INFORMATION RETRIEVAL SYSTEM PADA FILE PENCARIAN DOKUMEN TESIS BERBASIS TEXT MENGGUNAKAN METODE VECTOR SPACE MODEL

Ahmad Fauzi[1]; Ginabila[2]

[1, 2] Ilmu Komputer
STMIK Nusa Mandiri
www.nusamandiri.ac.id
[1]fauzi.aau@nusamandiri.ac.id, [2]14002151@nusamandiri.ac.id

Abstract—*Speed and density in the process of finding documents and information has become mandatory, contained in information systems, to facilitate the search process or find documents and information needed, it is called information retrieval or information retrieval system, implementation of the theory applied in this study using the model method vector space, the purpose of this study is to provide general exposure to the process of finding digital documents. With the token and indexing process so that the results of the masses are found in the database using keywords, so the system will search according to the keywords input into the system, and will be compared with the data contained in the database, so that it can produce the correct information.*

Keywords*: retrival information, vector space retrieval system model.*

Intisari— Kecepatan dan kepadatan dalam proses pencarian dokumen dan informasi telah menjadi wajib, terkandung dalam sistem informasi, untuk memudahkan proses pencarian atau menemukan dokumen dan informasi yang diperlukan, itu disebut informasi retrival atau sistem pengambilan informasi, implementasi Dari teori yang diterapkan dalam penelitian ini menggunakan metode model ruang vektor, tujuan penelitian ini adalah memberikan paparan umum tentang proses pencarian dokumen digital. Dengan proses token dan indexsing sehingga ditemukan hasil dari maskimal terdapat dalam database menggunakan kata kunci, sehingga sistem akan melakukan pencarian sesuai deanagn kata kunci yang di inputkan pada sistem, dan akan dibandingkan dengan data yang terdapat pada database, sehingga dapat menghasilkan informasi yang benar.

Kata Kunci*:* informasi retrival, model sistem pengambilan ruang vector.

PENDAHULUAN

Peningkatan arus informasi yang sangat cepat dalam mendukung kegiatan *browsing* dan *searching* bagi *user* untuk mempermudah aktivitas (Irmawati, 2017) Informasi tumbuh dengan sangat pesat dalam berbagai basis *content* seperti teks, *image, video, visual, audio* dan sebagainya. alat temu kembali *online public access catalog* (OPAC) sendiri sudah ada sejak tahun 1970. Sejak pertama kali diciptakan, pembuatan sistem temu kembali informasi telah mengalami proses perubahan sesuai perkembangannya (Lestari, 2016) Informasi tersebut tidak ada artinya bila informasi yang relevan tidak dapat ditemukan kembali guna memenuhi kebutuhan informasi pemustaka. Oleh karena itu, perpustakaan perguruan tinggi membutuhkan sistem temu kembali informasi *(information retrieval).*

perpustakaan perlu melakukan perubahan dalam pemeliharaan dan katalogisasi informasi, dari sistem tercetak menjadi *online* dalam bentuk digital agar dapat diakses dari mana saja (Amin, 2012). Perubahan sistem tersebut, terlihat pada pengembangan perpustakaan digital. Layanan perpustakaan digital menyediakan akses instan terhadap koleksi/dokumen, baik melalui metode pencarian *keyword*, penulis, maupun judul.(Sjaeful Afandi; Firman Ardiansya; Blasius Soedarsono, 2015) masalah utama dalam proses pencarian dokumen digital dibutuhkan waktu yang relatif lama karena pencariannya harus menyertakan isi judul dokumen secara lengkap dan benar pada aplikasi ELS-NURI, hal ini menjadi tidak relevan dalam sistem temu kembali informasi. Maka dari itu dibutuhkan sebuah *search engine* yang dapat

mencari dokumen-dokumen tersebut secara lebih cepat dan mudah serta menghasilkan informasi yang relevan tanpa perlu menyertakan judul dokumen secara lebih terperinci (Zain & Suswati, 2016) penulis mecoba menerapakan metode informasi retrival pada pembaharuan aplikasi ELS-NURI, guna memberikan informasi yang lebih baik dan akurat dalam proses pencarian dokumen tesis pada aplikasi ELS-NURI,sehingga mahasiswa dapat melakukan pencarian tanpa perlu mengetikan *keyword* secara lengkap dan terperinci, mahasiswa hanya perlu mengetikan kata kunci pada pencarian dokumen,makan semua isi yang berhubungan dengan katakunci yang sedang dicari akan ditampilkan secara lengkap.

Penelitian ini menggunakan *vektor space model* yang merupakan salah satu metode informasi retrival yang bertujuan untuk mempermudah dalam proses temu kembali informasi pada dokumen berbasis text digital, penelitian ini pernah dilakukan oleh (Zain & Suswati, 2016) pada perpustakaan fakultas Teknik universitas madurra menggunakan 3 data dan menghasilkan tiga rengking yang berbeda dari *query* yang di input pada sebuah *system*, Penelitian sebelumnya dilakukan oleh (Elektro et al., 2017) perhitungan kemiripan dokumen menggunakan vector space model. Sistem secara otomatis akan melakukan indexing secara offline dan temu kembali (retrieval) secara real time. Proses retrieval dimulai dengan mengambil query dari pengguna, kemudian sistem menghitung kemiripan antara keyword dengan daftar dokumen yang diwakili oleh term-term di dalam index. Dokumen akan ditampilkan diurutkan berdasarkan dokumen yang paling mirip. Penelitian sebelumnya mengenai sitem temu kembali yang di lakukan oleh (Putung et al., 2016) yang menjelaskan pencarian informasi dokumen skripsi. Terdapat dua proses utama dalam sistem temu kembali informasi yaitu indexing dan retrieval. Proses indexing adalah proses untuk memberikan bobot pada kata dalam dokumen, metode pembobotan pada penelitian ini menggunakan metode pembobotan TF-IDF. Prosesretrieval adalah proses untuk menghitung kemiripan query terhadap dokumen.

Tujuan penelitian ini Untuk mengimplementasikan retrieval system model pada aplikasi Pengambilan informasi menjadi bidang penelitian yang penting dibidang ilmu komputer. Dalam makalah ini, peneliti mewakili berbagai model dan teknik untuk pengambilan informasi. menjelaskan metode pengindeksan yang berbeda untuk mengurangi ruang pencarian dan teknik pencarian yang berbeda untuk mengambil informasi. Dari aplikasi repository.nusamandiri.ac.id yang menjadi bahan penelitian dalam penggunaan retrival information

BAHAN DAN METODE

Penulis Pengumpulan data dilakukan dengan cara mempelajari buku dan jurnal yang mendukung pada penelitian ini, termasuk di dalamnya literatur tentang penulisan dan mengenai hal-hal yang mendukung implementasi system temu kembali pada aplikasi.

Metadata koleksi dokumen tesis yang digunakan antara tahun 2010 – 2016 yang berjumlah 169 record. Data tersebut tidak berurutkan, Dari hasil penelusuran informasi, dihasilkan 6 dokumen tesis yang sering dilihat, pada tahap selanjutnya penelitian ini mengambil dari enam dokumen tesis sebagai semple pada penelitian kali ini.

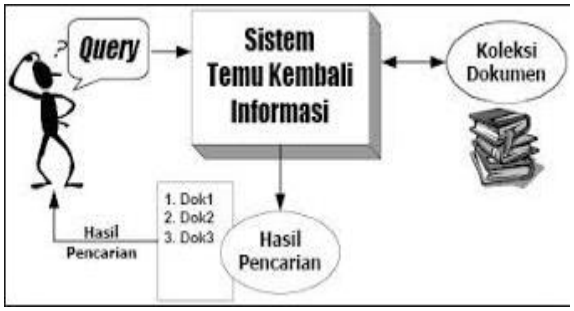
Information Retrieval System

Pengambilan informasi menunjukkan proses pencarian informasi yang diperlukan (Zhou, Liu, & Liu, 2012) Information retrieval (IR) umumnya berkaitan dengan pencarian dan pengambilan informasi berbasis pengetahuan (Sharma & Patel, 2013) sistem information retrieval (IR). Salah satu penerapan prinsip relevansi yang sejak dahulu digunakan dalam pengembanngan sistem (Lestari, 2016) Information Retrieval System menemukan informasi yang biasanya dalam bentuk dokumen dari sebuah data yang tidak terstruktur dalam bentuk teks untuk memenuhi kebutuhan informasi dari koleksi data yang sangat besar umumnya tersimpan dalam database computer (Amin & Purwatiningtyas, 2015)

Vector Space Model

Model ruang vektor memberikan sebuah kerangka pencocokan parsial Hal ini dicapai dengan menetapkan bobot non-biner untuk istilah indeks dalam *query* dan dokumen (Amin & Purwatiningtyas, 2015) Tidak hanya untuk pencarian teks, pencarian informasi juga dapat *query* elemen multimedia seperti gambar,suara,(Yulianto, Budiharto, & Kartowisastro, 2017) metode ini melihat tingkat kedekatan atau kesamaan (*smilarity*) *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*.(Zain & Suswati, 2016) dokumen dan *query* direpresentasikan sebagai *vektor* dan sudut antara keduanya.(V. K. Singh, Singh, Vishwavidyalaya,

Programmer, & Vishwavidyalaya, 2015) *vektor* dihitung menggunakan fungsi cosinus kesamaan. Efektivitas VSM sebagian besar tergantung pada istilah pembobotan yang diterapkan (Harcourt & Japheth, 2016) memungkinkan hasil penghitungan menjadi peringkat sesuai dengan ukuran kesamaan(J. N. Singh, 2012)



Sumber: (Afandi & Ardiansy, 2015)

Gambar1. Ilustrasi Model Sistem Temu Kembali Informasi

Langkah metode vector space model

a. Menghitung bobot dokumen dengan tf-idf Idf =log(D/df)
b. Menghitung jarak tiap dokumen dan query
   Sqrt (Q) = Sqrt (∑)
   Sqrt (D) = Sqrt ($\sum_j^n = 1 \quad Q_j^2$)
c. Menghitung Dot Product Sum
   (Q * D*i*) = ($\sum_j^n = DJ,_J^2$)
d. Menghitung Similaritas Cosine Θ
   $\frac{Q * D}{|Q| * |D|}$

Penelitian ini dilaksanakan dalam beberapa tahapan yang diilustrasikan pada Gambar 1. Data yang diproses dalam sistem ini adalah koleksi dokumen digital dan query yang telah disiapkan sebelumnya.

Dokumen yang relevan adalah yang paling dekat dengan *query* yang diberikan. Dengan cara yang sama, dua dokumen akan dianggap relevan jika mereka berada diwilayah tetangganya satu sama lain(R.K.Makhijani1, I.N.Bharambe2)

e. Membuat Ranking. Setelah menghitung nilai cosinus lalu di buat perangkingan dari dokumen-dokumen tersebut

Dokumen tesis yang sering dilihat menjadi sample pada penelitian ini.

Contoh : Query (Q) = Kajian metode Naive Bayes

1 (D1) = Kajian Penerapan Algoritma C45, Naive Bayes Dan Neural Network Untuk Memenuhi Penilaian Data Karyawan Service Level Agreement Di Bank

2 (D2) = Alternatif Pemilihan Sepeda Motor Dengan Metode Analytic Hierarchy Process(Ahp): Studi Kasus Pada Masyarakat Purwokerto

3 (D3) = Evaluasi Customer Knowledge Management Pada Situs E-Commerce

4 (D4) = Kajian Perbandingan Efektivitas pencarian lajur terpendek menggunakan algoritmatabu search ant colony optimization

5 (D5) = Knowledge Management System Pada Event Organizer Menggunakan Pendekatan Metode Specific Actions Berbasis Web-Mobile: Studi Kasus Kampus Amik Bsi Pontianak

6 (D6) = Penerapan Metode Adaptive-Network-Based Fuzzy Inference System (Anfis) Model Sugeno Untuk Memprediksi Index Saham : Studi Kasus Saham Lq45 Idx.

HASIL DAN PEMBAHASAN

Tabel 1. Perhitungan *tf (Term Frequency)*

Token	Q	Dokumen						DF	Token	Q	Dokumen						DF
		1	2	3	4	5	6				1	2	3	4	5	6	
Actiones	0	0	0	0	0	1	0	1	manage	0	0	0	1	0	1	0	2
Adaptive	0	0	0	0	0	0	1	1	masyarakat	0	0	1	0	0	0	0	1
Algoritma	0	1	0	0	1	0	0	2	prediksi	0	0	0	0	0	0	1	1
Alternatif	0	0	1	0	0	0	0	1	metode	1	0	1	0	0	1	1	3
Analytic	0	0	1	0	0	0	0	1	mobile	0	0	0	0	0	1	0	1
Banding	0	0	0	0	1	0	0	1	model	0	0	0	0	0	0	1	1
Based	0	0	0	0	0	0	1	1	motor	0	0	1	0	0	0	0	1
Bayes	1	1	0	0	0	0	0	1	naive	1	1	0	0	0	0	0	1
Cari	0	0	0	0	1	0	0	1	network	0	1	0	0	0	0	1	2
Colony	0	0	0	0	1	0	0	1	neural	0	1	0	0	0	0	0	1
Costomer	0	0	0	1	0	0	0	1	nilai	0	1	0	0	0	0	0	1
Data	0	1	0	0	0	0	0	1	optimization	0	0	0	0	1	0	0	1
Dekat	0	0	0	0	0	1	0	1	organizer	0	0	0	0	0	1	0	1
E-commers	0	0	0	1	0	0	0	1	pendek	0	0	0	0	1	0	0	1

Sumber: (Fauzi & Ginabila, 2019)

Sebelum melakukan perhitungan *tf* , perlu melakukan indexsing dan filtering terlebih dahulu dari semua dokumen yang ada, agar data yang di hasilkan dari setiap dokumen memiliki arti yang memiliki makna. D1, D2, D3, D4,D5,D6 = Dokumen tf = banyak kata yang dicari pada sebuah dokumen. D = total dokumen, df = Banyak dokumen yang mengandung kata yang dicar,Dari hasil Perhitungan *tf* , data sample dari jumlah dokumen yang ada dihasilkan 60 token dari 6 dokumen dan satu query, untuk mendapatkan jarak dokumen dan query, di perlukan perhitungan *idf* yang di hasilkan dari tokenasi hasil perhitungan pada table 2 berikut:

Table 2. Perhitungan *Term Frequency - Inverse Document Frequency*

Idf				**tf*idf**			
Log (D/df)	**Q**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
0.778	0	0	0	0	0	0.778	0
0.778	0	0	0	0	0	0	0.778
0.477	0	0.778	0	0	0.778	0	0
0.778	0	0	0.778	0	0	0	0
0.778	0	0	0.778	0	0	0	0
0.778	0	0	0	0	0.778	0	0
0.778	0	0	0	0	0	0	0.778
0.778	0.778	0.778	0	0	0	0	0
0.778	0	0	0	0	0.778	0	0
0.778	0	0	0	0	1	0	0
0.778	0	0	0	0.778	0	0	0
0.778	0	0.778	0	0	0	0	0
0.778	0	0	0	0	0	0.778	0
0.778	0	0	0	0.778	0	0	0

Sumber: (Fauzi & Ginabila, 2019)

TF-IDF (Term Frequency - Inverse Document Frequency) merupakan perhitungan statistik yang bertujuan untuk memberikan gambarkan seberapa penting sebuah kata terhadap sebuah koleksi dokumen yang tersedia. TF-IDF (Term Frequency - Inverse Document Frequency) digunakan untuk pembobotan dalam Information Retrieval dan text mining. Nilai TF-IDF (Term Frequency - Inverse Document Frequency) akan meningkat seiring dengan banyaknya jumlah kata yang sering muncul di dalam koleksi dokumen.

Table 3. Perhitungan Jarak Q-D

			Jara Q-D			
Q	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
0	0	0	0	0	0.605	0
0	0	0	0	0	0	0.605
0	0.605	0	0	0.605	0	0
0	0	0.605	0	0	0	0
0	0	0.605	0	0	0	0
0	0	0	0	0.605	0	0
0	0	0	0	0	0	0.605
0.605	0.605	0	0	0	0	0
0	0	0	0	0.605	0	0
0	0	0	0	1	0	0
0	0	0	0	0	0.605	0
0	0	0	0	0	0	0.605
0	0	0	0	0.605	0.605	0
2.422	6.660	7.871	3.027	7.660	9.082	8.099
SQRT (Q)			SQRT (D)			
1.556	2.580	2.805	1.739	2.767	3.013	2.845

Sumber: (Fauzi & Ginabila, 2019)

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 *query*,panjang dokumen cendrung memiliki frekuensi kemunculan kata yang besar.Setelah diketahui perhitungan jarak antara Q-D dengan menggunakan rumus Sqrt (D) = Sqrt ($\sum_{j}^{n} = 1 \quad Q_j^2$).

Tabel 4. Perhitungan *Dot Product*

Dont Produk					
Q*D1	**Q*D2**	**Q*D3**	**Q*D4**	**Q*D5**	**Q*D6**
0	0	0	0	0	0
0.366	0	0	0	0	0
0	0	0	0	0	0
0.366	0	0	0.366	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0.366	0	0	0.366	0.366
0	0	0	0	0	0
0.366	0	0	0	0	0
SUM (Q*D)					
1.099	0.366	0	0.366	0.366	0.366

Sumber: (Fauzi & Ginabila, 2019)

setelah mendapatkan nilai bobot TF-IDF dan jarak antara dokumen dengan *query* (Q-D) selanjutnya masing-masing kata, akan dihitung *dot product* untuk melihatkan derajat kemiripan antar dokumen teks yang tersedia. Hasil dari perhitungan *dot product* dapat dilihat pada table 4, hasilnya akan mempengaruhi tingkat *cosine similarity.*Perhitungan similaritas Langkah selanjutnya adalah menghitung nilai *Cosinus* sudut antara *vector query* dengan tiap dokumen dengan rumus

$$\text{Cosine } \emptyset \text{ D}i = \frac{Q*D}{|Q|*|D|}$$

$$\text{D1} = \frac{1{,}099961117}{1{,}556302501 * 2{,}580835728} = 0{,}273856438$$

$$\text{D2} = \frac{0{,}366653706}{1{,}556302501 * 2{,}805664233} = 0{,}083970428$$

$$\text{D3} = \frac{0}{1{,}556302501 * 1{,}739999093} = 0$$

$$\text{D4} = \frac{0{,}366653706}{1{,}556302501 * 2{,}767799316} = 0{,}085119187$$

$$\text{D5} = \frac{0{,}366653706}{1{,}556302501 * 3{,}013766834} = 0{,}078172214$$

$$\text{D6} = \frac{0{,}366653706}{1{,}556302501 * 2{,}845943865} = 0{,}082781966$$

Dari Analisa *Vector Space Model* diperoleh hasil untuk ketiga dokumen di atas adalah seperti table 5 dibawah ini:

Tabel 5. Pembuatan Ranking

PEMBUATAN RENGKING		
Dokumen	**Nilai**	**Rengking**
D1	0,27385643	1
D2	0,08397042	3
D3	0	6
D4	0,08511918	2
D5	0,07817221	5
D6	0,082781966	4

Sumber: (Fauzi & Ginabila, 2019)

KESIMPULAN

Penelitian ini bekerja sangat signifikan dengan mengambil informasi dari database yang besar cukup memakan waktu terutama jika informasi tersebut tidak terstruktur. Banyak algoritma dan teknik telah dikembangkan di bidang penambangan data dan pengambilan informasi namun mengambil data dari basis data besar terus menjadi masalah. Dalam penelitian ini, kami menggunakan model ruang *vektor* untuk melakukan Proses pencarian yang sebelunya harus menyertakan judul dokumen secara lengkap, setelah menerapkan konsep *information retrival system,* pencarian dapat dilakukan dengan lebih cepat tepat dan akurat, tanpa perlu melakukan pencarian judul dokumen secara terperinci, sistem akan menyamakan *keyword* yang di masukan dengan dokumen yang tersimpan pada aplikasi dengan mengambil informasi dari database Aplikasi ELS-NURI sebagai bahan uji coba. Pertama, menghitung skor kemiripan menggunakan rata-rata tertimbang dari setiap item ukuran kosinus kemudian menghitung ukuran kemiripan dan untuk menentukan sudut antara *vektor* dokumen dan vektor *query* karena VSM didasarkan pada geometri di mana setiap istilah memiliki dimensi sendiri dalam ruang multi-dimensi, pertanyaan dan dokumen adalah titik atau *vektor* dalam ruang ini. Ukuran kosinus sering digunakan. hal tersebut lebih memudahkan

pencarian dokumen pada perpustakaan *digital.* Dengan sistem informasi temu kembali dapat merancang sebuah alat yang akan memungkinkan pengguna untuk mengambil informasi secara lebih efisien dan efektif.

REFERENSI

Amin, F. (2012). Sistem Temu Kembali Informasi dengan Metode Vector Space Model, *02*, 78–83.

Amin, F., & Purwatiningtyas. (2015). Rancang Bangun Information Retrieval System (IRS) Bahasa Jawa Ngoko pada Palintangan Penjebar Semangad dengan Metode Vector Space Model (VSM). *Jurnal Teknologi Informasi DINAMIK*, *20*(1), 25–35.

Elektro, J. T., Studi, P., Telekomunikasi, T., Malang, P. N., Atmadja, M. D., Elektro, J. T., … Malang, P. N. (2017). INFORMATION RETRIEVAL TUGAS AKHIR DAN PERHITUNGAN KEMIRIPAN, *8*(1), 355–362.

Harcourt, P., & Japheth, R. B. (2016). Application of Vector Space Model to Query Ranking and Information Retrieval, *6*(5), 42–47.

Irmawati. (2017). Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model. *Jurnal Ilmiah Fifo*, *IX*(1), 74–80.

Lestari, N. P. (2016). Uji Recall and Precision Sistem Temu Kembali Informasi OPAC Perpustakaan ITS Surabaya SKRIPSI. *Universitas Airlangga*, 1. Retrieved from http://journal.unair.ac.id/LN@uji-recall-and-precision-sistem-temu-kembali-informasi-opac-perpustakaan-its-surabaya-article-10825-media-136-category-8.html

Putung, K. D., Lumenta, A., Jacobus, A., Informatika, T., Sam, U., & Manado, R. (2016). KUMPULAN DOKUMEN SKRIPSI, *8*(1).

Sharma, M., & Patel, R. (2013). A Survey on Information Retrieval Models, Techniques And Applications. *International Journal of Emerging Technology and Advanced Engineering*, *3*(11), 542–545.

Singh, J. N. (2012). Analysis of Vector Space Model in Information Retrieval, 14–18.

Singh, V. K., Singh, V. K., Vishwavidyalaya, G., Programmer, S. A., & Vishwavidyalaya, G. G. (2015). VECTOR SPACE MODEL : AN INFORMATION RETRIEVAL, 141–143.

Sjaeful Afandi; Firman Ardiansya; Blasius Soedarsono. (2015). Pengembangan Sistem Temu Kembali Informasi Digital Fulltext Artikel Jurnal Di Pdii – Lipi. *Baca: Jurnal Dokumentasi Dan Informasi*, *36*(1), 65–76. https://doi.org/http://dx.doi.org/10.14203/j.baca.v36i1.203

Yulianto, B., Budiharto, W., & Kartowisastro, I. H. (2017). The Performance of Boolean Retrieval and Vector Space Model in Textual Information, *11*(1), 33–39.

Zain, M. Y., & Suswati. (2016). Information Retrieval System Pada Pencarian File Dokumen Berbasis Teks Dengan Metode Vector Space Model Dan Algoritma ECS Stemmer. *Jurnal Insand Comtech*, *1*(1), 30–37.

Zhou, H., Liu, B. W., & Liu, J. (2012). Research on mechanism of the information retrieval based on ontology label. *Procedia Engineering*, *29*, 4259–4266. https://doi.org/10.1016/j.proeng.2012.01.654