PROSIDING

Senastik 2012

Seminar Nasional Teknologi Informasi & Komputasi

Sinergi Pengembangan Industri Kreatif dan Riset Teknologi Informasi Untuk Peningkatan Daya Saing Bangsa

Bangkalan, 13-14 Nopember 2012

Teknik Informatika

Universitas Trunojoyo

Scanned by CamScanner

PROSIDING

Senastik 2012

Seminar Nasional Teknologi Informasi & Komputasi

Sinergi Pengembangan Industri Kreatif dan Riset Teknologi Informasi Untuk Peningkatan Daya Saing Bangsa

Bangkalan, 13-14 November 2012

Teknik Informatika

Universitas Trunojoyo

Scanned by CamScanner

Teknik Informatika
Fakultas Teknik
Universitas Trunojoyo

PROSIDING

SENASTIK 2012

SEMINAR NASIONAL TEKNOLOGI INFORMASI & KOMPUTASI

Bangkalan, 13-14 Nopember 2012
Kampus Universitas Trunojoyo
Jl Raya POBOX 2 Kamal, Bangkalan

Editor :

1. **M. Kautsar Sophan, S.Kom., M.MT**
2. **Abdullah Basuki Rahmat, S.Si., MT.**
3. **Andharini Dwi Cahyani, S.Kom., M.Kom**
4. **Rima Triwahyuningrum, S.T.,M.T.**
5. **Iwan Santosa, ST.M.T**
6. **Arik Kurniawati, S.Kom, M.T.**
7. **Bain Khusnul Khotimah,S.T., M.Kom**

Scanned by CamScanner

SENASTIK 2012

Susunan Panitia

Keynote Speaker

1. **Prof. Drs. Ec. Ir. Riyanarto Sarno, M.Sc, Ph.D**
2. **DR. Romi Satrio Wahono, M.Eng., B.Eng.**

Mitra Bestari Eksternal

Nama	Institusi	Bidang
Dr. Ir. Yoyon K. Suprapto, M.Sc	Teknik Elektro ITS	Signal Processing
Dr. Ir. Risanuri Hidayat, M.Sc	Teknik Elektro dan Informatika UGM	
Ahmad Basuki S.Si, M.Kom, Ph.D.	Politeknik Negeri Surabaya (PENS) ITS	Image & Video Processing
Dr. Dra. Tatik Maftukhah M.T.	Pusat Penelitian Kalibrasi, Instrumentasi dan Metrologi (LIPI).	Computational Intelligence
Dr. Rahmad Syam, S.T., M.T.	Universitas Negeri Makasar	Biometrics
Dr. Syaiful Bukhori, S.T., M.Kom.	Universitas Jember	Data Mining
Dr. Yeni Herdiyeni Department of	Computer Science Bogor Agricultural University (IPB)	Computational Intelligence
Dr. Taufik Fuadi Abidin, S.Si., M.Tech.	Universitas Syah Kuala , Aceh	Web Mining
Paulus Insap Sentosa, M.Sc., Ph.D.	Teknik Elektro UGM	Software Engineering, HCI, Information System
Dr. Atris Suyantohadi, M.T.	Artificial Life Fakultas Pertanian Universitas Gajah Mada (UGM)	Artificial Life
Dr. Arif Muntasa, S.Si., M.T.	Teknik Informatika Universitas Trunojoyo	Computational Intelligence
Dr. Indah Agustien Siradjuddin,	Teknik Informatika Universitas Trunojoyo	Computational Intelligence

Penanggung Jawab : Firdaus Solihin, S.Kom., M.Kom.

Komite Pelaksana
Ketua: Hermawan, ST.M.Kom

Anggota :
Bain Khusnul Khotimah,S.T., M.Kom
Rima Triwahyuningrum, S.T.,M.T.
Ari Kusumaningsih, S.T, M.T
Fika Hastarita Rachman, ST., M.Eng
Cucun Very Angkoso, S.T, M.T
Eza Rahmanita, S.T, M.T
Mula'ab, S.Si., M.Kom.
Dwi Kuswanto, S.Pd., MT.

Scanned by CamScanner

KATA PENGANTAR

Seminar Nasional Teknologi Informasi dan Komputasi 2012 (SENASTIK) merupakan temu ilmiah nasional tahunan yang diselenggarakan oleh Program Studi Teknik Informatika Universitas Trunojoyo. Seminar ini kami adakan sebagai sarana desiminasi hasil – hasil pelatihan atau kajian – kajian pada bidang Teknologi Informasi dan Komputasi dengan skala nasional. Selain itu, kami berharap bahwa melalui seminar ini bisa mewadahi komunikasi antar peneliti, praktisi dan akademisi. Pada tahun pertama ini, kami mengangkat tema "Sinergi Pengembangan Industri Kreatif dan Riset Teknologi Informasi Untuk Peningkatan Daya Saing Bangsa". Tema tersebut bertujuan untuk menghimpun inovasi teknologi informasi terkait dengan pengembangan industri kreatif guna mendukung peningkatan daya saing bangsa Indonesia di kancah dunia.

Prosiding ini disusun untuk mendokumentasikan dan mengkomunikasikan hasil seminar nasional tersebut yang terangkum dalam makalah – makalah yang disajikan dalam seminar serta hasil rumusan seminar.

Pada kesempatan ini kami menyampaikan terima kasih kepada para penyaji dan penulis makalah, penyunting serta redaksi pelaksana yang telah bekerja keras sehingga prosiding ini dapat diterbitkan. Mudah – mudahan prosiding ini bermanfaat bagi pihak – pihak yang berkepentingan, utamanya bagi perkembangan dunia IT.

Tertanda

Ketua Panitia
Hermawan, S.T., M.Kom

Scanned by CamScanner

DAFTAR ISI

Scanned by CamScanner

Scanned by CamScanner

Scanned by CamScanner

ISSN : 2302-7088

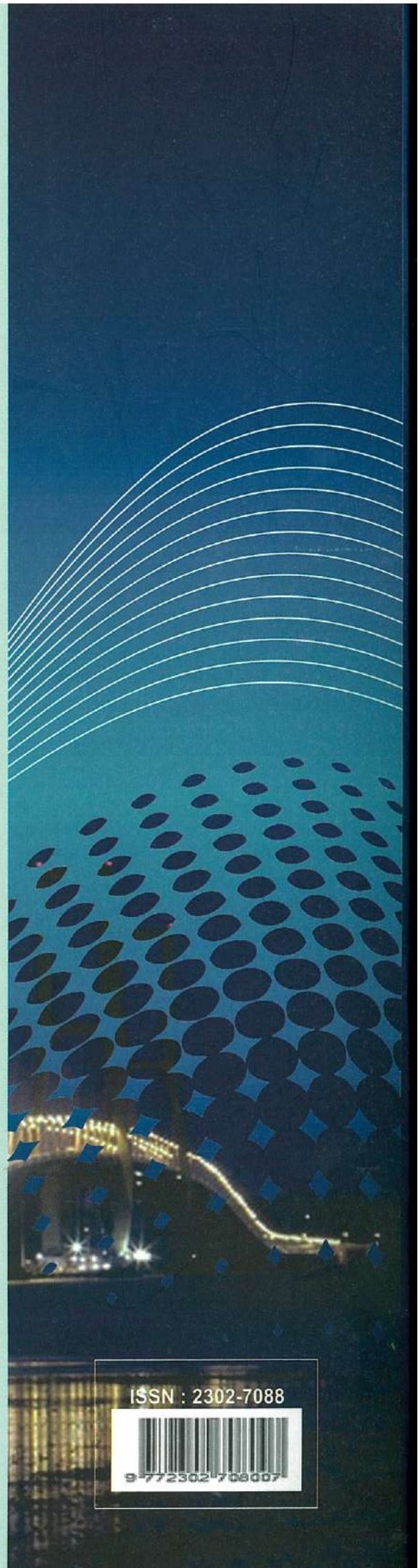
Scanned by CamScanner

Topik Seminar

- Sistem Informasi
- Komputasi
- Sistem Terdistribusi & Jaringan
- Multimedia
- Pendidikan IT

SENASTIK 2012

Seminar Nasional Teknologi Informasi & Komputasi



ISSN : 2302-7088

Program Studi Teknik Informatika
Universitas Trunojoyo

Jl Raya Telang POBOX 2 Kamal
Bangkalan

Telp 0313011147
senastik@if.trunojoyo.ac.id
http://senastik.trunojoyo.ac.id/

Implementasi Generalized Vector Space Model Menggunakan WordNet

Adi Wibowo*, Andreas Handojo, Charistian Widjaja*****
Jurusan Teknik Informatika
Fakultas Teknologi Industri, Universitas Kristen Petra
E-Mail: *adiw@petra.ac.id, **handojo@petra.ac.id, ***m26408061@john.petra.ac.id

Abstrak

Dengan pesatnya perkembangan dalam penggunaan teknologi komputer baik di perusahaan maupun di bidang pendidikan, maka semakin banyak pula dokumen-dokumen yang berbentuk digital yang dihasilkan. Metode yang sering dipergunakan untuk mencari dokumen adalah Vector Space Model (VSM). Kelemahan utama dari VSM adalah tidak mampu menemukan dokumen yang walaupun relevan dengan kata kunci tetapi tidak mengandung kata kunci tersebut. Oleh karena itu dibutuhkan sebuah metode search engine yang dapat memanfaatkan kemiripan makna antar kata untuk mengatasi masalah diatas.

Salah satu metode yang dipergunakan dalam perancangan search engine adalah Generalized Vector Space Model (GVSM). George Tsatsaronis dan Vicky Panagiotopolou mengembangkan metode GVSM dengan melakukan pemberian nilai kedekatan antar sense didapatkan dengan metode Semantic Relatedness yang mempergunakan database leksikal "WordNet".

Dari hasil pengujian yang dilakukan maka GVSM menghasilkan hasil pencarian dokumen-dokumen yang memiliki nilai recall yang sama atau lebih tinggi yaitu 0,4 ; 1 ; 0,7778 jika dibandingkan dengan VSM (0,4 ; 0 ; 0,2222). Sedangkan nilai precision dari hasil pencarian GVSM memiliki nilai yang lebih rendah yaitu 0,0526 ; 0,0588 ; 0,1707 jika dibandingkan dengan nilai precision dari hasil pencarian VSM yaitu 0,1333 ; 0 ; 0,2857 .

Kata kunci: Vector Space Model, GVSM, WordNet, Relasi Makna.

Abstract

With the rapid growth in the use of computer technology both in companies and in the field of education, more documents are generated in digital form. The method frequently used to search for documents is Vector Space Model (VSM). The main drawback of the VSM is not able to find relevant documents which do not contain the keyword terms. So we need a search method that can utilize the similarity of meaning between terms to overcome the above problems.

One of the methods used in the design of search engines is the Generalized Vector Space Model (GVSMGeorge and Vicky Tsatsaronis Panagiotopolou develop methods GVSM by scoring sense closeness between Semantic Relatedness obtained with the method that uses lexical databases "WordNet".

The test results produce that GVSM documents have the same recall value or higher at 0.4; 1; 0.7778 compared with VSM (0.4; 0; 0.2222). While the value of precision of the search results GVSM have a lower value is 0.0526; 0.0588; 0.1707 when compared with the value of precision of the search results VSM is 0.1333; 0; 0.2857.

Key words: Vector Space Model, GVSM, WordNet, Semantic Relatedness.

PENDAHULUAN

Dengan pesatnya perkembangan penggunaan teknologi komputer baik di perusahaan maupun di bidang pendidikan, maka semakin banyak pula dokumen yang berbentuk digital. Untuk mencari dokumen-dokumen tersebut dibutuhkan waktu yang relatif lama apabila pencariannya dilakukan secara manual. Maka dari itu dibutuhkan sebuah search engine yang dapat mencari dokumen-dokumen yang relevan secara lebih mudah. Salah satu metode yang dipergunakan dalam perancangan search engine adalah Vector Space Model.

Vector Space Model (VSM) sebagai metode yang mengukur kemiripan antara suatu dokumen dengan suatu query user dengan menggunakan cosinus dari sudut antar vektor yang dibentuk oleh dokumen dengan vektor dari kata kunci yang diinputkan oleh user [4]. Salah satu kelemahan dari VSM adalah metode ini menganggap bahwa setiap term pada dokumen bersifat independen, yaitu metode ini tidak melihat hubungan makna dengan term lain [2]. Sebagai contoh, apabila user melakukan pencarian dengan kata kunci "programming" maka hasil pencariannya adalah semua dokumen yang hanya memiliki kata "programming" saja, padahal masih banyak dokumen-dokumen yang masih berhubungan makna dengan kata "programming" seperti "PHP", "Java" , dan lain-lain. Dengan adanya kasus ini maka terjadi penurunan recall dari hasil pencarian. Karena itu dibutuhkan metode yang dapat mengembangkan VSM ini dengan menambahkan fungsi sense pada model ini yaitu GVSM (Generalized Vector Space Model).

Generalized Vector Space Model adalah model pencarian pengembangan dari *Vector Space Model* yang menambahkan fungsi sense dan penilaian terhadap hubungan makna antar term dalam dokumen [6*]. Generalized Vector Space Model (GVSM)* adalah Vector Space Model yang mempertimbangkan kedekatan sense antar term dalam merepresentasikan dokumen. Dalam GVSM ini pemberian nilai kedekatan antar sense didapatkan dengan metode *Semantic Relatedness*. Dimana metode *Semantic Relatedness* adalah metode yang menghitung nilai kedekatan sense dengan menggunakan kedalaman term dalam thesaurus dan banyaknya path yang dilalui antar dua term yaitu term yang ada di dokumen dan term pada kata kunci dari user. Dalam melakukan perhitungan dengan menggunakan metode *Semantic Relatedness* ini dibutuhkan thesaurus kata seperti "WordNet". Upaya penggunaan metode GVSM dan *Semantic Relatedness* ini dimaksudkan untuk meningkatkan recall dari hasil pencarian sehingga hasil pencariannya mencakup dokumen-dokumen yang relevan terhadap kata kunci dari user.

VECTOR SPACE MODEL

Vector Space Model adalah suatu model yang digunakan untuk mengukur kemiripan antara suatu dokumen dan suatu query dengan mewakili setiap dokumen dalam sebuah koleksi sebagai sebuah titik dalam ruang (vektor dalam ruang vektor) [7]. Poin yang berdekatan di ruang ini memiliki kesamaan semantik yang dekat dan titik yang terpisah jauh memiliki kesamaan semantik yang semakin jauh. Kesamaan antara vektor dokumen dengan vektor query tersebut dinyatakan dengan cosinus dari sudut antar keduanya [4].

Dalam metode Vector Space Model bobot dari setiap term yang didapat dalam semua dokumen dan query dari user harus dihitung lebih dulu. Term adalah suatu kata atau suatu kumpulan kata yang merupakan ekspresi verbal dari suatu pengertian. Perhitungan bobot tersebut dilakukan melalui persamaan nomor 1.

$$Term\ Weight: w_i = tf_i * \log \frac{D}{df_i} \quad (1)$$

tfi = frekuensi term atau banyak term i yang ada pada sebuah dokumen (Term Frequency)
dfi = frekuensi dokumen atau banyak dokumen yang mengandung term i (Inverse Document Frequency)
D = jumlah semua dokumen

Setelah itu untuk mengetahui tingkat kemiripan antar dokumen nilai cosinus dari sudut antar vektor dokumen dengan vektor query dihitung melalui persamaan nomor 2.

$$\cos \theta_{D_i} = \mathrm{Sim}(Q, D_i) \quad (2)$$

Dimana

$$\mathrm{Sim}(Q, D_i) = \frac{\sum_j w_{Q,j} w_{i,j}}{\sqrt{\sum_j w_{Q,j}^2} \sqrt{\sum_j w_{i,j}^2}}$$

Sim(Q,Di) = nilai kesamaan antara sebuah dokumen i dengan query Q
$w_{Q,j}$ = bobot term j pada query Q
$w_{i,j}$ = bobot term j pada dokumen i

Hasil cosinus tersebut diurutkan dari nilai kesamaan yang terbesar ke nilai yang terkecil. Hasil terbesar memiliki kedekatan yang lebih baik dengan user query dibandingkan nilai kesamaan yang lebih kecil [5].

GENERALIZED VECTOR SPACE MODEL

Generalized Vector Space Model (GVSM) adalah perkembangan dari *Vector Space Model* yang mempertimbangkan kedekatan *sense* antar

term dengan lebih akurat, dalam merepresentasikan dokumen. Wong et al. (1987) membuat GVSM pertama, yang memperkenalkan korelasi antar term, yang menganggap bahwa setiap term dinyatakan sebagai kombinasi linier dari vektor 2 dimensi. Pengukuran similiarity antara sebuah dokumen dengan sebuah query dilakukan dengan persamaan nomor 3.

$$\cos(\vec{d_k}, \vec{q}) = \frac{\sum_{j=1}^{\acute{n}} \sum_{i=1}^{\acute{n}} \acute{a}_{ki} \acute{q}_j \vec{t_i} \vec{t_j}}{\sqrt{\sum_{i=1}^{\acute{n}} \acute{a}_{ki}^2 \sum_{j=1}^{\acute{n}} \acute{q}_j^2}} \quad (3)$$

Dimana, t_i dan t_j adalah term vektor di sebuah ruang vektor 2 dimensi; d_k, dan q adalah vektor dokumen dan query; a_{ki} adalah bobot (*weight*) dari dokumen yang dihitung dengan rumus Term Weight; q_j adalah bobot (*weight*) dari query yang dihitung dengan rumus Term Weight; n adalah dimensi ruang [6].

SEMANTIC RELATEDNESS

t_i t_j menunjukkan besar relasi antara term I dan term j. Dalam *Semantic Relatedness* nilai dari t_i t_j dalam rumus GVSM Wong et al. dicari dengan rumus baru yang dikembangkan oleh George Tsatsaronis dan Vicky Panagiotopoulou dengan bantuan database leksikal "WordNet". Nilai t_i dan t_j dihitung melalui penghitungan SCM (*semantic compactness*), SPE (*semantic path elaboration*), dan SR (*semantic relatedness*). Langkah-langkah mencari nilai t_i dan t_j adalah [6]:

- Bila ada sebuah thesaurus O, sebuah bagan pembobotan (*weight*) yang menentukan *weight* e ∈ (0,1) untuk setiap edge, sepasang *senses* S=(s1,s2), dan sebuah path dengan panjang *l* yang menyambungkan 2 senses tersebut, maka *Semantic compactness* dari S dihitung menggunakan persamaan (4).

$$SCM(S,O) = \prod_{i=1}^{l} e_i \quad (4)$$

dimana e_1, e_2, e_3 adalah *path's edges*
Jika $s_1 = s_2$ maka SCM(S,O) = 1 dan jika tidak ada *path* antar keduanya maka SCM(S,O) = 0.

- Bila ada sebuah *thesaurus* O dan sepasang *senses* S=(s_1,s_2), dimana s_1,s_2 ∈ O dan $s_1 \neq s_2$ dan sebuah *path* dengan panjang *l* yang menyambungkan 2 *senses*, maka *Semantic path elaboration* dari S dihitung menggunakan persamaan (5).

$$SPE(S,O) = \prod_{i=1}^{l} \frac{2 d_i d_{i+1}}{d_i + d_{i+1}} \cdot \frac{1}{d_{max}} \quad (5)$$

dimana d_i adalah kedalaman *sense* s_i yang didasarkan pada O dan d_{max} adalah kedalaman maksimum dari O.
Jika $s_1 = s_2$ dan d =$d_1 = d_2$ maka SPE(S,O) = d/d_{max} dan jika tidak ada *path* antar keduanya maka SPE(S,O) = 0.

- Bila ada *thesaurus* O, sepasang *term* T=(t_1,t_2), dan semua pasang *senses* S=(s_{1i},s_{2j}), dimana s_{1i},s_{2j} merupakan *sense* dari t_1 dan t_2, maka *Semantic relatedness* dari T ditunjukkan dari persamaan (6).

$$SR(T,S,O) = \max\{SCM(S,O) \cdot SPE(S,O)\} \quad (6)$$

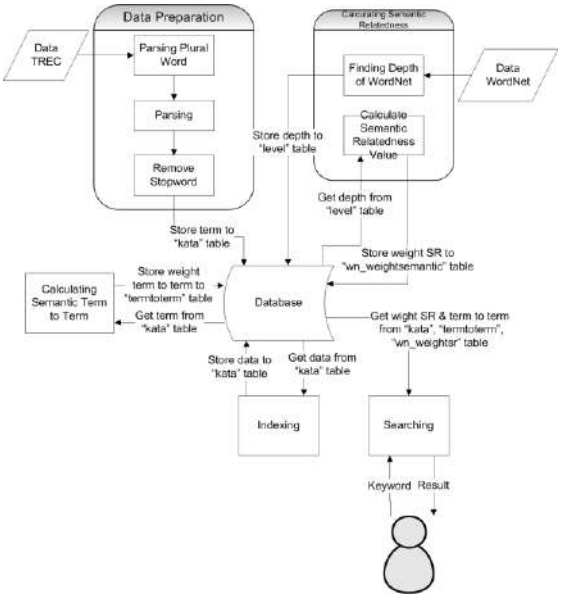
SR antar dua *terms* t_i,t_j dimana $t_i \equiv t_j \equiv t$ dan $t \notin O$ didefinisikan dengan 1. Jika $t_i \in O$ tapi $t_j \notin O$ atau $t_i \notin O$ tapi $t_j \in O$, SR=0.

IMPLEMENTASI DAN PENGUJIAN

Ada beberapa proses utama yang ada pada sistem, yaitu

1. *Data Preparation.*
2. *Indexing.*
3. *Calculating Semantic Relatedness.*
4. *Calculating Term to Term Cooccurence.*
5. *Searching.*

Gambar 1 menunjukkan blok diagram dari aplikasi ini.



Gambar 1. Blok Diagram dari Aplikasi

Data Preparation

Proses ini melakukan pengubahan terhadap file yang dipergunakan sebagai obyek pencarian yaitu "ClueWeb09_English_Sample.warc" yang didapatkan dari *Web Track* TREC (*The Text Retrieval Conference*). File tersebut berisi kumpulan file HTML menjadi beberapa file HTML yang terpisah. Setelah selesai akan dilakukan proses merubah HTML ke teks, yang kemudian diteruskan dengan proses *parsing* pada teks tersebut.

Indexing

Proses ini melakukan perhitungan *weight* pada setiap kata yang merupakan hasil *parsing* dari proses *data preparation* dengan menggunakan metode *Term Frequency* dan *Inverse Document Frequency* (TF-IDF) yang juga terdapat pada metode *Vector Space Model* (VSM). Hasil perhitungan *weight* untuk setiap kata/*term* ini nantinya dipergunakan dalam proses *Generalized Vector Space Model* (GVSM), yang nilainya dapat berpengaruh terhadap kemunculan dokumen yang diwakili oleh kata/*term* tersebut pada hasil pencarian.

Calculating Semantic Relatedness

WordNet adalah sebuah *thesaurus* yang menggambarkan hubungan antar term secara semantik/makna. Dalam WordNet hubungan antar term berupa relasi *synonym* (sama makna), *hyponym* (makna lebih sempit), *hypernym* (makna lebih luas), *meronym* (makna bagian lebih utuh), dan *holonym* (makna bagian dari sebuah benda). Tidak setiap term memiliki semua relasi di atas dengan term yang lain.

Proses ini melakukan perhitungan *semantic relatedness* dari tiap kata/*term* dalam *database* "*WordNet*" yang nilainya nanti dijadikan sebagai nilai kedekatan makna antara dua kata/*term*, yang dapat meningkatkan *recall* dari hasil pencarian. Nilai kedekatan makna ini nantinya dipergunakan dalam proses *Generalized Vector Space Model* (GVSM).

Kesulitan yang muncul adalah karena WordNet yang berbentuk graph sehingga sulit ditentukan term dengan level tertinggi, berbeda dengan misalnya WordNet berbentuk sebuah tree. Hal ini membuat kedalaman sebuah sense sulit untuk ditentukan. Untuk itu perlu dicari sebuah term yang dapati dianggap sebagai level yang paling tinggi dari hampir semua term, yaitu term "Thing".

Calculating Term to Term Cooccurence

Bila sebuah term tidak terdapat dalam WordNet, maka relasi makna antar term didapatkan dari term-to-term co-occurence matrix. Proses ini melakukan perhitungan terhadap nilai kedekatan makna dengan menghitung jumlah kemunculan bersama antara dua *term* yang berbeda. Jumlah kemunculan tersebut nantinya dinormalisasikan dengan mambagi setiap jumlah tersebut dengan jumlah terbesar. Nilai kedekatan makna dari *semantic term to term* nantinya dipergunakan sebagai nilai kedekatan makna yang menggantikan nilai *semantic relatedness* apabila kata/*term* tersebut tidak terdapat pada *database* "*WordNet*" atau nilai *semantic relatedness* menghasilkan nilai 0.

Searching

Proses ini berguna untuk mencari dokumen yang dicari oleh *user* sesuai dengan kata kunci yang dimasukkan oleh *user*. Pada proses ini menggabungkan nilai *weight* hasil dari proses *indexing* dengan nilai kedekatan makna, baik dari *semantic relatedness* ataupun dari *semantic term to term* dengan metode *Generalized Vector Space Model* (GVSM) perhitungan cosinus, untuk melakukan perankingan terhadap hasil pencarian.

Pengujian dilakukan dengan menggunakan data yang berjumlah 100 dokumen yang didapatkan dari TREC di atas yang seluruh datanya menggunakan bahasa Inggris sebagai obyek pencarian.

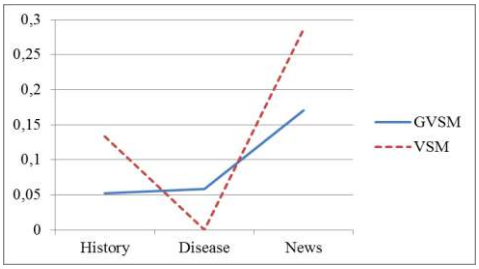
Pertama dilakukan pengujian program dengan memasukkan kata 'disease' dan 'news' sebagai kata kunci yang dipergunakan sebagai kata kunci untuk menguji hasil dari aplikasi pencarian dokumen berbasis *Generalized Vector Space Model* dan *Semantic Relatedness* ini. Hasil yang didapatkan dari proses *searching* dengan kata kunci 'disease' dan 'news' dapat dilihat pada Tabel 1 .

Tabel 1. Hasil Pencarian "Disease" dan "News"

Kata Kunci	Semua Dokumen Hasil Pencarian	Dokumen Relevan dari Hasil Pencarian	Dokumen Relevan dari Keseluruhan Dokumen
Disease	Dokumen 5, 80, 6, 76, 71, 26, 43, 94, 82, 28, 2, 17, 1, 64, 16, 62, 63	Dokumen 6	Dokumen 6
News	Dokumen 50, 70, 10, 61, 60, 79, 25, 77, 78, 96, 44, 52, 65, 29, 98, 58, 59, 18, 22, 69, 39, 26, 80, 99, 7, 40, 90, 100, 8, 83, 17, 16, 62, 13, 63, 14, 43, 64	Dokumen 7, 16, 40, 69, 77, 90, 100	Dokumen 7, 15, 16, 28, 40, 69, 77, 90, 100

Dari Tabel 1 dapat dilihat bahwa aplikasi ini dapat mengeluarkan hasil pencarian dokumen yang relevan.

Pengujian yang kedua dilakukan dengan membandingkan nilai *precision* dan *recall* dari pencarian dengan metode GVSM baru (GVSM & SR) dan VSM. Hasil yang didapatkan dari pengujian *precision* dari pencarian dengan metode GVSM dan VSM dengan kata kunci 'history', 'disease', dan 'news' dapat dilihat pada Gambar 2.

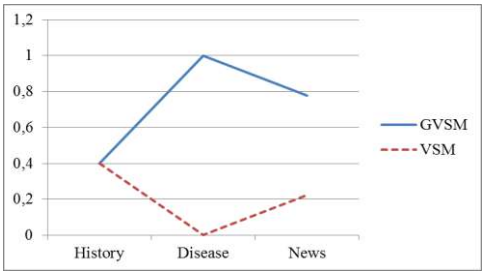


Gambar 2. Grafik perbandingan nilai *Precision* antara GVSM dan VSM

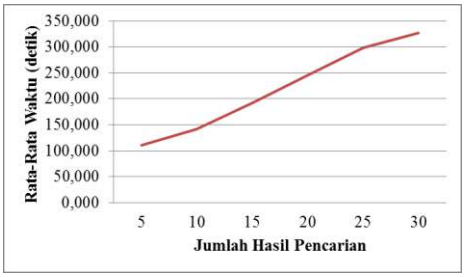
Dapat dilihat pada gambar 2 bahwa GVSM memiliki nilai *precision* yang lebih kecil jika dibandingkan dengan VSM. Nilai precision yang dihasilkan oleh GVSM adalah 0,0526 ; 0,0588 ; 0,1707 , sedangkan nilai precision yang dihasilkan oleh VSM adalah 0,1333 ; 0 ; 0,2857 . Hanya pada kata kunci "Disease" saja yang nilai *precision* GVSM-nya lebih tinggi jika dibanding dengan nilai *precision* VSM, dikarenakan tidak diketemukan sama sekali dokumen yang relevan pada hasil pencarian VSM.

Dapat dilihat pada gambar 3 bahwa GVSM memiliki nilai *recall* yang selalu lebih besar atau sama jika dibandingkan dengan VSM. Nilai *recall* yang dihasilkan oleh GVSM adalah 0,4 ; 1 ; 0,7778 , sedangkan nilai *recall* yang dihasilkan oleh VSM adalah 0,4 ; 0 ; 0,2222. Peningkatan recall terjadi karena Generalized Vector Space Model tidak hanya menampilkan dokumen yang mengandung keyword yang dimasukkan user saja, tetapi juga menampilkan dokumen yang mengandung keyword lain yang memiliki similiarity makna dengan keyword user.

Pengujian yang ketiga adalah pengujian waktu *Semantic Relatedness* (SR). Pengujian waktu SR ini dilakukan dengan menghitung rata-rata waktu proses pencarian nilai SR. Rata-rata waktu proses ini didapatkan dengan membagi total waktu yang dibutuhkan dalam sebuah proses dengan jumlah hasil yang didapatkan dari proses tersebut. Hasil pengujian tersebut disajikan pada Gambar 4.



Gambar 3. Grafik perbandingan nilai *Recall* antara GVSM dan VSM



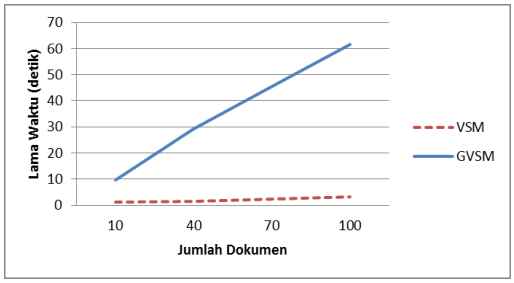
Gambar 4. Grafik rata-rata waktu proses pencarian nilai SR

Dari gambar 4 dapat kita lihat bahwa rata-rata waktu proses terus meningkat secara linear terhadap jumlah hasil pencarian. Jadi semakin banyak hasil pencarian yang dibutuhkan, maka semakin banyak pula rata-rata waktu untuk melakukan proses tersebut, sehingga semakin banyak waktu yang dibutuhkan untuk melakukan proses untuk mendapatkan hasil pencarian nilai SR tersebut.

Pengujian yang keempat adalah pengujian waktu proses *Searching*. Pengujian waktu *Searching* ini dilakukan dengan menghitung waktu setiap proses yang dilakukan dalam proses *seraching* dengan metode GVSM dan juga pada proses *searching* dengan metode VSM. Hasil pengujian tersebut disajikan pada Gambar 5.

Dari hasil perbandingan waktu *searching* pada gambar 5 maka dapat kita lihat bahwa proses *searching* dengan menggunakan metode GVSM memiliki waktu yang jauh lebih lama jika dibandingkan dengan waktu proses *searching* dengan menggunakan metode VSM. Hal ini bisa dilihat pada Gambar 5, dimana untuk melakukan *searching* dengan metode GVSM dengan 10 dokumen sebagai obyek pencariannya membutuhkan waktu yang lebih lama jika dibandingkan dengan melakukan *searching* dengan metode VSM dengan 100 dokumen sebagai obyek

pencarian. Hal ini dikarenakan pada GVSM terdapat proses mencari nilai kedekatan makna yang membutuhkan waktu yang lama dan waktu tersebut berpengaruh pada bertambahnya total waktu pencarian GVSM jika dibanding dengan pencarian dengan VSM.



Gambar 5. Grafik jumlah dokumen terhadap waktu *searching* GVSM & VSM

Pengujian yang terakhir adalah Pengujian jumlah *keyword user*. Pengujian jumlah *keyword user* ini dilakukan untuk menguji hasil pencarian yang dihasilkan oleh aplikasi, apabila *user* memasukkan *keyword* yang terdiri dari satu kata atau lebih. Proses pengujian ini dilakukan dengan membandingkan hasil pencarian yang diberikan oleh sistem dengan jumlah *keyword* 1 kata, 2 kata dan juga 3 kata. Hasilnya dapat dilihat pada Tabel 6.

Tabel 2. Hasil pengujian jumlah keyword

No	Kata Kunci	Hasil Pencarian	Jumlah Dokumen
1	Disease	Dokumen 5, 80, 6, 76, 71, 26, 43, 94, 82, 28, 2, 17, 1, 64, 16, 62, 63	17 Dokumen
2	Lethal	Tidak ada	0 Dokumen
3	Medicine	Dokumen 19, 11, 71, 20, 8, 64, 26, 80, 16	9 Dokumen
4	Lethal Disease	Dokumen 5, 80, 6, 76, 71, 26, 43, 94, 82, 28, 2, 17, 1, 64, 16, 62, 63	17 Dokumen
5	Disease Medicine	Dokumen 19, 11, 71, 5, 20, 80, 8, 6, 26, 76, 64, 16, 43, 94, 82, 28, 2, 17, 1, 62, 63	21 Dokumen
6	Lethal Disease Medicine	Dokumen 19, 11, 71, 5, 20, 80, 8, 6, 26, 76, 64, 16, 43, 94, 82, 28, 2, 17, 1, 62, 63	21 Dokumen
7	Common	Dokumen 67, 6, 74, 100, 89, 31, 66, 28, 87, 88, 17, 80, 63, 43	14 Dokumen
8	Common Disease	Dokumen 6, 5, 67, 80, 74, 100, 89, 31, 66, 28, 76, 87, 17, 88, 71, 43, 63, 26, 94, 82, 2, 1, 64, 16, 62	25 Dokumen

Dari tabel 2 dapat dilihat bahwa:

1. Kata kunci yang pertama "*Disease*" mendapatkan 17 dokumen sebagai hasil pencarian.
2. Kata kunci yang kedua "*Lethal*" tidak mendapatkan hasil pencarian
3. Kata kunci yang kedua "*Medicine*" mendapatkan 9 dokumen sebagai hasil pencarian.
4. Kata kunci keempat, dengan dua suku kata, yaitu "*Lethal Disease*" mendapatkan 17 dokumen sebagai hasil pencarian. Dari hasil ini dapat dilihat bahwa hasil pencarian "*Lethal Disease*" ini didapatkan dari hasil pencarian dengan kata kunci "*Lethal*" yang menghasilkan hasil pencarian sebesar 0 dokumen dan "*Disease*" yang menghasilkan hasil pencarian sebesar 17 dokumen. Sehingga hasil pencarian dengan kata kunci "*Lethal Disease*" sebesar 17 dokumen.
5. Kata kunci kelima, dengan dua suku kata, yaitu "*Disease Medicine*" mendapatkan 21 dokumen sebagai hasil pencarian. Dari hasil ini dapat dilihat bahwa hasil pencarian "*Disease Medicine*" ini didapatkan dari gabungan kata kunci "*Disease*" yang menghasilkan hasil pencarian sebesar 17 dokumen dan "*Medicine*" yang menghasilkan hasil pencarian sebesar 9 dokumen, serta 5 dokumen yang merupakan irisan dari

kedua hasil tersebut. Sehingga hasil pencarian dengan kata kunci "*Disease Medicine*" sebesar 21 dokumen.

6. Kata kunci keenam, dengan tiga suku kata, yaitu "*Lethal Disease Medicine*" mendapatkan 21 dokumen sebagai hasil pencarian. Dari hasil ini dapat dilihat bahwa hasil pencarian "*Lethal Disease Medicine*" ini didapatkan dari gabungan kata kunci "*Lethal*" yang menghasilkan hasil pencarian sebesar 0 dokumen, "*Disease*" yang menghasilkan hasil pencarian sebesar 17 dokumen dan "*Medicine*" yang menghasilkan hasil pencarian sebesar 9 dokumen, serta 5 dokumen yang merupakan irisan dari hasil pencarian "*Disease*" dan "*Medicine*". Sehingga hasil pencarian dengan kata kunci "*Lethal Disease Medicine*" sebesar 21 dokumen.
7. Kata kunci ketujuh "*Common*" mendapatkan 14 dokumen sebagai hasil pencarian.
8. Kata kunci "*Common Disease*" mendapatkan 25 dokumen sebagai hasil pencarian. Dari urutan perankingan terhadap *keyword* ini terdapat peningkatan peringkat dokumen nomor 6. Pada hasil pencarian dengan *keyword* "*Common*", dokumen nomor 6 terdapat pada peringkat kedua dan pada hasil pencarian dengan *keyword* "*Disease*", dokumen nomor 6 terdapat pada peringkat ketiga. Tetapi pada hasil pencarian dengan *keyword* "*Common Disease*", dokumen nomor 6 terdapat pada peringkat pertama. Dengan ini dapat kita lihat bahwa pencarian dengan *keyword* lebih dari satu dapat meningkatkan peringkat dokumen yang relevan.

Aplikasi pencarian dokumen berbasis Generalized Vector Space Model dan Semantic Relatedness dapat dilihat pada gambar 6 dan gambar 7.

Gambar 6. Tampilan halaman utama dari aplikasi

Gambar 7. Tampilan hasil pencarian dari aplikasi

KESIMPULAN

Berdasarkan hasil pengujian yang dilakukan pada sistem menggunakan data TREC dengan sampel 100 dokumen, maka dapat disimpulkan bahwa :

1. Dengan melakukan perbandingan antara *Generalized Vector Space Model* (GVSM) dan *Vector Space Model* (VSM), maka dapat dilihat bahwa *Generalized Vector Space Model* dapat membantu dalam meningkatkan *recall*.
2. Kelemahan dari *Generalized Vector Space Model* adalah kecilnya *precision* dari hasil pencarian jika dibandingkan dengan *Vector Space Model*.
3. Berdasarkan pegujian lama waktu pencarian nilai SR, dapat dilihat bahwa rata-rata waktu proses terus meningkat secara linear terhadap jumlah hasil pencarian. Jadi semakin banyak hasil pencarian yang dibutuhkan, maka semakin banyak pula rata-rata waktu untuk melakukan proses tersebut, sehingga semakin banyak waktu yang dibutuhkan untuk melakukan proses untuk mendapatkan hasil pencarian nilai SR tersebut.

4. Berdasarkan pengujian lama waktu *searching*, dapat dilihat bahwa jumlah dokumen berbanding lurus secara linear dengan lama waktu *searching*.
5. Berdasarkan pembandingan waktu *searching* antara *Generalized Vector Space Model* (GVSM) dan *Vector Space Model* (VSM), maka dapat dilihat bahwa lama proses *searching* dengan GVSM jauh lebih lama jika dibandingkan dengan lama proses *searching* dengan VSM. Dikarenakan proses *searching* dengan GVSM membutuhkan waktu untuk pencarian kedekatan makna antar *term*.
6. Kemampuan aplikasi ini sangat bergantung pada database "*WordNet*" yang dipergunakan.

DAFTAR PUSTAKA

[1] Dik L.L., Huei C., Kent E. S. Document ranking and the vector-Space Model.1997

[2] Harjono K.D. Perluasan Vektor pada Metode Search Vector Space. Integral, Vol. 10 No. 2, Juli 2005.

[3] Miller, G. A. WordNet : A Lexical Database for English. 1995

[4] Ning Liu et al. Learning Similarity Measures in Non-orthogonal Space. CIKM'04, November 8-13, 2004, Washington D.C., U.S.A.

[5] Garcia E. The Classic Vector Space Model. Retrieved URL:http://www.miislita.com/term-vector/term-vector-3.html, diakses tanggal 15 Maret 2012.

[6] Tsatsaronis, G., Panagiotopoulou V. A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness. The EACL 2009 Student Research Workshop, 70–78. 2009.

[7] Turney, P.D. & Pantel, P. From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research. 37: 141-188. 2010.