PENERAPAN ALGORITMA PROBABILISTICLATENT SEMANTIC INDEXING UNTUK MENDETEKSI DUPLIKASI JURNAL DI STIKI MALANG

Meivi Kartikasari
Program Studi Manajemen Informatika, STIKI - Malang
meivi.k@stiki.ac.id

Abstract

The impact of the development of information technology and computerization is to make easier for writers to compose a paper of scientific writing. This has an impact on the possibility of increased plagiarism. Such actions may harm the original author and plagiarist.
The manual plagiarism detection process is difficult because of the large number of manuscripts, so needed the system to detect plagiarism. To build a system that can detect plagiarism, certain algorithms are needed to measure the value of similarities between the documents compared. One of the algorithms used to detect similarities is Probabilistic Latent Semantic Analysis (PLSA). This method uses the term value reference and the result of singular value decomposition (SVD) to know the closeness between documents so that it can be known the value of similarity between documents.

Keywords: *Similarity, Probabilistic Latent Semantic Analysis,Singular Value Decompositio*

1.PENDAHULUAN

Seiring dengan berkembangnya teknologi informasi dan komputerisasi yang semakin pesat, maka proses pembuatan suatu karya penulisan dapat dilakukan dengan mudah dan cepat sehingga sangat rentan dalam melakukan tindakan *plagiarism*. Tindakan tersebut dapat memberikan kerugian bagi penulis asli dan pelaku plagiat karena melanggar Hak Kekayaan Intelektual (HKI). Untuk mencegah tindakan tersebut perludilakukan suatu cara untuk mengurangi tindakan plagiat.

Pencegahan dan pendeteksian dini merupakan cara yang dapat dilakukan untuk mengurangi plagiat. Untuk meminimalisasi praktik plagiat, diperlukan pendeteksian terhadap penulisan. Namun, proses pendeteksian secara manual sulit untuk dilakukan karena jumlah penulisan yang banyak, sehingga diperlukan system untuk mendeteksi plagiat. Untuk membangun suatu system yang dapat mendekteksi plagiat, diperlukan algoritma tertentu untuk mengukur nilai kemiripan antar dokumen yang dibandingkan.

Salah satu algoritma yang digunakan untuk mendeteksi kemiripan adalah *Probabilistic Latent Semantic Analysis* (PLSA). Metode ini menggunakan acuan nilai term dan nilai *singularvalued ecomposition* untuk mengetahui kedekatan antar dokumen sehingga dapat diketahui nilai kemiripan antar dokumen.

2. KAJIAN LITERATUR DAN PENGEMBANGAN HIPOTESIS

2.1 Plagiarisme

Plagiarisme merupakan tindakan penjiplakan atau pengambilan karangan, pendapat, dan sebagainya dari orang lain dan menjadikannya seolah karangan dan pendapat sendiri. Kegiatan plagiarism. Plagiarisme juga didefinisikan sebagai kegiatan dengan sengaja menyalin pemikiran atau kerja orang lain tanpa cara-cara yang sah. Plagiarisme adalah pencurian dan penggunaan gagasan atau tulisan orang lain (tanpa cara-cara yang

sah) dan diakui sebagai miliknya sendiri. Plagiarisme juga didefinisikan sebagai kegiatan dengan sengaja menyalin pemikiran atau kerja orang lain tanpa cara-cara yang sah. Pelaku plagiarisme dikenal juga dengan sebutan plagiat (Rosyidi, 2007).

2.2 *Information Retrieval*(IR)

Information Retrieval atau sistem temu balik informasi merupakan sistem yang mampu melakukan pencarian informasi pada kumpulan dokumen, pencarian dokumen itu sendiri, pencarian metadata untuk dokumen tersebut, atau pencarian teks, suara, gambar atau data dalam basis data dan pengambilan dokumen yang relevan dari sebuah koleksi dokumen sesuai dengan *query* pengguna sistem (Manning,Raghavan,&Schütze,2009).
Input dari suatu sistem temu balik informasi adalah *query* dari pengguna dan koleksi dokumen, dan *output*-nya adalah dokumen yang dianggap relevan oleh sistem. Sistem temu balik informasi ini digunakan untuk mengurangi informasi yang terlalu banyak sehingga sulit untuk dikelola. Dengan adanya sistem temu balik informasi maka diharapkan pencarian infomasi dapat dilakukan dengan efektif dan memberikan hasil pencarian yang tepat.

2.3 Ekstraksi Dokumen

Teks yang akan dilakukan proses teksmining,pada umumnya memiliki beberapa karakteristik diantaranya adalah memiliki dimensi yang tinggi, terdapat *noise* pada data dan terdapat struktur teks yang tidak baik. Cara yang digunakan dalam mempelajari suatu data teks adalah dengan terlebih dahulu menentukan fitur-fitur yang mewakili setiap kata untuk setiap fitur yang ada pada dokumen. Sebelum menentukan fitur-fitur yang mewakili,diperlukan tahap *preprocessing* yang dilakukan secara umum dalam teks mining pada dokumen,yaitu *tokenizing, filtering, stemming, tagging* dan *analyzing* (Triawati,2009).

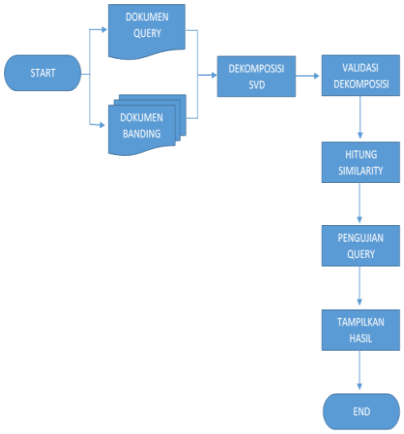
Gambar 1 Tahapan *Preprocessing*

2.4 Algoritma *Latent Semantic Indexing*

Latent Semantic Indexings(LSI) adalah metode pengindeksan dan pencarian yang menggunakan teknik matematika yang disebut Dekomposisi Nilai Singular (SVD) untuk mengidentifikasi pola hubungan antar istilah dan konsep-konsep yang terkandung dalam sebuah koleksi teks yang tidak terstruktur. LSI didasarkan pada prinsip bahwa kata-kata yang digunakan dalam konteks yang sama cenderung memiliki makna yang sama (Pavel,2011).

Hasil pencarian yang sesuai dengan kebutuhan dalam suatu koleksi dokumen yang besar merupakan hal sulit.Usaha pengguna secara manual untuk memilah-milah dokumen yang sesuai dengan kebutuhannya ternyata sangat besar. Hasil pencarian merupakan sejumlah dokumen yang relevan menurut sistem, namun relevansi merupakan hal yang subjektif.
Pada umumnya, dokumen dikatakan relevan dengan *query* apabila dokumen :

1. Memuat kata atau kalimat yangsama dengan *query*
2. Memuat kata atau kalimat yang bermakna sama dengan *query*



Gambar 2 Alur Proses Algoritma Latent Semantic Indexing

3. METODE PENELITIAN

3.1 Analisa Masalah

Semakin mudah pertukaran informasi dewasa ini tidak hanya membawa dampak positif bagi kemajuan teknologi, tetapi juga membawa dampak negatif yang hampir tidak dapat dihindari yaitu *plagiarism*.

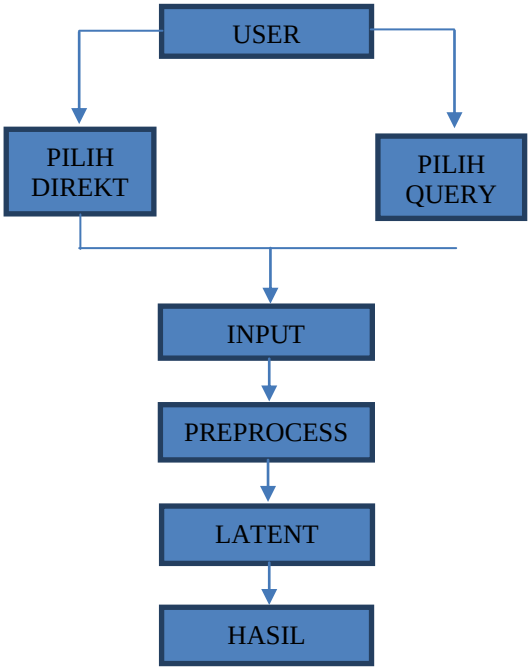
Berdasarkan permasalahan tersebut maka pada penelitianini akan dirancang suatu sistem yang dapat mengukur tingkat similaritas pada sebuah naskah jurnal. Teknik yang digunakan adalah dengan membandingkan kemiripan jurnal dengan dokumen banding jurnal yang didapat dari digital library STIKI Malang. Sebelum adanya aplikasi pengukuran tingkat similaritas dokumen ini, proses pengecekan dokumen membutuhkan waktu yang cukup lama karena harus membaca naskah dalam bentuk *hardcopy*. Dengan adanya aplikasi ini, sekumpulan file dokumen dapat diuji apakah antar dokumen memiliki kesamaan atau tidak, sehingga diharapkan praktek plagiarisme dapat dicegah.

3.2 Analisa Data

Data pengujian yang digunakan dalam penelitian ini adalah dokumen jurnal informatika di STIKI Malang dari perpustakaan STIKI Malang.

3.3 Deskripsi Sistem

Sistem yang dibangun ini adalah sistem untuk mengenali plagiat dokumen yang didalamnya dilakukan proses *text mining*. Tahapan dalam sistem plagiat dokumen ini ada 2, yaitu *preprocessing dan* analisa. Dari gambar dibawah ini (Gambar 3.1) dapat dilihat bahwa sistem memerlukan dua buah *input* yang pertama berupa dokumen yang akan diproses dan yang kedua berupa queryatau dokumen acuan yang akan dibandingkan kemudian akan dilakukan *preprocessing* dengan menggunakan langkah *text mining* dan dilakukan pembobotan, kemudian kedua input tersebut akan dibandingkan sehingga menghasilkan bobot yang baru yang akhirnya bobot tersebut akan digunakan sebagai acuan untuk menampilkan dokumen yang berhubungan dengan kata kunci.



Gambar 3 AlurProses Sistem

Dari proses tersebut didapatkan informasi apakah perbandingan antara dokumen *query* tersebut dapat dikatakan sebagai plagiat atau tidak, jika dokumen yang diuji mempunyai nilai LSI di atas 0.5 ke atas maka dapat di katakan bahwa dokumen tersebut adalah plagiat.

4. HASIL DAN PEMBAHASAN

4.1 Implementasi Sistem

Sesuai dengan perancangan sistem yang dibahas sebelumnya, pada sistem disediakan*interface* menu utama untuk memilih direktori dokumen dan dokumen query, menampilkan hasil preprocessing, nilai SVD dan nilai peringkat dokumen.

4.1.1 Tampilan Aplikasi

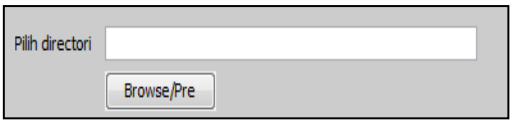
Adapun tampilan aplikasi dapat dilihat dalam Gambar 4

Pada tampilan menu diatas terdiri dari menu *preprocessing*, menu pilih direktori yaitu letak direktori dokumen query dan direktori dokumen banding. Setelah itu terdapat menu pilih query untuk menentuan dokumen query acuan yang ingin dibandingkan. Kemudian dilakukan preprocessing dokumen sehingga dihasilkan nilai term dokumen yang ditampilkan pada menu tab preprosesing. Selanjutnya dilakukan perhitungan nilai SVD dan peringkat dengan menekan button proses.

4.1.2 Menu Aplikasi

a. Pilih Direktori

Dokumen menu pilih direktori adalah menu yang berfungsi untuk memilih direktori dokumen yang digunakan sebagai dokumen query dan dokumen banding.

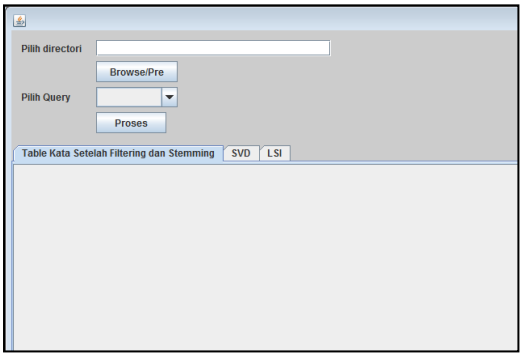


Gambar 5 Menu Pilih Direktori

b. Pilih Query

Setelah user memilih direktori dokumen, selanjutnya user memilih query dokumen yang digunakan sebagai acuanuntuk dibandingkan dengan dokumen yang terdapat pada direktori yang dipilih.

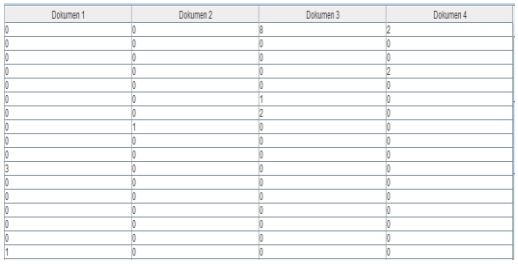
Gambar 6 Menu PilihQuery



Gambar 4 Tampilan aplikasi

c. Hasil Term Document

Hasil *preprocessing* menghasilkan tampilan list kata dan term dokumen dari masing-masing kata dalam direktori dokumen. Hasil term dokumen ditampilkan sebagai berikut.

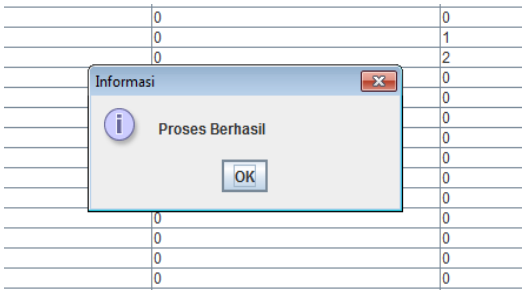


Gambar 7 Hasil Term Document

Hasil term dokumen menunjukkan nilai kemunculan kata dalam seluruh dokumen yang digunakan.

d. Nilai SVD

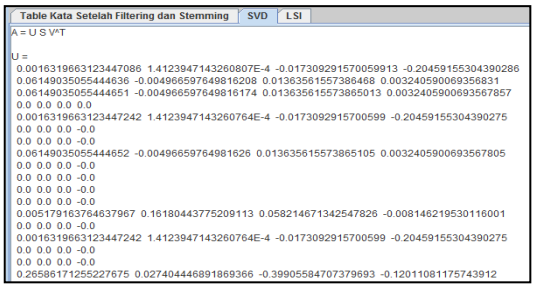
Nilai SVD dan hasil peringkat ditampilkan setelah user menekan button proses. Setelah *user* menekan button proses , muncul notifikasi bahwa proses berhasil.



Gambar 8 Informasi Proses Berhasil

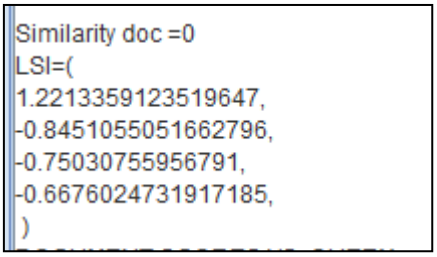
Informasi tersebut menunjukkan bahwa proses perhitungan dengan menghasilkan nilai SVD dan peringkat

telah berhasil dilakukan. Hasil dari proses tersebut ditampilkan pada tab menu SVD dan tab menu peringkat seperti tampak dalam Gambar 9



Gambar 9 Hasil Nilai SVD

Dari hasil nilai SVD diatas ditampilkan nilai matriks U, S dan V yang merupakan hasil perhitungan *singular value decomposition* dari term dokumen A.



Gambar 10 Hasil Nilai Similarity

Hasil peringkat diatas menampilkan nilai similarity hasil perbandingan dokumen query dengan dokumen banding. Dari hasil diatas dapat diketahu bahwa perbandingan yang dilakukan antara dokumen abstrak kampus (Dokumen 0) memilki kemiripan dengan dokumen abstrak banding (Dokumen 1) dengan nilai similarity dokumen 0 = 1.2213359 dan Dokumen 1 = 0.845105505.

Dokumen 1 = 0.845105505.

4.2 Pengujian Sistem

Hasil pengujian yang dilakukan menggunakan 6 dokumen yang terdiri dari 3 dokumen banding (abstrak kampus) dan 3 dokumen banding (abstrak banding) adalah sebagai berikut :

a. Dokumen 1

Tabel 1 Pengujian Dokumen 1

	Query : Dokumen 1	
No	Dokumen	Hasil
1	Dokumen 2	0.2919559357
2	Dokumen 3	0.2682159183
3	Dokumen 1	0.2588267211

Dari hasil pengujian dokumen 1 diatas dapat dilihat bahwa dokumen banding yang memilki nilai *similarity* tertinggi adalah dokumen 2 dengan nilai *similarity* 0.2919559357.

a. Dokumen 2

Tabel 2 Pengujian Dokumen 2

	Query : Dokumen 2	
No	Dokumen	Hasil
1	Dokumen 2	2.3957673463
2	Dokumen 1	1.5663860417
3	Dokumen 3	1.3660796411

Dari hasil pengujian dokumen 2 diatas dapat dilihat bahwa dokumen banding yang memilki nilai *similarity* tertinggi adalah dokumen 2 dengan nilai *similarity* 2.3957673463.

b. Dokumen 3

Tabel 3 Pengujian Dokumen 3

	Query : Dokumen 3	
No	Dokumen	Hasil
1	Dokumen 1	1.5435150822
2	Dokumen 3	0.9907814741
3	Dokumen 2	0.4139785256

Dari hasil pengujian dokumen 3 diatas dapat dilihat bahwa dokumen banding yang memilki nilai *similarity* tertinggi adalah dokumen 1 dengan nilai similarity 1.5435150822.

4.3 Hasil Pengujian Recall-Precision

a. *Matriks recall-precision* dari pengujian 1, pengujian 2 dan pengujian 3 adalah sebagai berikut :

Tabel 4 Pengujian Recall-Precision

	Relevan	Tidak relevan	Total
Memiliki Kemiripan	2	0	3
Tidak	0	1	0
Total	2	1	3

4.3 Hasil Pengujian Recall-Precision

b. *Matriks recall-precision* dari pengujian 1, pengujian 2 dan pengujian 3 adalah sebagai berikut :

Tabel 4 Pengujian Recall-Precision

	Relevan	Tidak relevan	Total
Memiliki Kemiripan	2	0	3
Tidak	0	1	0
Total	2	1	3

Dari hasil pengujian recall precision diatas dapat dilihat bahwa dari 3 pengujian diatas terdapat 2 dokumen yang sesuai dengan hasil pengujian *similarity* dan 1 dokumen yang tidak sesuai. Pengujian recall precision menunjukkan hasil sebagai berikut:

Recall : 2/2 x 100% = 100%

Precision : 2/3 x 100% = 66.7%

5. KESIMPULAN

5.1 Kesimpulan

a. Dari hasil pengujian query yang dilakukan pada dokumen kampus dan dokumen banding dapat disimpulkan bahwa masing-masing dokumen kampus memiliki tingkat kemiripan berbeda-beda dengan dokumen banding.

b. Berdasarkan pengujian sistem yang dilakukan, diperoleh nilai precision dan recall dari pengujian 1 , pengujian 2 dan pengujian 3 yaitu mencapai nilai recall 100% dan nilai precision 66,7%. Hal ini disebabkan karena metode latent semantic indexing mengambil nilai semantik pada dokumen sehingga masih terdapat kemungkinan dokumen query dan dokumen banding tidak relevan.

5.2 Saran

a. Pada setiap pengujian sebaiknya terdapat minimal 10 dokumen sehingga dapat meningkatkan nilai recall dan precision.

b. Data *stoplist* dan *wordlist* cukup berpengaruh pada proses *text mining*, semakin lengkap data tersebut maka proses mining akan menghasilkan informasi yang tepat.

c. Perlunya suatu sistem yang secara otomatis menampilkan hasil term dokumen , nilai SVD dan nilai peringkat secara bersamaan.

6. REFERENSI

[1] Abdul Kadir, 2008. *Dasar Pemrograman Java 2*, Penerbit Andi Yogyakarta

[2] Devita,2012. *Temu Kembali Informasi dengan keyword,* Universitas Airlangga

[3] Roger,2011. *Why LSI? Latent Semantic Indexing and Information Retrieval.* Agilex Technologies, Inc., Chantilly, Virginia

[4] Pavel,2011. *Latent Semantic Indexing for Image Retrieval Systems.* Society for Industrial and Applied Mathematics

[5] Wahana,2009. *Menguasai Java Programming,* Penerbit Salemba Empat

[6] Yureska,2011. *Algoritma Pencocokan Objek Geometri Citra Berbasis Graph Untuk Pemilihan Kembali (Retrieval).*Jurusan Teknik Elektro-FTI, Institut Teknologi Sepuluh Nopember

[7] Zoran,2011.*Information Retrieval using Latent Semantic Indexing.* Institute of Technology Lausanne