PENGEMBANGAN SISTEM PENDETEKSI KEMIRIPAN KARYA PADA INAICTA 2013

Cadea Mikha Pasma[1], Ulla Delfana Rosiani,ST.,MT[2], Rudy Ariyanto,ST. MCs[3]
[1,2] Jurusan Teknik Elektro, Program Studi Teknik Informatika, Politeknik Negeri Malang
[1]cadeamikha@gmail.com, [2]ullarosi@gmail.com, [3] ariyantorudi@gmail.com

Abstrak

Indonesia ICT Award (INAICTA) 2013 merupakan ajang lomba karya cipta kreativitas dan inovasi di bidang TIK (Teknologi Informasi dan Komputer) terbesar di Indonesia yang bertujuan untuk terus mendorong berkembangnya produk-produk TIK (Teknologi Informasi dan Komputer) lokal dengan peningkatan kualitas maupun inovasi produk. Semakin tahun, jumlah kontestan yang mengikuti INAICTA semakin bertambah. Hal tersebut berpengaruh terhadap tingkat kesulitan bagi para juri atau tim penilai untuk mengetahui kemiripan dari inovasi-inovasi para kontestan. Dibutuhkan suatu aplikasi yang dapat membantu dalam pendeteksian kemiripan tiap hasil karya yang diikutsertakan oleh para kontestan. Oleh karena itu dilakukan pengembangan sistem pendeteksi kemiripan karya pada INAICTA 2013 dengan membandingkan penjelasan ringkas karya para kontestan. Dalam pengembangan sistem pendeteksi ini menggunakan Algoritma *Term Frequency– Inversed Document Frequency* (*TF-IDF*) untuk proses pembobotan karya. Dengan *TF-IDF* sistem akan menghitung berdasarkan *term* pada setiap karya. Sedangkan untuk melihat tingkat kedekatan atau kesamaan *(similarity)* karya, sistem ini menggunakan Algortitma *Vector Space Model* (*VSM*). Dengan *VSM* data karya dipandang sebagai sebuah vektor yang memiliki *magnitude* (jarak) dan *direction* (arah). Sehingga sistem pendeteksi kemiripan karya pada INAICTA 2013 ini akan menghasilkan urutan tingkat kemiripan karya INAICTA 2013.

Katakunci: pendeteksi kemiripan, *Term Frequency-Inversed Document Frequency* (*TF-IDF*)*, Vector Space Model* (*VSM*).

7. Pendahuluan

Kemajuan teknologi pada saat ini mendorong semua kalangan untuk mencari dan memberikan informasi dengan cepat dan mudah. Berbagai kegiatan yang dulu dilakukan secara manual pun sekarang dapat dilakukan secara online dan sistematis. Hal tersebut mendorong orang untuk mendapatkan dan menyalurkan berbagai informasi dan inovasi. Tidak salah jika saat ini banyak ditemui pemikiran-pemikiran baru yang mempunyai nilai guna tinggi. Saat ini banyak ditemui hal-hal baru, yang awalnya dianggap hanya hal yang tidak mungkin, namun dengan berkembangnya ilmu teknologi hal-hal tersebut dapat menjadi mungkin. Saat ini inovasi yang sangat berguna untuk membantu kehidupan manusia banyak dicari. Dengan berbagai cara, para ahli mencari temuan-temuan dan ciptaan-ciptaan baru yang diluar pemikiran manusia pada umumnya. Dari latar belakang tersebut, terciptalah suatu kompetisi online yang bergerak dibidang inovasi yaitu Indonesia ICT Award (INAICTA).

Indonesia ICT Award (INAICTA) sudah diadakan sejak tahun 2007 yang merupakan ajang lomba karya cipta kreativitas dan inovasi di bidang TIK terbesar di Indonesia. Tujuan diselenggarakannya INAICTA adalah untuk mendorong terus berkembangnya produk-produk TIK lokal dengan peningkatan kualitas maupun inovasi produk. Tidak hanya bagi pengembang individu, tapi juga bagi komunitas untuk pemberdayaan masyarakat (www.inaicta.web.id/inaicta/). Ajang lomba ini dapat diikuti oleh kalangan pelajar (SD, SMP, SMA/SMK), mahasiswa perguruan tinggi, maupun kalangan profesional (www.inaicta.web.id/lomba). Untuk dapat berpartisipasi dalam INAICTA, para kontestan harus memiliki Hak Atas Kekayaan Intelektual dari tiap karya yang diikutsertakan (www.inaicta.web.id/lomba/syarat-dan-ketentuan).

Semakin tahun, jumlah kontestan yang mengikuti INAICTA semakin bertambah. Hal tersebut berpengaruh terhadap tingkat kesulitan bagi para juri atau tim penilai untuk mengetahui kemiripan dari inovasi-inovasi dari para kontestan. Dibutuhkan suatu aplikasi yang dapat membantu dalam pendeteksian kemiripan tiap hasil karya yang diikutsertakan oleh para kontestan. Algoritma *Term Frequency – Inversed Document Frequency* (*TF-IDF*) serta *Vector Space Model* digunakan dalam pengembangan sistem pendeteksi kemiripan

karya pada INAICTA dengan menggunakan kata kunci dari tiap dokumen penjelasan ringkas karya

Dasar Teori

7.1 *Text Mining*

Text mining adalah suatu proses menggali informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen yang menggunakan *tools* analisis yang merupakan komponen-komponen dalam *data mining* yang salah satunya adalah kategorisasi. Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen sehingga sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur (Triawati,2009:1).

7.2 Vector Space Model (VSM)

Vector Space Model (VSM) metode untuk melihat tingkat kedekatan atau kesamaan term dengan cara pembobotan term. Dokumen dipandang sebagai sebuah vektor yang memiliki jarak dan arah.

Proses penghitungan VSM dilakukan dengan dengan beberapa langkah perhitungan, antara lain yang diawali dengan persamaan (1)

$$tf = tf_{ij} \quad (1)$$

Dengan *tf* adalah *term frequency,* dan tf_{ij}adalah banyaknya kemunculan *term* t_i dalam dokumen d_j, *Termfrequency (tf)* dihitung dengan menghitung banyaknya kemunculan term t_i dalam dokumen d_j.

Perhitungan *Inverse Document Frequency (idf),* menggunakan persamaan (2)

$$idf_i = log\frac{N}{dfi} \quad (2)$$

Dengan idf_i adalah *inverse document frequency, N* adalah jumlah dokumen yang terambil oleh sistem, dan df_i adalah banyaknya dokumen dalam koleksi dimana $termt_i$ muncul di dalamnya, maka perhitungan idf_i digunakan untuk mengetahui banyaknya *term* yang dicari (df_i) yang muncul dalam dokumen lain yang ada pada *database*.

Perhitungan *termfrequency Inverse Document Frequency (tfidf)*, menggunakan persamaan (3)

$$W_{ij} = tf_i \,.\, idf_i \quad (3)$$

Dengan W_{ij}adalahbobot dokumen. Bobot dokumen (W_{ij}) dihitung untuk didapatkannya suatu bobot hasil perkalian atau kombinasi antara *termfrequency* (tf_i) dan *inverse document frequency* (idf_i).

7.3 Vector Space Model

Vector Space Model (VSM) adalah metode untuk melihat tingkat kedekatan atau kesamaan *(similarity)* term dengan cara pembobotan *term*. Dokumen dipandang sebagi sebuah vektor yang memiliki *magnitude* (jarak) dan *direction* (arah). Pada *Vector Space Model,* sebuah istilah direpresentasikan dengan sebuah dimensi dari ruang vektor. Relevansi sebuah dokumen ke sebuah *query* didasarkan pada similaritas diantara vektor dokumen dan vektor *query* (Baeza, 1999) (Amin, 2012).

VSM memberikan sebuah kerangka pencocokan parsial adalah mungkin. Hal ini dicapai dengan menetapkan bobot non-biner untuk istilah indeks dalam query dan dokumen. Bobot istilah yang akhirnya digunakan untuk menghitung tingkat kesamaan antara setiap dokumen yang tersimpan dalam sistem dan permintaan user.Dokumen yang terambil disortir dalam urutan yang memiliki kemiripan, model vektor memperhitungkan pertimbangan dokumen yang relevan dengan permintaan user.Hasilnya adalah himpunan dokumen yang terambil jauh lebih akurat (dalam arti sesuai dengan informasi yang dibutuhkan oleh *user*) (Amin, 2012).

Perhitungan jarak *query* menggunakan persamaan (4) dan untuk dokumen menggunakan persamaan (5).

$$|q| = \sqrt{\sum_{j=1}^{t}(W_{iq})^2} \quad (4)$$

Dengan |q| adalah jarak *query* dan W_{iq}adalah bobot *query* dokumen ke-i, maka jarak *query* |q| dihitung untuk didapatkan jarak *query* dari bobot *query* dokumen (W_{iq}) yang terambil oleh sistem. Jarak *query* bisa dihitung dengan persamaan akar jumlah kuadrat dari *query* (Amin, 2012).

$$|d_j| = \sqrt{\sum_{i=1}^{t}(W_{ij})^2} \quad (5)$$

Dengan $|d_j|$ adalah jarak dokumen dan W_{ij} adalah bobot dokumen ke-i, maka jarak

dokumen ($|d_j|$) dihitung untuk didapatkan jarak dokumen dari bobot dokumen dokumen (W_{ij}) yang terambil oleh sistem. Jarak dokumen bisa dihitung dengan persamaan akar jumlah kuadrat dari dokumen (Amin, 2012).

Perhitungan pengukuran *similarity query document* (*inner product*), menggunakan persamaan (6)

$$\boldsymbol{sim(q, d_j) = \sum_{i=1}^{t} W_{iq} . W_{ij}} \quad \textit{(6)}$$

Dengan W_{ij} adalah bobot *term* dalam dokumen, W_{iq} adalah bobot *query*, dam $sim(q,d_j)$ adalah similaritas antara *query* dan dokumen. Similaritas antara *query* dan dokumen atau *inner product* / $sim(q,d_j)$ digunakan untuk mendapatkan bobot dengan didasarkan pada bobot *term* dalam dokumen (W_{ij}) dan bobot *query* (W_{iq}) atau dengan cara menjumlahkan bobot *q* dikalikan dengan bobot dokumen (Amin, 2012).

Pengukuran *cosine similarity* (menghitung nilai kosinus sudut antara dua vektor) menggunakan persamaan berikut

$$\boldsymbol{sim(q, d_j) = \frac{q.d_j}{|q| * |d_j|} = \frac{\sum_{i=1}^{t} W_{iq}.W_{ij}}{\sqrt{\sum_{j=1}^{t}(W_j)^2} * \sqrt{\sum_{i=1}^{t}(W_i)^2}}}$$

Similaritas antara *query* dan dokumen $sim(q,d_j)$ berbanding lurus terhadap jumlah bobot *query* (q) dikali bobot dokumen (d_j) dan berbanding terbalik terhadap akar jumlah kuadrat q ($|q|$) dikali dengan akar jumlah kuadrat dokumen ($|d_j|$). Perhitungan similaritas menghasilkan bobot dokumen yang mendekati nilai 1 atau menghasilkan bobot dokumen yang lebih besar dibandingkan dengan nilai yang dihasilkan dari perhitungan *inner product* (Amin, 2012).

8. Metode Rancangan Sistem

8.1 Langkah dalam pengimplementasian ke dalam *Vector Space Model*

a. *Tokenizing*
Proses ini memotong setiap kata dalam teks, dang mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf "a" sampai "z" yang diterima, sedagkan karakter selain huruf dihilangkan. Jadi hasil dari proses *tokenzing* adalah kata-kata yang merupakan penyusun kalimat/string yang dimasukkan.

b. *Filtering*
Pada tahap ini dilakukan proses *filter* atau penyaringan kata hasil dari proses *tokenizing*, dimana kata yang tidak relevan dibuang. Proses ini menggunakan pendekatan *stoplist*. Yang termasuk *stoplist* adalah "yang", "dari", "di", dan lain-lain.

c. *Stemming*
Stemming merupakan proses untuk menghubungkan atau memecahkan setiap varian-varian menjadi suatu kata dasar. *Stem* (akar kata) adalah bagian dari akar yang tersisa setelah dihilangkan imbuhannya (awalan dan akhiran) (Hatta, Ramadijanti & Helen, 2010).

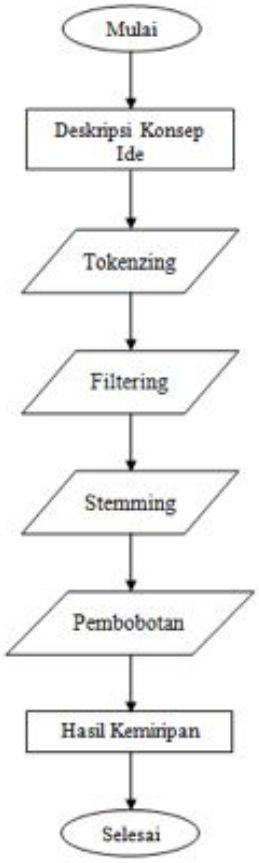
d. Pembobotan
Dalam tahap ini dilakukan dengan dua cara, yaitu perhitungan menggunakan *TF IDF*. Kemudian menggunakan *Vector Space Model* yang digunakan untuk mendapatkan bobot pada tiap karya.

8.2 Analisis data

Pada tahap ini dilakukan pengumpulan fakta-fakta yang mendukung perancangan sistem dengan mengadakan konsultasi dengan pakar dan membandingkan hasil penelitian dengan yang ada pada buku penuntun perancangan *Flowchart* dan penentuan bentuk *tree*.

8.3 Flowchart

Flowchart merupakan bagan yang memperlihatkan urutan dan hubungan antar proses beserta instruiksinya. Gambaran ini dinyatakan dengan simbol, setiap simbol menggambarkan proses tertentu.



Gambar (1) Flowchart Pendeteksi Kemiripan

9. Hasil

Hasil implementasi dari pencarian kemiripan menggunakan metode *TF-IDF* dan *VSM* adalah peringkat atau rangking nilai tertinggi sampai terendah suatu kemiripan data terhadap data pembanding.

10. Kesimpulan dan saran

10.1 Kesimpulan

Dari pembahasan dan analisa di atas, maka penulis menyimpulkan bahwa:

1. Dalam pengembangan sistem pendeteksi kemiripan karya pada INAICTA menggunakan algoritma *Term Frequency – Inversed Document Frequency* (TF-IDF) sebagai penentu bobot tiap karya dan *Vector Space Model* sebagai penentu peringkat urutan kemiripan karya.
2. Dari perhitungan meggunakan algoritma *Term Frequency – Inversed Document Frequency* (TF-IDF) dan *Vector Space Model*, maka pengguna atau tim penilai dapat mengetahui tingkat kemiripan karya pada INAICTA 2013 berdasarkan jumlah kemunculan kata yang mirip.

10.2 Saran

Pada penelitian pengembangan sistem pendeteksi kemiripan karya ini terdapat beberapa saran sebagai berikut:

1. Penggunaan algoritma *Term Frequency – Inversed Document Frequency* (TF-IDF) dan *Vector Space Model* perlu dikembangkan lagi agar digunakan untuk mendeteksi kemiripan dokumen lain.
2. Proses *stemming* perlu dikembangkan untuk kata-kata yang masih belum termasuk kata dasar seperti, mempertanggungjawabkan, menganakemaskan dan sejenisnya.

Daftar Pustaka:

www.inaicta.web.id
Triawati, Candra: Text Mining, 2009
Herwansyah Adhit: Aplikasi Pengkategorian Dokumen dan PengukuranTingkat Similaritas Dokumen Menggunakan Kata KunciPada Dokumen Penulisan IlmiahUniversitas Gunadarma, 2009
Agusta Ledy: Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia, November 14, 2009
Amin Fatkhul: Sistem Temu Kembali Informasi dengan Metode Vector Space Model ,2012
Khairunnisa Nova, Syarif Dadang SS, Wibowo Ardianto : Aplikasi Pendeteksi Plagiat dengan Menggunakan Metode *Latent Semantic Analysis* (Studi Kasus : Laporan TA PCR)", PCR.Vol 1.95.2012
Subekti Agus, Susanto Hendri Murti:E-library dengan fungsi searching menggunakan text mining Vol 3 No. 1 - 2013