PERANCANGAN APLIKASI PENGUKURAN SIMILARITAS PADA ARTIKEL DENGAN ALGORITMA TF-IDF DI CV. DIGIMEDIA CORPINDO

SKRIPSI

Diajukan Untuk Penulisan Skripsi Guna Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer (S.Kom)
Program Studi Teknik Informatika FT UN PGRI Kediri

OLEH :

REFI RIDHO SEPTIAN
NPM : 12.1.03.02.0077

FAKULTAS TEKNIK (FT)

UNIVERSITAS NUSANTARA PERSATUAN GURU REPUBLIK INDONESIA

UN PGRI KEDIRI

2016

Skripsi oleh :

REFI RIDHO SEPTIAN

NPM : 12.1.03.02.0077

Judul :

"PERANCANGAN APLIKASI PENGUKURAN SIMILARITAS PADA ARTIKEL DENGAN ALGORITMA TF-IDF DI CV. DIGIMEDIA CORPINDO "

Telah disetujui untuk diajukan kepada

Panitia Ujian/Sidang Skripsi

Program Studi Teknik Informatika FT UN PGRI Kediri

Tanggal : 11Agustus 2016

Pembimbing I	Pembimbing II
Juli Sulaksono, Ir. M.Kom., M, M.	**Ratih Kumalasari N, S. ST., M.Kom.**
NIDN. 0707076505	NIDN. 0710018501

ii

Skripsi oleh :

REFI RIDHO SEPTIAN

NPM : 12.1.03.02.0077

Judul :

"PERANCANGAN APLIKASI PENGUKURAN SIMILARITAS PADA ARTIKEL DENGAN ALGORITMA TF-IDF DI CV. DIGIMEDIA CORPINDO "

Telah dipertahankan di depan Panitia Ujian/Sidang Skripsi
Program Studi Teknik Informatika FT UN PGRI Kediri

Pada Tanggal : 11Agustus 2016

Panitia Penguji :

1. Ketua Penguji: **Juli Sulaksono, Ir. M.Kom., M, M.** _____
2. Penguji I : **Ardi Sanjaya, M.Kom.** _____
3. Penguji II : **Ratih Kumalasari N, S. ST., M.Kom.** _____

Mengetahui,
Dekan Fakultas Teknik

Dr. Suryo Widodo, M.Pd
NIP. 19640202 199103 1 002

iii

PERANCANGAN APLIKASI PENGUKURAN SIMILARITAS PADA ARTIKEL DENGAN ALGORITMA TF-IDF DI CV. DIGIMEDIA CORPINDO

Refi Ridho Septian
12.1.03.02.0077
Teknik - Informatika
refiridho@ymail.com
Juli Sulaksono, Ir. M.Kom., M, M. dan Ratih Kumalasari N, S. ST., M.Kom.
UNIVERSITAS NUSANTARA PGRI KEDIRI

ABSTRAK

Penelitian ini dilatarbelakangi oleh hasil pengamatan peneliti pada CV. Digimedia Corpindo yang masih melakukan penyeleksian artikel dari calon content writer secara manual. Hal ini sangatlah tidak efektif untuk mendapatkan calon content writer yang kompeten, pada akhirnya artikel yang dibuat oleh content writer baru kurang efektif untuk meningkatkan traffik website.

Permasalahan penelitian ini adalah (1) Bagaimana menerapkan metode text mining untuk menentukan similarity artikel yang dibuat oleh calon content writer? (2) Bagaimana meranking artikel yang dibuat oleh calon content writer sesuai dengan kriteria perusahaan?. Dengan menggunakan text mining dapat memberikan alternatif dari permasalahan yang terjadi di CV. Digimedia Corpindo.

Kesimpulan hasil penelitian ini adalah dengan menggunakan metode TF-IDF dapat memberikan alternative untuk memutuskan calon content writer yang terbaik melalui pengukuran similaritas artikel sesuai dengan topik berita yang dibuat.

Kata kunci *: text mining, tfidf, similaritas artikel.*

I. Latar Belakang

Startup atau disebut perusahaan rintisan merujuk pada semua perusahaan yang belum lama beroperasi. Perusahaan-perusahaan ini sebagian besar merupakan perusahaan yang baru didirikan dan berada dalam fase pengembangan dan penelitian untuk menemukan pasar yang tepat. Definisi lebih konkrit dijelaskan oleh *David McClure* yang disetujui oleh 1.500 masyarakat *Quora* bahwa *Startup* adalah perusahaan yang belum menemukan target konsumen, produk & cara menjualnya.

CV. Digimedia Corpindo adalah salah satu contoh startup yang sedang berkembang di bidang web blog. CV. Digimedia Corpindo yang beralamatkan di bangsongan Kabupaten Kediri ini mengelola banyak *web blog* seperti *smeaker.com, newsth.com dan oketekno.com*. Tiga website tersebut adalah webstie utama yang dikelolanya. Keuntungan yang didapat dari *web blog* tersebut adalah dari *Google Adsense* dan iklan-iklan lokal lainnya. Dengan mempekerjakan 15 orang *content writer* dan 5 orang *SEO, web* dari CV. Digimedia Corpindo mampu menduduki *index* pertama *google* dan mempunyai rata-rata 150.000 pengunjung setiap hari. Hal ini tidak lepas dari kerja keras *content writer* dalam membuat artikel yang berkualitas.

Content writer adalah seseorang yang bertugas menulis sebuah artikel pada blog tertentu. Seorang *content writer* haruslah mempunyai ide-ide yang menarik untuk dituangkan dalam sebuah artikel. Tugas dari *Content writer* inilah, harus memilih kata-kata yang nantinya akan membuat informasi tersebut terlihat seperti fresh dan unik. Dengan menjaga kualitas artikel yang dipublikasikan maka traffik website akan selalu terjaga dan tidak menurun. Kesulitan pertama yang di alami *content writer* yaitu sulit dalam merangkai kata menjadi sebuat artikel yang bisa dikonsumsi untuk publik. Jika artikel yang dibuat tidak bisa menarik akan menurunkan traffik dari website tersebut. Dan jika traffik dari website menurun, maka keuntungan yang didapat juga dipastikan akan menurun. Apalagi jika tidak segera dilakukan evaluasi akan terjadi hal yang sangat tidak diinginkan dengan berhentinya kontrak iklan-iklan yang menggunakan jasa website tersebut. Hal ini menjadi problema tersendiri bagi *startup* untuk pengembangan website yang dikelolanya.

Untuk mencari seorang *content writer* yang handal tidaklah mudah, ada beberapa poin yang harus dimiliki oleh seorang *content writer*, antara lain adalah siap dibebani tanggung jawab, selalu *mengupgrade* kemampuan, paham dan mengerti apa yang ditulis, mencintai dunia tulis dan percaya diri. Dalam menyusun sebuah artikel yang berkualitas tidak hanya sekedar menulis dan mengarang, tetapi juga dibutuhkan optimasi dan keakurasian artikel sesuai dengan topik berita.

Saat ini cara perekrutan tenaga *content writer* di CV. Digimedia Corpindo masih manual. Proses yang dilakukan saat *test* masuk di perusahaan ini adalah dengan memberi tugas calon *content writer* untuk membuat 3 artikel yang mempunyai topik berita yang berbeda dengan waktu yang ditentukan. Setelah test selesai, personalia akan mengecek artikel yang dibuat oleh calon *content writer*. Hal tersebut akan memakan waktu dan kurang efisiennya menentukan keakurasian artikel yang diteliti karena masih menggunakan tenaga manusia. Dan yang terjadi *traffik* dari website tersebut setelah melakukan perekrutan tidak mengalami peningkatan yang efektif.

Masalah yang terjadi dalam perekrutan calon *content writer* baru tersebut dapat diminimalisir dengan adanya sarana aplikasi untuk penyeleksian artikel dengan nilai similaritas topik berita. Dari beberapa penelitian yang telah dilakukan sebelumnya (Ari Andika, 2015) "*Perancangan Aplikasi Pengukuran Similaritas Pada Dokumen Dengan Metode Semantic*", menjelaskan metode yang digunakan dalam perancangan aplikasi yaitu *Metode Semantic* sebagai pencocokan antara dokumen dengan kata kunci yang telah dibuat untuk menentukan atau memilih kategori suatu dokumen. Dan (Radiant Victor Imbar, Adelia, Mewati Ayub, Alexander Rehatta, 2014) "*Implementasi Cosine Similarity Dan Algoritma Smith-Waterman Untuk Mendeteksi Kemiripan Teks*" digunakan pendeteksian kemiripan teks untuk karya-karya tulis akademis dengan menggunakan aplikasi, sehingga keorisinalannya karya tulis akademis dapat diketahui dengan cepat. Dengan penelitian diatas dapat dilakukan pembobotan nilai similaritas, namun masih diperlukan penelitian dengan menggunakan metode lain untuk mendapatkan hasil yang lebih optimal.

Salah satu alternatif dalam menyelesaikan masalah kemudahan dalam pengklasifikasian tersebut adalah dengan menggunakan metode *text*

minning. *Text minning* merupakan varian dari *data mining* yang berusaha untuk *mengekstraksi* dan *menemukan* sesuatu yang belum diketahui sebelumnya dan mengandung informasi yang sangat berguna dari sekumpulan data tekstual yang besar jumlahnya (Hearts, 2003).

Dengan adanya kelemahan yang terdapat di CV. Digimedia Corpindo tersebut, penulis tertarik melakukan penelitian skripsi dengan judul "*Perancangan Aplikasi Pengukuran Similaritas Pada Artikel Dengan Algoritma TF-IDF Di CV. Digimedia Corpindo*". Dengan diadakannya penelitian tersebut, diharapkan dapat membantu menyelesaikan masalah yang ada di CV. Digimedia Corpindo dan dapat memaksimalkan keuntungan yang didapat.

II. Algoritma *TF/IDF* (*Term Frequency – Inversed Document Frequency*)

Metode TF-IDF merupakan metode untuk menghitung bobot setiap kata yang paling umum digunakan pada *information retrieval*. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat (Abdul Aziz Ma'arif., 2015). Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu, *frekuensi* kemunculan sebuah kata di dalam sebuah dokumen tertentu dan *inverse frekuensi* dokumen yang mengandung kata tersebut (Ari Andika., 2015).

Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata tersebut di dalam dokumen tersebut. *Frekuensi* dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antar sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan *frekuensi* keseluruhan dokumen yang ada mengandung kata tersebut yang rendah pada kumpulan dokumen (*database*) (Ari Andika., 2015).

a. Perhitungan *TF/IDF*

Pada algoritma *TF/IDF* digunakan rumus untuk menghitung bobot (W) masing-masing dokumen

terhadap kata kunci dengan rumus yaitu : (Adhit Herwansyah., 2012).

$$W_{dt} = tf_{dt} * IDF_t$$

Dimana:

d = dokumen ke-d

t = kata ke-t dari kata kunci

W = bobot dokumen ke-d terhadap kata ke-t

tf = banyaknya kata yang dicari pada sebuah dokumen

IDF =*InversedDocumentFrequency*

IDF = log2 (D/df)

D = total dokumen

df = banyak dokumen yang mengandung kata yang dicari

Setelah bobot (W) masing-masing dokumen diketahui, maka dilakukan proses pengurutan di mana semakin besar nilai *W*, semakin besar tingkat *similaritas* dokumen tersebut terhadap kata kunci, demikian sebaliknya. (Adhit Herwansyah., 2012).

III. Hasil Perhitungan Program

Dari hasil *testing* program, didapatkan program berjalan dengan lancar dan menghasilkan output yang sesuai dengan perhitungan. Adapun perhitungan data testing yang dilakukan sebagai berikut :

Kata kunci = menang liga pssi.

Dokumen 1 = Kesebelasan aremania telah kalah dalam pertandingan liga presiden melawan kesebelasan persebaya

Dokumen 2 = Kemenangan kesebelasan Persija telah diatur oleh wasit dalam tanding melawan Persik Kediri.

Dokumen 3 = PSSI Indonesia telah melarang segala macam pertandingan sepakbola dalam liga apapun sebelum presiden baru PSSI dipilih.

Setelah dokumen diinputkan, program melakukan *tokenizing* yaitu pemotongan string input berdasarkan tiap kata yang menyusunnya dan *stopword* yaitu pembuangan kata yang tidak penting dan tanca baca, sehingga menjadi :

Kata kunci = menang liga pssi.

Dokumen 1 = kesebelasan aremania kalah pertandingan liga presiden melawan kesebelasan persebaya.

Dokumen 2 = kemenangan kesebelasan persija diatur wasit tanding melawan persik kediri.

Dokumen 3 = pssi Indonesia melarang macam pertandingan sepakbola liga presiden pssi dipilih.

Selanjutnya melakukan *stemming* yaitu pengubahan kata menjadi kata dasar dengan menghilangkan kata imbuhan sesuai dengan *rule-rule morfologi* bahasa indonesia, sehingga menjadi :

Kata kunci = menang liga pssi.

Dokumen 1 = kesebelasan aremania kalah tanding liga presiden lawan kesebelasan persebaya.

Dokumen 2 = menang kesebelasan persija atur wasit tanding lawan persik diri.

Dokumen 3 = pssi Indonesia larang macam tanding sepakbola liga presiden pssi pilih.

Setelah proses *stemming* selesai selanjutnya adalah penerapan algoritma *TF-IDF* untuk melakukan pembobotan dokumen terhadap kata kunci. Dalam pembobotan ini akan didapat nilai pada setiap dokumen.

Term	tf				df	d/df	idf	Wdt=tf.idf			
	q	d1	d2	d3				q	d1	d2	d3
kesebelasan		2	1		2	1.5	0.176	0.000	0.352	0.176	0.000
aremania		1			1	3	0.477	0.000	0.477	0.000	0.000
kalah		1			1	3	0.477	0.000	0.477	0.000	0.000
tanding		1	1	1	3	1	0.000	0.000	0.000	0.000	0.000
liga	1	1		1	2	1.5	0.176	0.176	0.176	0.000	0.176
presiden		1		1	2	1.5	0.176	0.000	0.176	0.000	0.176
lawan		1	1		2	1.5	0.176	0.000	0.176	0.176	0.000
persebaya		1			1	3	0.477	0.000	0.477	0.000	0.000
menang	1		1		1	3	0.477	0.477	0.000	0.477	0.000
persija			1		1	3	0.477	0.000	0.000	0.477	0.000
atur			1		1	3	0.477	0.000	0.000	0.477	0.000
wasit			1		1	3	0.477	0.000	0.000	0.477	0.000
persik			1		1	3	0.477	0.000	0.000	0.477	0.000
kediri			1		1	3	0.477	0.000	0.000	0.477	0.000
pssi	1			2	1	3	0.477	0.477	0.000	0.000	0.954
indonesia				1	1	3	0.477	0.000	0.000	0.000	0.477
larang				1	1	3	0.477	0.000	0.000	0.000	0.477
macam				1	1	3	0.477	0.000	0.000	0.000	0.477
sepakbola				1	1	3	0.477	0.000	0.000	0.000	0.477
pilih				1	1	3	0.477	0.000	0.000	0.000	0.477
Total								1.130	2.312	3.215	2.738

Dari tabel di atas didapat nilai dari masing-masing dokumen sesuai dengan kata kunci yang telah digunakan, yaitu :

Bobot (*W*) untuk D1 = 0.176 + 0 + 0

= **0.176**

Bobot (*W*) untuk D2 = 0 + 0.477 + 0

= **0.477**

Bobot (*W*) untuk D3 = 0.176 + 0 + 0.954

= **1.130**

Setelah hasil didapat, maka perangkingan dokumen dapat dilakukan seperti tabel di bawah ini :

Dokumen 1	Dokumen 2	Dokumen 3
0.176	**0. 477**	**1.130**
Rank 3	**Rank 2**	**Rank 1**

a. Kesimpulan

Berdasarkan penelitian yang dilakukan oleh dalam membuat pengukuran similaritas pada artikel di CV. Digimedia Corpindo, maka dapat ditarik kesimpulan sebagai berikut :

1. Telah berhasil dibuatnya sebuah aplikasi untuk mengukur tingkat similaritas artikel dengan algoritma *TF-IDF* disetakan dengan hasil nilai bobot dari tiap dokumen yang ditemukan.
2. Aplikasi dapat memberikan *alternatif* pilihan untuk penerimaan calon *content writer* berdasarkan perangkingan dokumen sesuai dengan topik berita yang telah dipilih.

IV. DAFTAR PUSTAKA

1) Abdul, Ma'arif Azis. 2015. *Penerapan Algoritma TF-IDF Untuk Pencarian Karya Ilmiah.* Skripsi, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro.

2) Afuan, Lasmedi. 2013. *Stemming Dokumen Teks Bahasa Indonesia Menggunakan Algoritma Porter*. Prodi Teknik Informatika, Fakultas Sains dan Teknik, Purbalingga 2 Agustus.

3) Andika, Ari. 2015. *Perancangan Aplikasi Pengukuran Similaritas Pada Dokumen Dengan Metode Semantic*. Program Studi Teknik Informatika STMIK Budi Darma Medan, Medan 3 Pebruari.

4) Harlian, Milkha. 2006. *Machine Learning Text Kategorization*. Austin : Universityof Texas.

5) Hearst, Marti. 2003. *What Is Text Minning*?. SIMS, UC Berkeley. http://www.sims.berkeley.edu/~hearst/text.mining.html .

6) Herwansyah, Adhit. 2012. *Aplikasi Pengkategorian Dokumen Dan Pengukuran Tingkat Similaritas Dokumen Menggunakan Kata Kunci Pada Dokumen Penulisan Ilmiah Universitas Gunadarma*. Jurusan Sistem Informasi, Fakultas Ilmu Komputer Universitas Gunadarma.
7) Imbar, Radiant Victor., Adelia, Mewati Ayub, Alexander Rehatta. 2014. *Implementasi Cosine Similarity Dan Algoritma Smith-Waterman Untuk Mendeteksi Kemiripan Teks*. Jurnal Informatika, hal 31- 42.
8) Mustofa, Aziz . 2009. *Klasifikasi Otomatis Dokumen Berita Kejadian Berbahasa Indonesia*. Skripsi, Fakultas Sains dan Teknologi.
9) Tala, Fadillah Z. 2003. *A Study of Stemming Efects on Information Retrieval in*
10) *Bahasa Indonesia*. Institute for Logic, Language and Computation Universite Itvan Amsterdam The Netherlands.