operator precedence).

Urutanya adalah:

1. () Prioritas yang berada dalam tanda kurung
2. NOT
3. AND
4. OR

Jika kita ingin mencari dari cerita karangan shakespheare yang mengandung kata Brutus AND Caesar AND NOT Calpurnia, salah satu caranya adalah mencari dari semua teks dari awal sampai akhir, komputer juga dapat melakukan hal ini dan disebut grepping.

Kalau kita melakukan grepping terhadap ribuan dokumen kita perlu melakukan sesuatu yang lebih baik, karena ada beberapa masalah yang harus diselesaikan yaitu kecepatan dalam pemrosesan dokumen yang jumlahnya sangat banyak, fleksibilitas dam perangkingan, salah satu pemecahan adalah membangun index.

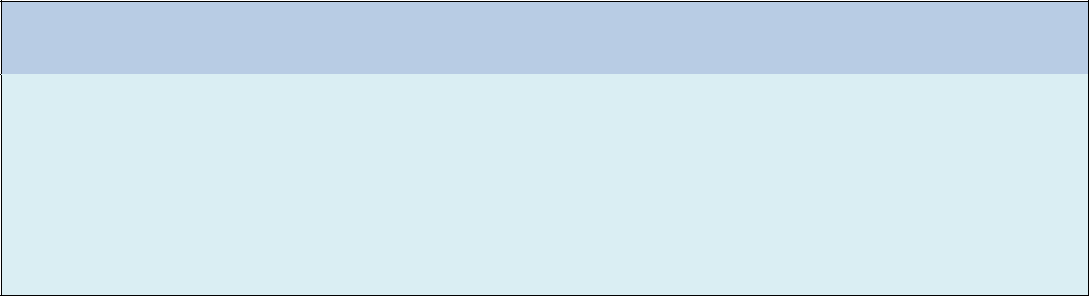
**Incidence Matrik**

Incidence Matriks adalah suatu matrix yang terdiri dari kolom (dokumen) dan baris (token/terms/kata), Pembangunan index akan berbeda untuk tiap metode Retrieval, Untuk boolean model, salah satunya kita akan menggunakan Incidence matrix sebagai index dari korpus (kumpulan dokumen) data kita, Dokumen yang ada di kolom adalah semua dokumen yang terdapat pada korpus data kita, Token/Terms/Kata pada baris adalah semua token unik (kata yang berbeda satu dengan yang lainnya) dalam seluruh dokumen yang ada, Saat suatu token(t) ada dalam dokumen(d), maka nilai dari baris dan kolom (t,d) adalah 1. Jika tidak ditemukan, maka nilai kolom (t,d) adalah 0, Dari sudut pandang kolom, kita bisa tahu token apa saja yang ada di satu dokumen (d), Dari sudut pandang barisnya, kita bisa tahu di dokumen mana saja token (t) ada (posting lists).

**Inverted Index**

Inverted index mempunyai vocabulary, yang berisi seluruh term yang berbeda pada masing-masing dokumennya (unik), dan tiap-tiap term yang berbeda ditempatkan pada inverted list.

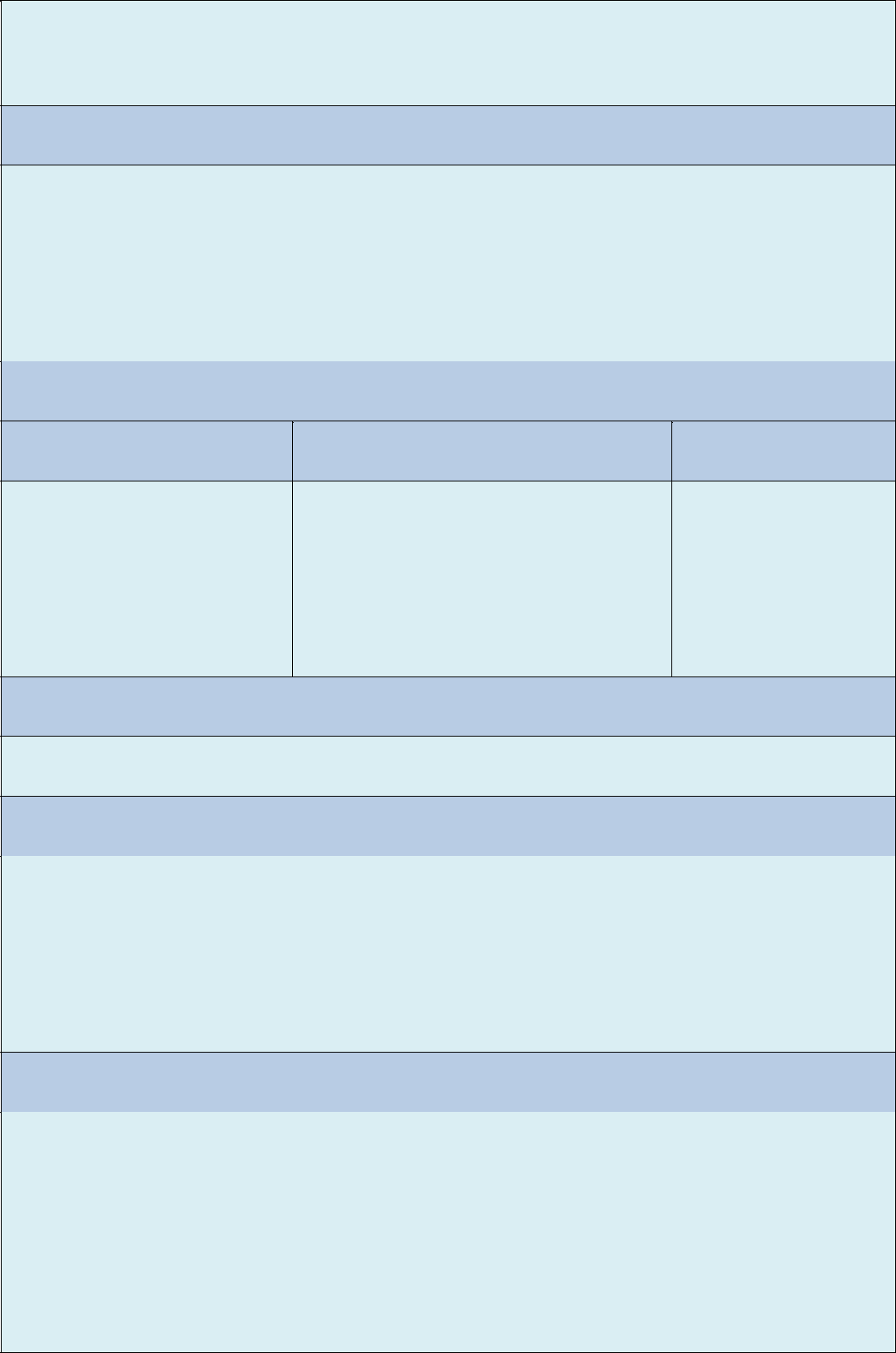
**2.3 Kerangka Pemikiran**



**Masalah**

Internet yang maju menjadikan jumlah informasi yang didapat seseorang tak terhingga, menjadikan dokumen gampang didapat dan sulit dicari, jika seseorang yang penulis diharuskan menulis beberapa karya maka penulis menggunakan internet sebagai alat plagiat.

8



Sebagai seseorang atau instansi yang mengawasi dan mengeditnya itu menjadi masalah karena karya harus original.

**Tujuan**

Tujuan dari penelitian ini adalah mendeteksi perbedaaan tulisan penulis yang sama dengan menggunakan tanda didalam gaya tulis dan kesalahan penulis, sehingga saat pengeditan pengawas bisa membedakan karya sang penulis dengan karya sang penulis dengan tulisan sebelumnya.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Eksperimen** |  |
|  | **Data** | **Metode** | **Tool** |
| 1. | Abstrak laporan | *Boolean Retrival Model (BRM)* | *Python* |
|  | akhir penelitian |  |  |
| 2. | Abstrak proposal |  |  |
|  | Penelitian |  |  |
|  |  | **Pengujian** |  |

**Hasil**

* Penelitian ini menghasilkan pembedaan beberapa dokumen dengan dokumen lain yang sedang diuji

**Manfaat**

* Terciptanya fitur deteksi pembeda beberapa dokumen / tulisan beda penulis.
* Berkurangnya sebuah karya hasil plagiat.

**BAB III**

**METODE PENELITIAN**

**3.1** **Instrumen Penelitian**

Beberapa komponen yang dibutuhkan untuk melakukan penelitian ini sebagai berikut :

**3.1.1 Kebutuhan Perangkat Lunak (Software)**

Perangkat lunak yang digunakan untuk membangun sistem ini antara lain :

1. Sistem Operasi : Windows 7
2. Python 3.7.3
3. Notepad++ sebagai teks editor dll

**3.1.2 Kebutuhan Perangkat Lunak (Software)**

Perangkat keras yang digunakan untuk membangun sistem ini antara lain :

a. Processor Core i3

b. RAM 2 GB

**3.2** **Prosedur Pengambilan dan Pengumpulan Data**

Pada penelitian ini penulis mengumpulkan data untuk memperbanyak rule yang digunakan sebagai menyempurnakan algorithm, database term atau kata yang sama juga. Beberapa metode yang digunakan sebagai berikut :

1. Observasi

Observasi merupakan metode pengumpulan data yang dilakukan dengan cara pengamaan langsung terhadap data-data yang berkaitan dengan penilitian.

7

1. Wawancara

Wawancara merupakan metode pengumpulan data yang hampir sama dengan metode observasi hanya saja dilakukan dengan cara Tanya jawab langsung kepada narasumber berkaitan dengan data-data yang diperlukan pada penelitian.

1. Studi Pustaka

Studi pustaka merupakan metode pengumpulan data dengan mencari referensi dari buku-buku dan media internet yang berkaitan dengan data-data yang diperlukan pada penelitian.

**3.3** **Teknik Analisis Data**

Setelah proses pengumpulan data, tahapan selanjutnya yaitu proses analisa data. Beberapa tahapan yang dilakukan untuk analisa data antara lain :

1. Memilih data sesuai dengan keperluan dari data-data mentah kemudian diolah menjadi data primer. Data primer yang didapatkan antara lain :
   1. Data identitas usulan
   2. Data abstrak proposal
2. Membuat kamus yang nanti akan digunakan sebagai pengenalan term/kata yang dianggap sama.
3. Data primer yang telah didapatkan kemudian akan melalui tahap *pre-processing* untuk mendapatkan data yang siap digunakan.
4. Data yang telah berisi kata-kata tersebut dimasukkan ke dalam database.

**3.4** **Model/Metode yang Diusulkan**

Penelitian ini menggunakan model/metode *Boolean Retrival Model* dimana untuk menyambungkan setiap term yang **dianggap** sama kemudian melakukan indecence matriks.

Proses program dalam penelitian

Mulai

memasukan dokumen uji coba dan pengambilan database term/kata yang dianggap sama

hasil pembedaan dokumen

pembedaan setiap dokumen menggunakan indicence matriks

Selesai

Penjelasan alur proses pada gambar diatas sebagai berikut :

1. Dataset

Dataset yang digunakan adalah database kata yang dianggap sama,

* 1. Tahap Preprocessing

Dalam Sistem Temu Kembali processing digunakan untuk mempercepat proses, tetapi dalam tulisan ini Prepocessing tidak digunakan karena tulisan ini memerlukan setiap kata dan susunan penulisanya apa adanya.

1. Pembedaan setiap dokumen terhadap dokumen yang lain

2.1 Rule

Kunci utama dalam tulisan ini adalah kesempurnaan rule (penanda perbeda dokumen), rule yang ada dalam tulisan ini belum sempurna, rule pembeda antar dokumen satu penulis ada beberapa yaitu:

1. Menggunakan dua kata berbeda yang memiliki arti yang sama, contoh : cabai dan cabai.
2. Menggunakan pemilihan kata penulis, cotoh : saya(nada sopan) dan aku(nada biasa).
3. Kata yang tidak disingkat menjadi disingkat, contoh : yang dan yg.
4. Penggunaan kata dalam me- dalam kata dasar di setiap dokumen, contoh : memukul dan pukul.

2.2 Incidence Matrix

Setiap kata digunakan dalam incidence matrix karena tidak di token dan stemming jadi jumlahnya baris sangat banyak dan memiliki kolom dari beberapa dokumen dan selanjutnya adalah dua baris yang memiliki kata yang dianggap sama akan dilakukan Boolean Logic agar ditentukan apakah ada dokumen yang dianggap berbeda.

2.3 Boolean Logic  
 Tulisan ini memang dapat diselesaikan menggunakan Vektor Space Model (VSM) tetapi penulis menggunakan Boolean Retrival Model, boolean logic digunakan untuk setiap baris yang dianggap sama. dan Boolean Logic Pembeda dokumen adalah:

( a or not b ) and ( not a or b) or ( a and b)

a : baris yang dianggap sama 1

b : baris yang dianggap sama 2

1. Hasil pembedaan dokumen

Dokumen yang dianggap berbeda adalah dokumen yang memiliki nilai beda yang melebihi tertentu, untuk menentukan nilai beda adalah dalam setiap Boolean logic sebelumnya hasil Boolean logic yang bernilai nol di kolom hasil maka nilai beda di dokumen yang memiliki kolom tersebut akan bertambah.