# *AUTOMATIC SEARCH IN A SINGLE DOCUMENT THE CORE SENTENCES INDONESIAN LANGUAGE*

*[1] Syamsul Arifin(0910651057), [2] Lutfi Ali Muharom S.Si, [3] Ratih Ayuninghem, S.ST, M.Kom*
*Jurusan Teknik Informatika, Fakultas Teknik*
*Universitas Muhammadiyah Jember*
*Email : syamsul_arifin19@yahoo.com*

## *ABSTRACT*

*Search this sentence is automatically nuclei using tetx mining and weighting document to find the core sentence of a single document. Core search process automatically sentence was originally a form of document that will then be in the form of sentences and then of the sentences will be the core of a document sentence. document in this study is a collection of sentences Obtained from a sports site on the internet. This research is the development of an application that generates an automatic search phrase core document in Indonesian. The method used in this study is the Term Frequency-Inverse Document Frequency (TF-IDF) in roomates to calculate the weight of each word. By applying Text Mining and Term Frequency-Inverse Document Frequency (TF-IDF), result abtained from this research is the application that is able to produce a summary document of the document article.*

*Keyword : search, core sentence, automatically, text mining, tf-idf.*

## ABSTRAK

*[1] Syamsul Arifin(0910651057), [2] Lutfi Ali Muharom S.Si, [3] Ratih Ayuninghem, S.ST, M.Kom*
*Jurusan Teknik Informatik,a Fakultas Teknik*
*Universitas Muhammadiyah Jember*
*Email : syamsul_arifin19@yahoo.com*

Pencarian kalimat inti secara otomatis ini adalah menggunakan *tetx mining* dan pembobotan dokumen untuk menemukan kalimat inti dari sebuah dokumen tunggal. Proses pencarian kalimat inti secara otomatis ini awalnya adalah berupa dokumen yang kemudian akan di bentuk menjadi kalimat-kalimat yang kemudian dari kalimat-kalimat tersebut akan di dapat kalimat inti dari sebuah dokumen. dokumen di penelitian ini adalah kumpulan dari kalimat-kalimat yang didapat dari situs olahraga di internet. Penelitian ini adalah pengembangan sebuah aplikasi yang menghasilkan pencarian otomatis kalimat inti dokumen dalam Bahasa indonesia. Metode yang digunakan dalam penelitian ini adalah *Term Frekuensi-Frekuensi Dokumen Invers* (TF-IDF) yang mana dengan menghitung bobot setiap kata. Dengan menerapkan *Text Mining* dan *Term Frekuensi-Frekuensi Dokumen Invers* (TF-IDF) ini, hasil yang diperoleh dari penelitian ini adalah aplikasi yang mampu menghasilkan summarization dokumen dari dokumen berita.

Kata Kunci*: pencarian, kalimat inti, otomatis,text mining, tf-idf.*

## PENDAHULUAN

### Latar Belakang

Perkembangan teknologi di bidang komputasi dan telekomunikasi telah memberi kemudahan kepada setiap orang untuk menyampaikan dan mendapatkan informasi. Kemudahan ini menyebabkan informasi menjadi semakin banyak dan beragam. Informasi dapat berupa dokumen, berita, surat, cerita, laporan penelitian, data keuangan dan lain-lain.

Ketersediaan informasi yang semakin banyak menjadikan ringkasan sebagai kebutuhan yang sangat penting. Dengan adanya ringkasan, pembaca dapat dengan cepat dan mudah memahami makna sebuah teks tanpa harus membaca keseluruhan teks. Hal ini dapat menghemat waktu pembaca karena dapat menghindari pembacaan teks yang tidak relevan dengan informasi yang diharapkan oleh pembaca, terutama jika informasi tersedia di internet cukup banyak.

Berbagai metode telah diterapkan dan masih terus dikembangkan oleh para peneliti tentang peringkasan. Salah satunya adalah peringkasan dokumen yang kemudian akan menghasilkan informasi yang penting atau inti kalimat dari dokumen dengan menggunakan komputer. Tujuannya adalah mengambil sumber informasi dengan mengutip sebagian besar isi yang penting untuk mencari kalimat inti dari sebuah dokumen artikel dan menampilkan kepada pembaca dalam bentuk yang ringkas sesuai dengan kebutuhan pembaca. Dengan demikian teknologi ini dapat membantu pembaca untuk menyerap informasi yang ada dalam artikel melalui ringkasan tanpa harus membaca seluruh isi dokumen.

Pencarian otomatis kalimat inti ini dibutuhkan pembobotan pada masing-masing term sehingga metode *term frequency – invers document frequency* (*tf-idf*) digunakan sebagai penentu bobot masing-masing term tersebut.

### Rumusan Masalah

Dari latar belakang yang sudah dijelaskan di atas diperoleh rumusan masalah sebagai berikut

a. Bagaimana mengolah dokumen dengan teks mining.
b. Bagaimana menghitung bobot yang dimiliki suatu dokumen dalam keseluruhan teks menggunakan metode TF-IDF.
c. Bagaimana mengurutkan bobot yang diperoleh dari metode TF-IDF.

### Batasan Masalah

a. Input hanya dari satu dokumen atau satu artikel saja.
b. Dokumen berupa artikel dari website/situs olahraga.
c. Dokumen yang digunakan sebagai data set mengambil dari situs Goal.com.
d. Database untuk kata-kata yang tidak relavan mengambil dari kamus bahasa indonesia.
e. Input dan output dalam bahasa indonesia.
f. Tidak menangani kesalahan penulisan kata.
g. Singkatan di anggap satu kata.
h. Hanya melayani list dokumen.

### Tujuan Penelitian

a. Mengolah dokumen dengan teks mining dan melakukan pembobotan dengan tf-idf.
b. Melakukan pencarian otomatis kalimat inti dalam sebuah dokumen tunggal berbahasa indonesia.
c. Mengambil sumber informasi dengan mengutip sebagian besar isi yang penting dan menampilkan kepada pembaca dalam bentuk yang ringkas sesuai dengan kebutuhan pembaca.

### Manfaat Penelitian

a. Dapat membantu pembaca untuk menyerap informasi yang ada dalam artikel melalui ringkasan tanpa harus membaca seluruh isi dokumen.
b. Dapat menghemat waktu pembaca karena dapat menghindari pembacaan teks yang tidak relevan dengan informasi yang diharapkan oleh pembaca.

## TINJAUAN PUSTAKA

### Kalimat Inti

Kalimat inti sering juga di katakan sebagai kalimat utama ataupun kalimat topik. Pengertian dari kalimat sendiri ini adalah kalimat yang paling penting dalam sebuah *paragraph* karena merupakan ide utama dalam *paragraph* tersebut. Kalimat inti atau kalimat topik bertugas memberitahukan kepada pembaca gagasan pokok yang akan dibicarakan dalam alinea dimaksud. Karena karangan terdiri dari alinea-alinea yang memiliki kesatuan yang utuh, maka kalimat inti yang biasanya terletak di awal paragraf juga berfungsi menghubungkan alinea itu dengan alinea sebelumnya. Jika kalimat tersebut sengaja dihilangkan, maka isi paragraf tersebut akan hilang. Hal ini tidak terjadi dengan kalimat-kalimat lain yang memang hanya berfungsi sebagai penjelas. Kalimat utama berupa ringkasan dari sebuah paragraf yang rupakan pandangan mendalam dari ide pokok penulis dalam paragraf tersebut.

### Dokumen

Kata dokumen berasal dari bahasa latin yaitu docere, yang berarti mengajar. Pengertian dari kata dokumen ini menurut Louis Gottschalk (1986; 38) seringkali digunakan para ahli dalam dua pengertian, yaitu pertama, berarti sumber tertulis bagi informasi sejarah sebagai kebalikan daripada kesaksian lisan, artefak, peninggalan-peninggalan terlukis, dan petilasan-petilasan arkeologis. Pengertian kedua diperuntukan bagi surat-surat resmi dan surat-surat negara seperti surat perjanjian, undang-undang, hibah, konsesi, dan lainnya. Lebih lanjut, Gottschalk menyatakan bahwa dokumen (dokumentasi) dalam pengertiannya yang lebih luas berupa setiap proses pembuktian yang didasarkan atas jenis sumber apapun, baik itu yang bersifat tulisan, lisan, gambaran, atau arkeologis.

### Artikel

Artikel *(article)* adalah tulisan tentang suatu masalah berikut pendapat penulis tentang masalah tersebut yang dimuat di media *online* ataupun media cetak seperti surat kabar, majalah, atau buletin. Secara umum dapat dikatakan bahwa artikel tidak terlalu panjang, hanya beberapa halaman. Artikel termasuk tulisan kategori *views* (pandangan), yaitu tulisan yang berisi ide, opini, atau penilaian penulisnya terhadap suatu peristiwa atau masalah.

Penulis artikel bertujuan menyampaikan gagsan dan fakta guna meyakinkan, mendidik, menawarkan pemecahan suatu masalah, atau menghibur. Biasanya, tahap-tahap penulisan artikel meliputi latar belakang masalah, identifikasi masalah, analisis masalah, dan kesimpulan.

***Text Mining***

*Text mining* adalah salah satu bidang khusus dari *data mining*. Sesuai dengan buku the *text mining handbook*, *text mining* dapat didefinisikan suatu proses untuk menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan *tool* analisis yang merupakan komponen-komponen dalam data mining yang salah satunya adalah kategorisasi.

*Text mining* bisa dianggap subyek riset yang tergolong baru. *Text mining* dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian atau pengelompokan dan menganalisa *unstructured* teks dalam jumlah besar. Dalam memberikan solusi, *text mining* mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti *data mining*, *Information retrieval*, *Statistik* dan *Matematik, Machine Learning, Linguistic, Natural Languange Processing,* dan *Visualization*. Kegiatan riset untuk *text mining* antara lain ekstraksi dan penyimpanan teks, *preprocessing* akan konten teks, pengumpulan data statistik dan *indexing* dan analisa konten.

***Tokenizing***

Penggunaan istilah kata dan token sering kali saling dipertukarkan. Sebuah token dapat di definisikan sebagai unit terkecil dari sebuah teks atau dapat juga merupakan suatu kumpulan dari string alphanumerik [Konchady,2006]. Suatu unit terkecil yang digunakan disini adalah sebuah kata tungggal. Sebuah kata tunggal dapat berisi kumpulan dari string alphanumerik. Pada proses *tokenizing* ini, akan terjadi proses pemotongan dokumen menjadi daftar kata yang berdiri sendiri sebelum dilakukan proses selanjutnya. Tahap *Tokenizing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Lebih tepatnya adalah proses pemecahan dokumen yang ada dalam sebuah dokumen menjadi kata.

***Stoplist***

Dalam sebuah dokumen sangat banyak didapatkan jenis-jenis kata seperti, kata sambung, kata depan, kata ganti, kata sifat dan lain sebagainya. Sebagian besar kata-kata itu merupakan kata yang tidak berpotensi dalam mengidentifikasikan isi sebuah dokumen. Selain itu juga pendefinisian dokumen dengan berdasarkan frekuensi kata menjadi kurang efektif bila jenis kata-kata diatas tidak dilakukan proses penyaringan.

Stoplist adalah proses penyaringan (*filtering*) terhadap kata-kata yang tidak layak untuk dijadikan sebagai pembeda atau kata kunci sehingga kata-kata tersebut dapat dihilangkan dari dokumen.

***Stemming***

Setelah kata-kata yang terdapat dalam dokumen menjalani proses *tokenizing* dan *stoplist*, maka selanjutnya kata-kata yang tersisa akan menjalani proses stemming. Proses stemming betujuan untuk mengubah atau mengembalikan kata menjadi kata/bentuk dasarnya dengan menghilangkan imbuhan-imbuhan pada kata dalam dokumen. Proses stemming dilakukan dengan mengecek kata apakah mengandung imbuhan atau tidak, terutama imbuhan yang bersifat *suffix*. Proses stemming kata dalam Bahasa Inggris memiliki karakteristik tersendiri [Porter, 2001], yang tidak lepas dari pengaruh tata bahasanya. Selain itu, yang perlu diperhatikan pada proses ini adalah konsistensi dalam memproses suatu kata, meskipun tidak sepenuhnya sempurna dalam memberikan hasil.

**Term Frequency-InverseDocument Frequency**

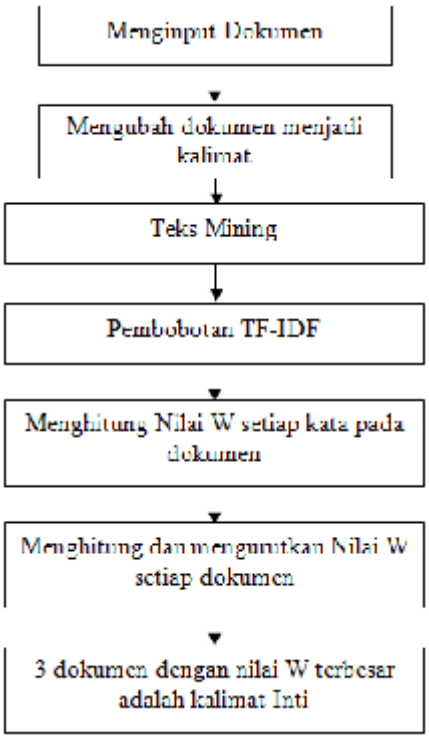
MetodeTerm Frequency-InverseDocument Frequency(TF-IDF)adalahcarapemberian bobot hubungan suatu kata (term) terhadap dokumen. Untuk dokumen tunggaltiap kalimat dianggap sebagai dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu Term frequency (TF) merupakan frekuensi kemunculan kata (t) pada kalimat (d). Document frequency(DF) adalah banyaknya klaimatdimana suatukata(t) muncul. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut.

Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Bobot kata semakin besar jika sering muncul dalam suatu dokumen dansemakin kecil jika muncul dalam banyak dokumen (Robertson, 2005). Pada Metode ini pembobotan kata dalam sebuah dokumen dilakukan dengan mengalikan nilai TF dan IDF.

## METODE PENELITIAN

**Desain Sistem**

Model yang akan di kembangkan dalam pencarian kalimat inti dalam penelitian dapat di lihat pada gambar di bawah ini:




Gambar Desain Sistem

Adapun langkah-langkah pencarian otomatis kalimat inti adalah sebagai berikut:

1. Menginput dokumen yang akan dicari kalimat intinya
2. Melakukan *Scanning*untuk menelusuri setiap kata pada dokumen
3. Proses *casefolding* yaitu proses merubah huruf menjadi huruf kecil dan selain karakter akan dibuang menjadi dokumen
4. Memilah dokumen menjadi beberapa kalimat . Pemilahan kalimat dilakukan dengan memecah string teks dari

dokumen yang panjang menjadi kalaimat-kalimat mengunakan fungsi *split()*,dengan tanda titik ”.”, tanda tanya ”?” dan tanda seru ”!” sebagai delimiter untuk memotong string dokumen

5. Memilah kalimat yang terbentuk menjadi kata yaitu dengan *tokenizing*, dilanjutkan dengan proses *stoplist* yaitu membuang kata-kata yang tidak relavan dan proses mengembalikan kata ke bentuk dasar dengan menggunakan *stemming.*
6. Pembobotan TF-DF
   Pembobotan diperoleh berdasarkan jumlah kemunculan term dalam kalimat (TF) dan jumlah kemunculan term pada seluruh kalimat dalamdokumen ( IDF). Bobot suatu istilah semakin besar jika istilah tersebut sering muncul dalam suatu dokumen dan semakin kecil jika istilah tersebut muncul dalam banyak dokumen ( Grossman, 1998) . Nilai IDF sebuah term dihitung menggunakan persamaan berikut :

   $$= \log\frac{}{}$$

   Dengan
   *N* : jumlah kalimatyang berisi *term*(t)
   *Dfi* : jumlah kemunculan kata(*term*)terhadap D
7. Menghitung bobot (*W*) masing-masing kata pada dokumen dengan persamaan berikut (Mustaqhfiri,2011) :

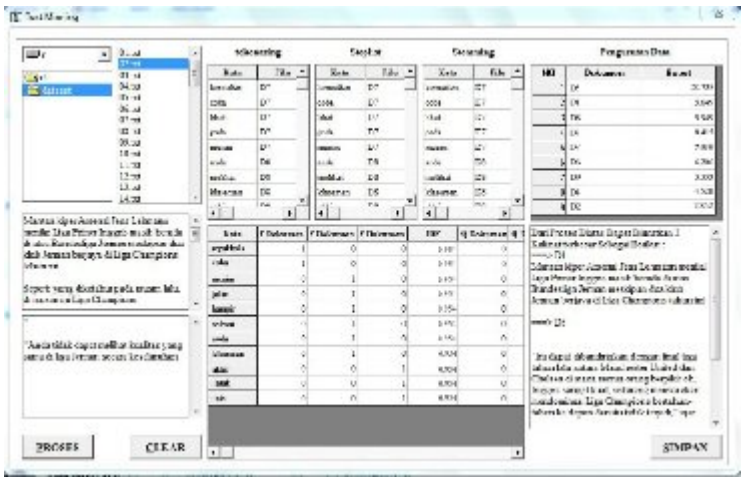
   $$W_{d,t} = TF_{d,t} * IDF_t$$

   Dengan
   $d$ = kalimat ke-**d**
   $t$ =kata(*term*) ke –t
   $TF$ = *term frequency*
   $W$ = bobot kalimat ke-d terhadap kata(*term*)ke- t
   $IDF$ = *inverse document frequency*
8. Melakukan proses pengurutan (*sorting*) nilai kumulatif dari W untuk setiap kalimat
9. Tiga kalimat dengan nilai W terbesar dijadikan sebagai hasil dari ringkasan atau sebagai output dari pencarian otomatis kalimat inti.

## HASIL DAN PEMBAHASAN

### Pencarian Otomatis Kalimat Inti

Pada Penelitian ini pengembangkan aplikasi peringkasan teks otomatias digunakan
bahasa pemrograman Microsoft Visual Basic. Aplikasi yang dikembangkan dibuat untuk single dokumen secara dinamis artinya hanya membuat Pencarian dari satu dokumen tetapi dapat digunakan untuk dokumen yang berbeda.
Berikut tampilan sistem pencarian otomatis kalimat inti dokumen tunggal berbahasa indonesia :

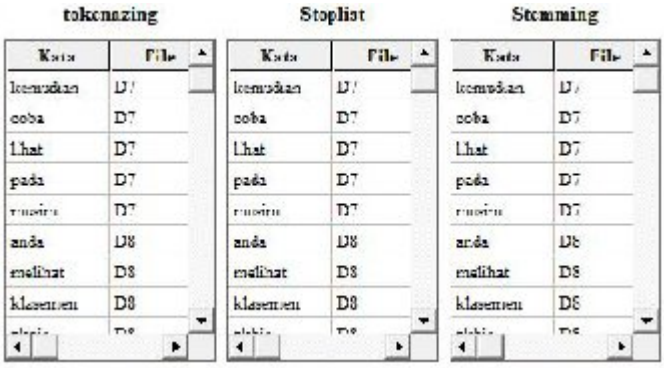



**Gambar Sistem Pencarian Otomatis Kalimat inti**

Gambar diatas adalah sistem pencarrian otomatis kalimat inti dengan menggunakan metode tf –idf. Berikut pembahasan dari sistem pencarian kalimat inti :

**Proses Text Mining**

Pada proses ini, memperlihatkan dimana dokumen dipecah menjadi kata dan kemudian dihitung nilai kemunculannya.




**Gambar Proses *Text Mining***

Proses TF-IDF

Pada proses ini, memperlihatkan nilai IDF yang ditemukan pada proses *text mining*.

| Kata | f Dokumen | f Dokumen | f Dokumen | IDF | ij Dokumen | ij D |
|---|---|---|---|---|---|---|
| sepakbola | 1 | 0 | 0 | 0.954 | 0 | |
| coba | 1 | 0 | 0 | 0.954 | 0 | |
| musim | 0 | 1 | 0 | 0.954 | 0 | |
| jalan | 0 | 1 | 0 | 0.954 | 0 | |
| hampir | 0 | 1 | 0 | 0.954 | 0 | |
| selesai | 0 | 1 | 0 | 0.954 | 0 | |
| once | 0 | 1 | 0 | 0.954 | 0 | |
| klasemen | 0 | 1 | 0 | 0.954 | 0 | |
| akhir | 0 | 0 | 1 | 0.954 | 0 | |
| jarak | 0 | 0 | 1 | 0.954 | 0 | |
| lain | 0 | 0 | 1 | 0.954 | 0 | |

**Gambar Pembobotan TF-IDF**

**Pengurutan Bobot (*W*)**

Pada tahap inilah yang paling penting karena tahap inilah yang menentukan hasil dari pencarian kalimat inti.

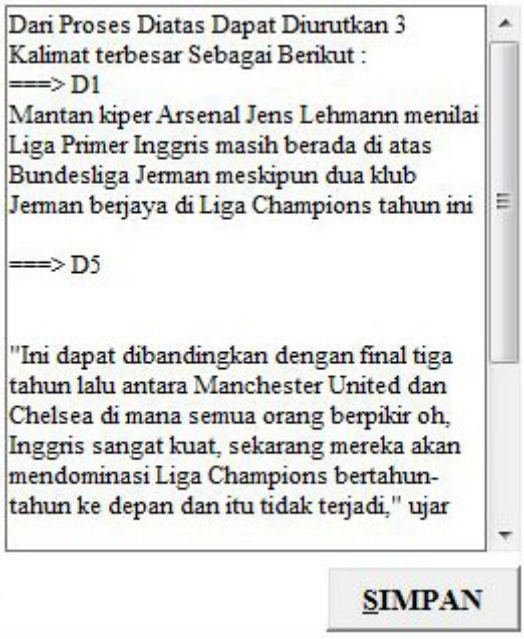
**Pengurutan Data**

| NO | Dokumen | Bobot |
|---|---|---|
| 1 | D5 | 20.738 |
| 2 | D1 | 15.649 |
| 3 | D8 | 9.949 |
| 4 | D3 | 8.415 |
| 5 | D7 | 7.030 |
| 6 | D6 | 6.206 |
| 7 | D9 | 5.553 |
| 8 | D4 | 4.520 |
| 9 | D2 | 2.862 |

**Gambar Pengurutan Bobot (*W*)**

**Hasil Pencarian**

Tahap inilah yang menampilkan hasil dari pencarian otomatis kalimat inti.




**Gambar hasil pencarian**

## KESIMPULAN DAN SARAN

### KESIMPULAN

Dari uji coba dan analisa yang telah dijelaskan dalam bab sebelumnya, maka penulis mengambil beberapa kesimpulan sebagai berikut :

1. Program ini berhasil mengolah data atau dokumen artikel dengan pra-proses *Text mining* dan pembobotan TF-IDF, serta dapat melakukan pencarian secara otomatis kalimat inti dalam sebuah dokumen berbahasa indonesia.
2. Program ini berhasil mengambil sumber informasi dengan mengutip sebagian besar isi yang penting dan menampilkan kepada pembaca dalam bentuk yang ringkas sesuai dengan kebutuhan pembaca dengan tidak mengubah produk teks yang memiliki / mengandung bagian-bagian yang penting dari artikel asli meskipun secara tata bahasa belum baik. Selain itu, sistem yang dikembangkan ini mampu menghasilkan *summarization* dokumen dari sebuah dokumen artikel.

### SARAN

Penulis ingin memberikan beberapa saran yang mungkin dapa tmembantu dalam pengembangan Tugas Akhir ini adalah sebaga berikut :

1. Untuk penelitian selanjutnya pengembang diharapkan dapat menggunakan dokumen dalam bahasa inggris atau bahasa lainnya, bukan hanya menggunakan bahasa Indonesia.Sehingga dapat menghasilkan program yang lebih meluas lagi dan berguna bagi pembaca.
2. Penulis menyarankan bagi pengembang agar memadukan dengan metode lainnya bukan hanya tf-idf agar hasil yang diperoleh lebih bagus lagi.
3. Diharapkan pengembang mampu menghitung akurasi dari hasil percobaan.

## DAFTAR PUSTAKA


Arifin, A Z., (2002), *Penggunaan Digital Tree Hibrida pada Aplikasi Information Retrieval untuk Dokumen Berita*, Jurusan Teknik Informatika, FTIF, Institut Teknologi Sepuluh Nopember, Surabaya.

Gottschalk, L., (1986), *Understanding History: A Primer of Historical Method* (terjemahan Nugroho Notosusanto). UI Press,Jakarta.

Grossman, D., Ophir, F., (1998), *Information Retrieval : Algorithm and Heuristics*. Kluwer Academic Publisher,USA.

Jones K, S., (1998), *Charting a New Course: Natural Language Processing and Information Retrieval : Essays in Honour of Karen Spärck Jones.* Penerbit: Springer,New York.

Kunchady, M., (2006), *Text Mining Application Programming.* Thomson Learning Inc. ISBN 1-58450-460-9,USA.

Mustaqhfiri, M., Abidin Z., Kusumawati,R.(2011). Peringkasan Teks Otomatis Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance.Ejournal Matics,4,4,135-147. Dimbali 5 Januarii 2012 dari basis data saintek.

Robertson, S., (2004). "*Understanding Inverse Document Frequency: On theoretical arguments for IDF*", Journal of Documentation,Vol.60, no.5, pp. 503-520,New York.

Widjono, H.S., (2007). "*Bahasa Indonesia Mata Kuliah Pengembangan Kepribadian di PT".* Penerbit : grasindo, Jakarta.