PENCARIAN PASAL PADA KITAB UNDANG-UNDANG HUKUM PIDANA (KUHP) BERDASARKAN KASUS MENGGUNAKAN METODE *COSINE SIMILARITY* DAN *LATENT SEMANTIC INDEXING* (LSI)

Setyoko Yudho Baskoro[1], Achmad Ridok[2], Muhammad Tanzil Furqon[3]

Universitas Brawijaya
Email: [1] setyokoyudhobaskoro@gmail.com, [2] acridokb@ub.ac.id @ub.ac.id,
[3]m.tanzil.furqon@gmail.com

ABSTRACT

Indonesia is a country of law. As law states, Indonesian have regulations that govern the relationship between the communities, one of them is criminal law. Set of rules of criminal law is written in the Kitab Undang-undang Hukum Pidana (KUHP), which contains hundreds of clause which regulate the relationship between the community based on values, norms, and specific rules that focuses on the interests of the public. In this paper, information retrieval used to search the clause of the KUHP based on a description of the crime, using Latent Semantic Indexing (LSI). LSI adopts techniques in mathematical dimension reduction process Singular Value Decomposition (SVD). This system use 60 clause as training data, and 6 query or crime description as test data. In each of the data clause of the KUHP contained data such as clause number, clause, and the clause contents. The system will calculate and determine the relevant clause is based on query or description of the crimes that has been entered. Cosine similarity used to calculate the similarity or proximity clause KUHP with query. The performance of the system is shown by the test results of Mean Average Precision (MAP) value at each k-rank is 5, 10, 20, 30, 40, 50, and 59, with the highest performance is in k-rank 40 with MAP 0.8944.

Keywords: *Information Retrieval, Latent Semantic Indexing, Singular Value Decomposition, Cosine Similarity, Text Mining, KUHP*

1. PENDAHULUAN

Negara Indonesia adalah negara hukum. Sebagai negara hukum Indonesia memiliki peraturan-peraturan yang mengatur hubungan antar masyarakat, salah satunya adalah hukum pidana. Kumpulan peraturan hukum pidana tersebut tertulis secara tegas bersertakan ketentuannya di dalam Kitab Undang-undang Hukum Pidana (KUHP) yang berisi ratusan pasal yang mengatur hubungan antar masyarakat berdasarkan nilai, norma ,dan kaidah tertentu yang menitik beratkan kepada kepentingan publik (Prasetyo 2014). Namun fakta menunjukan, Indonesia sampai sekarang belum juga sampai ke tahap cita-cita negara hukum (Atiq 2015). Sebagian besar masyarakat di Indonesia tidak mengerti dan memahami peraturan - peraturan dalam bernegara, terutama pada Kitab Undang-Undang Hukum Pidana (KUHP).

Perkembangan teknologi memunculkan revolusi dan inovasi dalam ilmu pengetahuan, khususnya dalam teknologi *Information Retrieval* atau Sistem Temu Kembali Informasi. Dengan bantuan teknologi ini, permasalahan akan buta hukum di Indonesia dapat dikurangi dengan mengembangkan *Information Retrieval* terhadap pasal pada KUHP sehingga pencarian pasal berdasarkan suatu kasus dapat dengan mudah dan cepat dilakukan.

Pada penelitian ini digunakan proses temu kembali menggunakan metode *Latent Semantic Indexing* (LSI) yang memanfaatkan reduksi dimensi dari *Singular Value Decomposition* (SVD) dengan menggunakan objek data berupa pasal-pasal dalam KUHP Buku Kedua. Proses pengolahan yang pertama kali dilakukan yaitu *preprocessing* pada data pasal KUHP yang merupakan salah satu cabang dari *Natural Language Processing* (NLP). Hasil numerik dari proses pembobotan TF-IDF setelah di-*preprocesing* diolah menggunakan SVD. Kemudian dilakukan reduksi dimensi data dengan *k-rank*. Hasil reduksi SVD diolah kembali menggunakan LSI. Kemudian

digunakan Cosine Similarity untuk menghitung kedekatan antara vektor dari *corpus* dan vektor dari *query*. Hasil dari sistem ini berupa pasal yang berkaitan dengan deskripsi kasus kejahatan yang dimasukkan, sehingga diketahui akurasi dari LSI dalam proses pencarian pasal KUHP berdasarkan kasus.

2. ISTILAH HUKUM

2.1. Hukum Pidana

Menurut Martiman Prodjohamidjojo (Prasetyo 2014), Hukum Pidana adalah bagian dari keseluruhan hukum yang berlaku di suatu negara, yang mengadakan dasar - dasar dan aturan – aturan untuk menentukan perbuatan – perbuatan mana yang tidak boleh dilakukan, yang dilarang, dengan disertai ancaman atau sanksi pidana tertentu bagi siapa saja yang melanggarnya. Menentukan kapan dan dalam hal apa kepada mereka yang telah melakukan larangan –larangan itu dapat dikenakan atau dijatuhi pidana sebagaimana yang telah dicantumkan. Menentukan dengan cara bagaimana pengenaan pidana itu dapat dilaksanakan apabila orang yang diduga telah melanggar ketentuan tersebut (Prasetyo 2014).Persamaan matematika harus diberi nomor urut dalam kurung biasa dan harus diacu dalam tulisan.

2.2. KUHP

Hukum pidana di Indonesia tertulis dalam sebuah kitab undang-undang. Hukum pidana tersebut dikodifikasikan beserta ketentuan-ketentuannya di dalam Kitab Undang-Undang Hukum Pidana (KUHP) yang berasal dari zaman pemerintah penjajahan Belanda (Prasetyo 2014).

Kitab undang-Undang Hukum Pidana (KUHP) terdiri atas 569 pasal, secara sistematik dibagi dalam (Prasetyo 2014):

- Buku I : Memuat tentang Ketentuan-ketentuan Umum. Pasal 1-103.
- Buku II : Mengatur tentang Kejahatan. Pasal 104-488.
- Buku III : Mengatur tentang Pelanggaran. Pasal 489-569.

3. *TEXT MINING*

Text mining memiliki definisi menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antara dokumen. Pada cara yang sejalan dengan *data mining*, *text mining* berusaha mengutip informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi pola yang unik. Akan tetapi pada text mining, sumber data adalah kumpulan dokumen dan pola yang menarik tidak ditemukan pada record database yang terbentuk melainkan pada data kata per kata yang tidak terstruktur pada kumpulan dokumen tersebut (Feldman 2007).

3.1. *Text Processing*

Suatu data dokumen teks, mempunyai struktur kata yang tidak teratur, maka dalam melakukan proses *text mining* diperlukan beberapa proses tambahan yang bertujuan untuk menyiapkan dan mengubah data teks mentah menjadi lebih terstruktur. Salah satu tahapan dari *text mining* adalah *preprocessing text*. Pada tahap *preprocessing text* ini dilakukan proses seleksi data berupa text pada setiap dokumen, yang kemudian akan dilakukan proses selanjutnya. Pada tahap *preprocessing* ini melalui beberapa tahapan yaitu *case folding*, *tokenizing*, *filtering*, dan *stemming.* (Nugroho 2011)

3.2. TF-IDF

TF-IDF merupakan metode pembobotan *term* yang banyak digunakan sebagai metode pembanding terhadap metode pembobotan baru. Pada metode ini, perhitungan bobot *term* dalam sebuah dokumen dilakukan dengan mengalikan nilai *Term Frequency* dengan *Inverse Document Frequency.*

Term Frequency (TF) adalah faktor yang menentukan bobot *term* pada suatu dokumen berdasarkan jumlah kemunculannya dalam dokumen tersebut. Untuk menghitung bobot *term frequency* menggunakan persamaan (1):

$$w_{tf_{t,d}} = \begin{cases} 1 + \log 10\, tf_{t,d}, & if\ tf_{t,d} > 0 \\ 0 & otherwise \end{cases} \quad (1)$$

 P-ISSN:2356-3109 E-ISSN: 2356-3117

dimana $w_{tf_{t,d}}$ adalah bobot *term frequency* (*tf weight*), $tf_{t,d}$ adalah banyaknya kemunculan *term*/kata t dalam dokumen d.

Inverse Document Frequency (IDF) adalah pengurangan dominansi *term* yang sering muncul di berbagai dokumen. *Term* yang banyak muncul di berbagai dokumen, dapat dianggap sebagai *term* umum (*common term*) sehingga tidak penting nilainya. Kata yang muncul pada sedikit dokumen harus dipandang sebagai kata yang lebih penting (*uncommon term*) daripada kata yang muncul pada banyak dokumen. IDF dihitung menggunakan persamaan (2) (Zafikri 2008):

$$idf_t = \log_{10} N/df_t \qquad (2)$$

dimana idf_t adalah bobot *inverse document frequency*, N adalah banyaknya dokumen yang ada, dan *dft* adalah banyaknya dokumen yang mengandung term/ kata *t*.

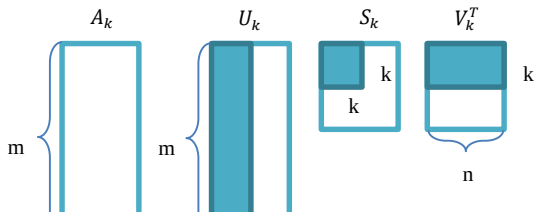
Kemudian rumus untuk menghitung bobot *tf-idf* untuk setiap kata *t* dalam dokumen d, merupakan hasil perkalian antara *tf weight* dengan *idf* dihitung menggunakan persamaan (3) (Schutze 2011):

$$w_{t,d} = w_{tf_{t,d}} \times idf_t \qquad (3)$$

dimana $w_{t,d}$ adalah bobot *tf-idf* kata t dalam dokumen d, $w_{tf_{t,d}}$ adalah bobot *term frequency* (*tf weigh*t), dan idf_t adalah bobot *inverse document frequency*.

3.3. *Singular Value Decomposition* (SVD)

Singular Value Decomposition (SVD) merupakan model matematis yang digunakan untuk reduksi dimensi data. Proses SVD dilakukan dengan mendekomposisi matriks



Gambar 1. Ilustrasi Matriks SVD

menjadi tiga bagian (Peter 2009), seperti pada gambar 1.

Matriks *U* dan *V* adalah matriks *othonormal*, dimana baris pada matriks *U* menggambarkan banyaknya baris pada matriks *A*, sementara kolom pada matriks *V* menggambarkan banyaknya kolom pada matriks *A*. *K-rank* digunakan untuk mereduksi dimensi dari matriks *A*. Matriks S merupakan matriks simetris yang berisi nilai positif di sepanjang diagonal, daerah selain diagonal berisi 0 (Sari 2012).

3.4. *Latent Semantic Indexing* (LSI)

Reduksi dari SVD digunakan dalam LSI. LSI merupakan salah satu bentuk teknik proses temu kembali dengan menggunakan *Vector Space Model* (VSM), untuk menemukan informasi yang relevan. Keterkaitan makna dalam LSI sifatnya tersembunyi. Fungsi matematis di dalam LSI mampu menemukan hubungan semantik antar kata (Sari 2012). Representasi dari LSI dapat dilihat pada persamaan (4):

$$q' = q^T.U_k.S_k^{-1} \qquad (4)$$

dimana *q'* adalah *query vector* representasi dari LSI, q^T adalah *transpose* dari pembobotan TF-IDF *query*, *Uk* adalah reduksi dimensi k dari matriks U, dan S_k^{-1} adalah inverse dari reduksi dimensi k matriks S (Sari 2012).

3.5. *Vector Space Model (VSM)*

Vector Space Model (VSM) adalah cara konvensional yang biasa digunakan dalam proses temu kembali informasi. Prosesnya dengan menghitung kemiripan dua buah vektor, yaitu antara vektor dari *corpus* dan vektor dari *query* (Kontostathis 2007). Untuk melakukan perhitungan terhadap kemiripan antar vektor digunakan rumus *Cosine Similarity* pada persamaan (5) (Parsons 2009):

$$\text{CosSim}(d_i,q) = \frac{d_i.q}{|d_i||q|} \qquad (5)$$

dimana d_i adalah dokumen vector ke i yang diambil dari nilai matriks V, q adalah kata kunci/query vector hasil perhitungan LSI.

3.6. *Mean Average Precision (MAP)*

Precision, recall, dan F-Measure merupakan kumpulan evaluasi untuk

mengetahui keakuratan sistem temu kembali secara *unranked retrieval*, atau dengan pengembalian dokumen tanpa perangkingan. Tipe evaluasi yang digunakan untuk mengevaluasi sistem temu kembali dengan *ranked retrieval* pada penelitian ini digunakan *Mean Average Precission* (MAP). *Average Precission* (AP) hanya mengambil nilai presisi dari dokumen-dokumen yang relevan dan kemudian hasilnya dibagi dengan jumlah dokumen yang dilibatkan (Strehl 2000). Pengukuran dari MAP merupakan hasil perhitungan rata-rata dokumen relevan yang *retrieved* dari setiap *query* yang terlibat di dalam sistem, sedangkan dokumen yang tidak relevan nilainya adalah 0 (Blanken 2007). Rumus dari Mean Average Precission pada persamaan (6) berikut (Manning 2009):

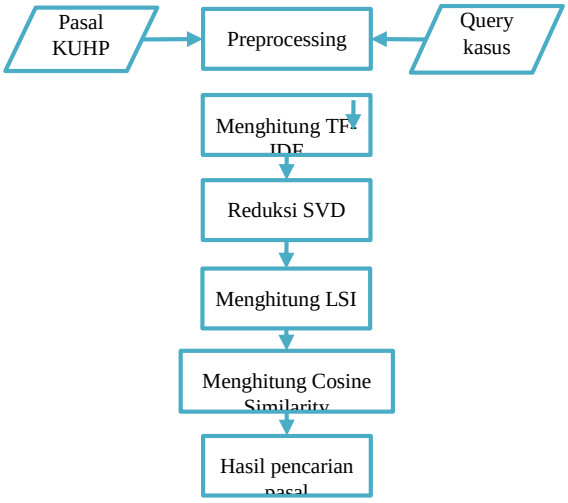
$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m}\sum_{k=1}^{m_j} Precision(R_{jk}) \quad (6)$$

dimana nilai Q merupakan kumpulan *query* atau menyatakan banyaknya *query* yang diinputkan dan R_{jk} adalah nilai *precission* dari kumpulan *file* pasal KUHP yang di *retrieved* oleh sistem dan relevan yang telah diranking. Nilai MAP mempunyai rentang nilai 0 sampai 1, dan dalam sebuah system dikatakan baik jika nilai MAP mendekati 1 (Manning 2009).

4. EKSPERIMEN

Penelitian dilakukan melalui langkah langkah yang diilustrasikan pada Gambar 2 sebagai berikut:

a. Memasukkan data berupa pasal-pasal KUHP Buku Kedua. Kumpulan pasal tersebut disebut sebagai *corpus*. Inputan sistem terdiri atas *corpus*(pasal KUHP) dan *query* (deskripsi kasus kejahatan).
b. *Preprocessing file corpus* dan *query*.
c. Membentuk struktur data *inverted index* pada *corpus.*
d. Membentuk matriks pembobotan TFIDF pada *corpus* dan *query*.
e. Mendekomposisi matriks pembobotan *corpus* dengan SVD.
f. Reduksi dimensi dari hasil dekomposisi matriks SVD dengan *k-rank*.
g. Menghitung *query vector* yang merupakan representasi dari LSI.
h. Mencari kemiripan antara *corpus* dan *query* dengan *cosine similarity*.
i. Pengurutan nilai *cosine similarity* secara *descending order*.
j. Pengambilan *top-n* teratas nilai *cosine similarity* hasil pengurutan.
k. Melakukan evaluasi dari hasil penelitian dengan *Mean Average Precission* (MAP). Data pasal dari pakar hukum mengenai pasal yang terkait dengan query (deskripsi kasus kejahatan) dibandingkan dengan hasil pencarian pasal sistem. Hasil pencarian sistem yang relevan antara *query* dan *corpus* adalah pasal yang sama dengan keputusan pakar hukum.



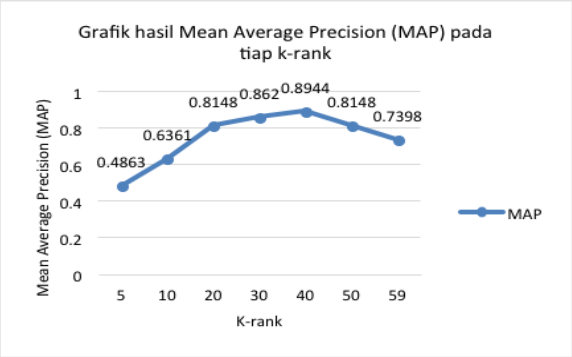
Gambar 2. Alur Kerja Sistem

5. HASIL DAN PEMBAHASAN

Data yang digunakan adalah pasal KUHP dengan jumlah pasal pada data training yaitu sebanyak 60 pasal, menggunakan deskripsi kasus kejahatan (*query*) sebanyak 6 kasus, yang terdiri dari berbagai macam kasus kejahatan yang terkait dengan pasal KUHP pada Buku Kedua tentang kejahatan, dengan nilai k-rank untuk masing-masing kasus yaitu 5, 10, 20, 30, 40, 50 dan 59.

P-ISSN:2356-3109 E-ISSN: 2356-3117

Analisa hasil dari nilai *Average Precision* (AP) dari masing-masing pengujian, tergantung dari nilai *k-rank* yang digunakan. Pada Gambar 3, menunjukkan bahwa akurasi sistem akan mengalami peningkatan jika nilai *k-rank* yang dimasukkan semakin besar, tetapi pada saat nilai *k-rank* melebihi nilai 40, performa sistem mengalami penurunan kembali.



Gambar 3. Grafik MAP pada tiap *k-rank*

Tabel 1 menunjukkan bahwa hasil dari evaluasi *Mean Average Precision* (MAP) pada masing-masing nilai *k-rank* yaitu 5, 10, 20, 30, 40, 50 dan 59. Pada Terlihat bahwa hasil *Mean Average Precision* (MAP) memiliki nilai yang rendah pada *k-rank* = 5 dengan nilai MAP = 0.4863. Sedangkan pada *k-rank* = 10, nilai MAP mengalami sedikit peningkatan, yaitu dengan nilai MAP = 0.6361. Terjadi peningkatan nilai MAP yang cukup signifikan yaitu dari MAP = 0.6361 pada *k-rank* =10, menjadi MAP = 0.8148 pada *k-rank*=20. Nilai MAP semakin membaik dan meningkat saat nilai semakin bertambah yaitu *k-rank* = 30, *k-rank* ini menghasilkan peningkatan nilai MAP yaitu dengan nilai MAP = 0.8620. Dari hasil analisis data diatas, nilai MAP pada *k-rank* 5, 10, 20, dan 30 kurang optimal. *K-rank* mereduksi dimensi fitur, semakin kecil nilai *k-rank* semakin besar data yang direduksi, sehingga pada *k-rank* 5, 10, 20, dan 30, banyak data atau informasi yang hilang akibat reduksi dimensi fitur. Reduksi dimensi fitur dapat menghilangkan *noice* pada data, tetapi juga dapat membuat informasi atau data yang dibutuhkan hilang.

Sistem menunjukkan peningkatan MAP yang terbaik dan tertinggi pada saat nilai *k-rank* bernilai 40 dengan nilai MAP yang hampir mendekati 1 dengan nilai MAP yaitu sebesar 0.8944. Saat nilai *k-rank* ditingkatkan menjadi 50 dan 59, sistem mengalami penurunan nilai MAP yang cukup signifikan yaitu menjadi 0.8148 dan 0.7398. Pada *k-rank* 50 dan 59, penambahan nilai *k-rank* memungkinkan untuk menghasilkan informasi yang lebih baik. Tetapi penambahan *k-rank* - dapat mengurangi keakurasian informasi yang diberikan, karena masih terdapat banyak *noise* yang terdapat pada data. Sehingga pada penelitian ini, *k-rank* dengan nilai 40, sistem memiliki nilai MAP paling optimal, sehingga sistem ini dapat mengembalikan kebutuhan informasi yang dibutuhkan dengan baik dan akurat.

Tabel 1. MAP pada masing-masing *k-rank*

Query	k-rank						
	5	**10**	**20**	**30**	**40**	**50**	**59**
Q1	0.5556	0.3333	0.3333	0.7556	1	0.9167	0.4667
Q2	0.4667	0.3333	0.5556	0.9167	0.9167	0.6389	0.6389
Q3	1	1	1	0.5	0.5	0.3333	0.3333
Q4	0.5	0.5	1	1	1	1	1
Q5	0.3333	1	1	1	1	1	1
MAP	**0.5711**	**0.6333**	**0.7778**	**0.8344**	**0.8833**	**0.7778**	**0.6878**

6. KESIMPULAN

Sistem yang dikembangkan dalam pencarian pasal KUHP berdasarkan kasus kejahatan pada penelitian ini menunjukkan hasil yang cukup baik, dimana nilai Mean Average Precision (MAP) yang dihasilkan mendekati nilai 1. Pada penelitian ini digunakan data berupa pasal pada Kitab Undang-Undang Hukum Pidana Buku Kedua tentang Kejahatan. Sehingga mengakibatkan banyaknya kata yang dihasilkan pada tahap *preprocessing* pasal. Jika dimensi antar jumlah kata dan banyaknya pasal semakin besar maka waktu komputasi yang dihasilkan juga semakin lama. Sehingga digunakan proses reduksi dimensi *Singular Value Decomposition* (SVD) yang dapat mengurangi jumlah dimensi. Proses *Latent Semantic Indexing* (LSI) yang menggunakan SVD juga digunakan untuk mencari keterkaitan makna antar kata tersembunyi. Proses matematis dalam SVD mampu menunjukkan hubungan semantik antar kata.

Pada penelitian ini *k-rank* dengan nilai 40, sistem memiliki nilai MAP paling optimal, sehingga sistem ini dapat mengembalikan kebutuhan informasi yang dibutuhkan dengan baik dan akurat. Tetapi pemilihan *k-rank* yang optimal tidak dapat ditentukan secara pasti karena banyaknya jumlah kata dan dokumen yang berbeda akan memungkinkan untuk menghasilkan *k-rank* optimal yang berbeda pula.

6. DAFTAR PUSTAKA

ATIQ ,DIAN ASTITATUL. 2015. *Indonesia Negara Hukum*. www.academia.edu/6234004/Indonesia_Negara_Hukum. 25/4/2015

BLANKEN, H., VRIES, ARJEN P.DE, BLOK, HENK ERNST, DAN FENG, LING ,.2007. *Multimedia Retrieval*.Springer Berlin Heidelberg New York

FELDMAN, RONEN AND JAMES SANGER. 2007. *The Text Mining Handbook*. Cambridge: Cambridge University Press. Cambridge.

KONTOSTATHIS, APRIL. 2007. *Essential Dimensions of Latent Semantic Indexing (LSI).* Departemen Matematika dan Ilmu Komputer Universitas Ursinus.USA

MANNING, CHRISTOPER.D, RAGHAVAN, PRABHAKAR, DAN SCHUTZE, H. 2009. *An Introduction to Information Retrieval*. Cambridge.

NUGROHO, EKO. 2011. *Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoritma Rabin-Karp*. Program studi Ilmu Komputer, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Brawijaya. Malang.

PARSONS, KATHRYN, MCCORMAC, A., BUTAVICIUS, M, DENNIS*,S, DAN FERGUSON, L. 2009. *The Use of Context-Based Information Retrieval Technique.* Australia. Defence Science and Technology Organization.

PETER, RAKESH, G, SHIVAPRATAP, DVYA G,DAN SOMAN KP. 2009. *Evaluation of SVD and NMF Fungsis for Latent Semantic Analysis*. India. Amrita University.

PRASETYO , TEGUH. 2014. *Hukum Pidana.* Jakarta: Rajawali Pers

SARI, YUITA ARUM. RIDOK, ACHMAD. MARJI. 2012. *Penentuan Lirik Lagu Berdasarkan Emosi Menggunakan Sistem Temu Kembali Informasi Dengan Metode Latent Semantic Indexing (Lsi)*. Teknik Informatika, Institut Teknologi Sepuluh Nopember (ITS) Dan Program Teknik Informatika dan Ilmu Komputer, Surabaya dan Malang.

SCHUTZE, HINRICH. 2011. Introduction to Information Retrieval. Institute for Natural Language Processing. University of Stuttgart : Jerman

STREHL, A, ET AL.2000.*Impact of Similarity Measures on Web-Page Clustering*. Proceeding of the Workshop of Artificial Intelligent for Web Search, 17th National Conference on Artificial Intelligence,2000.

ZAFIKRI, ATIKA. 2008. Implementasi Metode Term Frequency Inverse Document Frequency (TF-IDF) Pada Sistem Temu Kembali Informasi. Tugas Akhir Program Studi Ilmu Komputer Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Sumatera Utara : Medan