Perbandingan Metode Latent Semantic Analysis, Syntactically Enhanced Latent Semantic Analysis, dan Generalized Latent Semantic Analysis dalam Klasifikasi Dokumen Berbahasa Inggris

Gilbert Wonowidjojo
Bina Nusantara University
School of Computer Science
Jakarta, Indonesia
gwonowidjojo@gmail.com

Frendy
Bina Nusantara University
School of Computer Science
Jakarta, Indonesia
frendyguo@gmail.com

Michael Sean Hartono
Bina Nusantara University
School of Computer Science
Jakarta, Indonesia
michael.s.hartono@gmail.com

Abstract

Teknik Latent Semantic Analysis (LSA) banyak digunakan dalam Natural Language Processing. Contoh penggunaan LSA adalah dalam penilaian esai, sumarisasi, dan klasifikasi dokumen secara otomatis. Penelitian ini membandingkan tiga teknik yang berbasiskan LSA untuk melakukan klasifikasi dokumen secara otomatis, yaitu LSA (Landauer, Foltz & Laham, 1998), Syntactically Enhanced LSA (Kanejiya, Kumar & Prasad, 2003), dan Generalized LSA (Islam & Hoque, 2012). Ketiga teknik ini akan menyusun *word by document matrix (n-gram by document matrix* untuk *Generalized LSA*) yang kemudian didekomposisi dengan teknik *Singular Value Decomposition* (SVD) dan diukur kedekatannya dengan menggunakan *Cosine Similarity*. Hasil penelitian menunjukkan bahwa ketiga teknik ini dapat melakukan klasifikasi dokumen berbahasa Inggris secara otomatis dengan akurasi rata-rata yaitu 77%.

1. Introduction

Kemampuan untuk melakukan klasifikasi terhadap dokumen sehingga dapat dikelompokkan menjadi kategori tertentu memegang peran penting dalam menyelesaikan permasalahan mengorganisir, mengklasifikasi, dan merepresentasikan informasi dalam jumlah yang besar. Setiap dokumen akan memiliki nilai guna yang lebih tinggi apabila dikelompokkan dengan dokumen lain yang memiliki jenis yang sama. Hal tersebut dikarenakan pengelompokkan dokumen dapat memudahkan pengguna dalam membaca dokumen sejenis dan dapat mempercepat pengumpulan informasi .

Dalam Natural Language Processing (NLP), terdapat teknik-teknik untuk melakukan klasifikasi dokumen secara otomatis. Teknik-teknik tersebut didasarkan kepada hipotesis bahwa dokumen yang berbeda kategori dapat dibedakan oleh fitur dari bahasa alami (*Natural Language*) yang terkandung dalam setiap dokumen. Fitur Klasifikasi Dokumen tersebut dapat berupa pengukuran struktur kata, frekuensi kata dan struktur bahasa alami dalam setiap dokumen.

Pada penelitian ini, klasifikasi dokumen dilakukan terhadap artikel-artikel berbahasa Inggris yang berdasarkan 4 topik yaitu Kesehatan, Politik, Olahraga dan Teknologi yang didapatkan dari berita-berita dunia terkini. Tujuannya adalah untuk membandingkan ketepatan klasifikasi antara metode *Latent Semantic Analysis* (LSA), *Syntactically Enhanced*

Latent Semantic Analysis (SELSA) dan Generalized Latent Semantic Analysis (GLSA). Dokumen-dokumen yang sudah diklasifikasi secara manual oleh manusia akan dibutuhkan pada tahap awal penelitian sehingga dapat dilakukan *Supervised Classification.*

2. Tinjauan Pustaka

Banyak sekali teknik-teknik yang sudah pernah dicoba oleh para peneliti untuk menciptakan sebuah sistem pengklasifikasi dokumen secara otomatis dengan harapan dapat mencapai akurasi yang tinggi, salah satunya adalah *Naive Bayes Classifier*, yaitu pendekatan probabilitas yang menciptakan asumsi kuat tentang bagaimana suatu data dihasilkan dan berdasarkan model probabilitas untuk mewujudkan asumsi-asumsi tersebut. Dengan menggunakan teknik *Supervised Learning* pada data training sampel untuk memprediksi suatu parameter, *Bayes' rule* akan melakukan klasifikasi dokumen baru dengan memilih kategori yang paling mungkin menghasilkan contoh dokumen tersebut. (McCallum & Nigam, 1998).

Nigam, Lafferty dan McCallum (1999) menggunakan teknik *Maximum Entropy* untuk melakukan klasifikasi dokumen. *Maximum Entropy* melakukan estimasi distribusi probabilitas dari data. Prinsipnya adalah, ketika tidak terdapat informasi apa-apa, maka distribusi probabilitas dalam pemilihan kategori adalah seimbang. Contoh yang diberikan adalah sebagai berikut, misalkan terdapat suatu klasifikasi 4 kategori dimana yang hanya diketahui bahwa pada rata-rata 40% dokumen dengan kata "profesor" didalamnya, merupakan kelas "fakultas". Maka, ketika diberikan dokumen yang terdapat kata "profesor" didalamnya, dapat dikatakan bahwa kemungkinan dokumen tersebut masuk kedalam kelas "fakultas" adalah 40%, dan 20% pada 3 kelas yang lain. Namun jika suatu dokumen tidak terdapat kata "profesor", maka dengan melakukan *uniform class distribution* (distribusi kelas beraturan), akan terdapat peluang 25% pada setiap kelas. *Maximum Entropy* mengestimasi distribusi kondisional dari kategori kelas berdasarkan dokumen yang diberikan. Suatu dokumen disajikan dengan fitur penghitungan kata, data yang telah dilakukan *training* digunakan untuk mengestimasi nilai ekspektasi dari penghitungan kata pada basis kelas ke kelas.

Landauer dan Dumais (1997) mengajukan teknik yang bernama *Latent Semantic Analysis* (LSA). Cara kerja LSA ialah dengan menghasilkan sebuah model yang didapat dengan mencatat kemunculan-kemunculan kata dari tiap-tiap dokumen yang direpresentasikan dalam sebuah matriks yang dinamakan *term-document matrix*, setelah itu dilakukan proses *Singular Value Decomposition* (SVD) yang akan digunakan untuk mendapatkan *Cosine Similarity* (nilai kemiripan) antara satu dokumen dengan dokumen yang lain (Landauer, Foltz, & Laham, 1998).

Syntactically Enhanced Latent Semantic Analysis (SELSA) dikembangkan oleh Kanejiya, Kumar dan Prasad (2003), untuk memperbaiki performa dari LSA. Dasar dari teknik SELSA adalah bahwa LSA mengabaikan informasi sintaktik berupa urutan kata (*word order*) yang terdapat pada suatu dokumen, karena pendekatan yang digunakan oleh LSA adalah pendekatan *bag-of-words*. SELSA menginkorporasikan informasi sintaktik berupa urutan kata tersebut dalam

penyusunan matriks. Sebagai contoh, sebuah kata yang akan disimpan kedalam matriks, akan dipasangkan terlebih dahulu informasi *part-of-speech* (POS) dari kata yang mendahului kata tersebut (disebut *prevtag*). Hal ini dilakukan untuk membedakan kata yang sama namun memiliki arti yang berbeda dikarenakan *prevtag* yang berbeda. Dengan menelusuri perbedaan tersebut, SELSA dapat mengurangi ambiguitas dari suatu kalimat dalam dokumen apabila akan dibandingkan dengan dokumen lain untuk mendapatkan nilai kemiripannya.

Berbeda dengan LSA yang hanya memanfaatkan kemunculan satu kata, Islam dan Hoque (2012) melakukan modifikasi terhadap cara kerja LSA yang dikenal dengan *Generalized Latent Semantic Analysis* (GLSA), dimana *n-gram* digunakan untuk membentuk matriks yang tidak hanya terdiri dari satu kata tetapi juga bisa lebih dari satu kata. Sebuah *n-gram* berukuran 1 dinamakan "*Unigram*", ukuran 2 "*Bigram*", ukuran 3 "*Trigram*" dan seterusnya. Ide dibalik *n-gram* ini adalah untuk memastikan bahwa kata-kata yang muncul bersamaan akan mendapatkan nilai yang lebih tinggi dibandingkan kata tunggal. Tidak berbeda dengan teknik LSA, teknik SELSA dan GLSA juga akan melakukan SVD terhadap matriks yang sudah dibentuk untuk mendapatkan nilai kemiripan antar dokumen.

3. Metodologi

Penelitian ini berfokus pada performa tiga metode (LSA, SELSA, dan GLSA) dalam melakukan klasifikasi dokumen berbahasa Inggris. Dokumen yang akan diklasifikasikan adalah artikel-artikel yang membahas tentang topik tertentu. Dalam penelitian ini topik yang digunakan berjumlah 4 topik, yaitu *Sports*, *Politics*, *Health*, dan *Technology*. Artikel yang digunakan untuk membuat model LSA, SELSA, dan GLSA serta testing semuanya berhubungan dengan topik-topik tersebut.

Setiap artikel yang dikumpulkan, dilakukan preprocessing secara manual untuk menghilangkan huruf-huruf yang tidak standar, misalnya huruf yang memiliki tanda ~ atau tanda .. diatas huruf tersebut, dan juga menghilangkan simbol-simbol lain yang tidak dapat diproses dalam aplikasi. Setelah melalui tahapan *preprocessing* secara manual barulah dilakukan *preprocessing* secara otomatis. Tahapan ini meliputi *stopwords removal*, *token checking* (memastikan setiap kata hanya terdiri dari huruf saja), dan membuat seluruh huruf menjadi *lowercase*.

Setelah itu, dilakukan penyusunan *term-document matrix (ngram-document matrix* untuk GLSA), dengan cara yang berbeda untuk setiap metode dalam penentuan term / n-gram nya, yaitu :

- Pada LSA, setiap kata yang ada setelah melalui preprocessing langsung menjadi term.
- Pada SELSA, setiap kata yang ada setelah melalui preprocessing akan dipasangkan dengan *Part-of-Speech Tag* dari kata yang mendahuluinya (disebut *prevtag*), sehingga setiap kata yang sama dapat memiliki beberapa makna yang berbeda, kemudian baru menjadi term.
- Pada GLSA, setiap kata yang ada setelah melalui preprocessing akan dikelompokkan satu-satu (*unigram*), dua-dua (*bigram*), dan tiga-tiga (*trigram*), setelah itu akan dikelompokkan dan dijadikan term.

Setelah matriks tersusun, dilakukan *Singular Value Decomposition* untuk memecah matriks menjadi ukuran yang lebih kecil, yang menyatakan model LSA, SELSA, dan GLSA yang sudah dibuat. Kemudian, untuk setiap query yang masuk (berupa dokumen yang akan diklasifikasikan), akan dipreprocessing dengan cara yang sama, lalu dilakukan penghitungan *Cosine Similarity* antara *query* dengan model LSA, SELSA, dan GLSA pada setiap topik. Hasil klasifikasi dokumen dapat ditentukan oleh dua kriteria, yaitu :

- *Maximum Similarity*, yaitu mencari dokumen yang memiliki *similarity* tertinggi dengan query yang dimasukkan. Topik dari dokumen tersebut akan menjadi hasil klasifikasi dokumen.
- *Average Similarity*, yaitu mencari topik yang memiliki *similarity* tertinggi dengan query yang dimasukkan. Caranya adalah dengan menjumlahkan nilai kedekatan dari setiap dokumen dengan *query*, yang akan menjadi nilai kedekatan dari topik tersebut dengan *query*.

4. Hasil dan Pembahasan

Semua dokumen yang digunakan dalam penelitian ini didapat dari berbagai sumber di internet, diantaranya termasuk dari *British Broadcasting Corporation* (http://www.bbc.com/) dan *Cable News Network* (http://edition.cnn.com/) . Untuk membentuk model LSA, SELSA, dan GLSA digunakan sebanyak 40 dokumen, yang sudah diklasifikasikan terlebih dahulu (*pre-classified*) secara manual ke dalam 4 topik yaitu *Sports, Politics, Health,* dan *Technology* (masing-masing 10 dokumen). Sedangkan untuk melakukan testing terhadap aplikasi yang telah dibuat digunakan sebanyak 100 dokumen yang juga sudah *pre-classified* secara manual, yang terdiri dari 40 dokumen dengan satu topik utama, dan 60 dokumen dengan dua topik utama. Rinciannya adalah sebagai berikut :

- 10 dokumen dengan topik "*Health*",
- 10 dokumen dengan topik "*Politics*",
- 10 dokumen dengan topik "*Sports*",
- 10 dokumen dengan topik "*Technology*",
- 10 dokumen dengan topik "*Health*" dan "*Politics*",
- 10 dokumen dengan topik "*Health*" dan "*Sports*",
- 10 dokumen dengan topik "*Health*" dan "*Technology*",
- 10 dokumen dengan topik "*Politics*" dan "*Sports*",
- 10 dokumen dengan topik "*Politics*" dan "*Technology*", dan
- 10 dokumen dengan topik "*Sports*" dan "*Technology*".

Berikut adalah hasil ujicoba dari klasifikasi dokumen dengan satu topik utama, yang berjumlah sebanyak 40 dokumen (10 dokumen untuk masing-masing topik). Akurasi klasifikasi dengan menggunakan *Maximum Similarity* terdapat pada tabel 1, sedangkan akurasi klasifikasi dengan menggunakan *Average Similarity* terdapat pada tabel 2.

Topik	LSA	SELSA	GLSA
Health	50%	50%	70%
Politics	50%	60%	60%
Sports	60%	60%	80%
Technology	40%	60%	30%

Tabel 1 : Akurasi klasifikasi dokumen dengan satu topik utama untuk setiap metode, menggunakan *Maximum Similarity*

Topik	LSA	SELSA	GLSA
Health	80%	70%	70%
Politics	90%	60%	80%
Sports	90%	90%	90%
Technology	20%	30%	20%

Tabel 2 : Akurasi klasifikasi dokumen dengan satu topik utama untuk setiap metode, menggunakan *Average Similarity*

Berikut adalah hasil ujicoba dari klasifikasi dokumen dengan dua topik utama, yang berjumlah sebanyak 60 dokumen (sesuai dengan rincian diatas). Akurasi klasifikasi dengan menggunakan *Maximum Similarity* terdapat pada tabel 3, sedangkan akurasi klasifikasi dengan menggunakan *Average Similarity* terdapat pada tabel 4. Klasifikasi dokumen dengan dua topik utama untuk GLSA masih dalam tahapan testing, sehingga hasilnya belum dapat ditampilkan.

Topik	LSA	SELSA
Health - Politics	60%	80%
Health - Sports	90%	80%
Health - Technology	100%	90%
Politics - Sports	80%	70%
Politics - Technology	100%	90%
Sports - Technology	80%	90%

Tabel 3 : Akurasi klasifikasi dokumen dengan dua topik utama untuk metode LSA dan SELSA, menggunakan *Maximum Similarity*

Topik	LSA	SELSA
Health - Politics	80%	80%
Health - Sports	90%	100%
Health - Technology	100%	100%
Politics - Sports	100%	100%
Politics - Technology	100%	100%
Sports - Technology	90%	100%

Tabel 4 : Akurasi klasifikasi dokumen dengan dua topik utama untuk metode LSA dan SELSA, menggunakan *Average Similarity*

Tabel-tabel hasil ujicoba di atas dapat disimpulkan dalam tabel 5.

	1 Topik	2 Topik
LSA_Max	50%	85%
LSA_Avg	70%	93.3%
SELSA_Max	57.5%	83.3%
SELSA_Avg	62.5%	96.7%
GLSA_Max	60%	
GLSA_Avg	65%	

Tabel 5 : Kesimpulan akurasi klasifikasi dokumen untuk setiap metode dan setiap kriteria *similarity*

Dari hasil uji coba, dapat dilihat bahwa *Average Similarity* lebih cocok digunakan untuk melakukan klasifikasi dokumen dibandingkan dengan *Maximum Similarity*. Hal ini dikarenakan *Average Similarity* merupakan cerminan model dari sebuah topik, sedangkan *Maximum Similarity* hanya melihat dokumen-dokumen dalam setiap topik secara individual, bukan secara keseluruhan.

Apabila hanya memperhitungkan *Average Similarity*, maka LSA memiliki akurasi yang tertinggi, yaitu 84%, sedangkan apabila memperhitungkan kedua kriteria similarity maka SELSA memiliki akurasi yang tertinggi, yaitu 78%. Dari hasil uji coba juga dapat dilihat bahwa klasifikasi dokumen dengan topik *Technology* akurasinya sangat kecil. Dapat disimpulkan bahwa topik tersebut sulit untuk berdiri sendiri, sehingga klasifikasi dokumen secara otomatis dengan metode LSA dalam topik tersebut cenderung kurang tepat.

Selain itu, klasifikasi dokumen dengan satu topik utama masih memiliki akurasi yang kurang memuaskan dibandingkan dengan dokumen dengan dua topik utama. Hal ini disebabkan jumlah dokumen *training* yang masih terlalu sedikit, sehingga model LSA, SELSA, dan GLSA yang dibuat masih kurang mewakili topik yang direpresentasikan. Harapannya dengan menambah jumlah dokumen *training*, model yang dibuat akan lebih baik sehingga dapat melakukan klasifikasi dokumen dengan lebih baik.

5. Simpulan

Simpulan yang dapat diperoleh dari penelitian yang dilakukan ini adalah sebagai berikut :

- Untuk melakukan klasifikasi dokumen secara otomatis lebih baik menggunakan *Average Similarity* dibandingkan dengan *Maximum Similarity*, karena *Average Similarity* mencerminkan hubungan dokumen kepada model topik secara keseluruhan.
- Topik-topik yang dipilih harus topik yang kuat, sehingga ketika model dari topik tersebut dibuat dapat dibedakan dengan mudah dari topik-topik yang lain.
- Untuk mendapatkan akurasi yang tinggi, model LSA, SELSA, dan GLSA dari setiap topik harus kuat. Salah satu cara untuk memperkuatnya adalah dengan menambah jumlah dokumen yang membentuk model tersebut.

Beberapa perbaikan dan usulan yang bisa dikerjakan untuk penelitian lanjutan adalah :

- Mengoptimisasi algoritma GLSA agar dapat bekerja secara lebih cepat (tidak memakan waktu lama saat testing).
- Menambah jumlah dokumen pembentuk model LSA, SELSA, dan GLSA agar setiap topik dapat lebih mudah dibedakan oleh aplikasi.
- Mencari kriteria *similarity* yang lain disamping *Maximum Similarity* dan *Average Similarity*.

Reference

Islam,M.M., & Hoque,A.S.M.L. (2012). Automated essay scoring using generalized latent semantic analysis. *Journal of Computers*, Academy Publisher, vol.7, no,3, pp.616-626.

Kanejiya, D., Kumar, A., & Prasad, S. (2003). Automatic evaluation of students' answers using syntactically enhanced LSA. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing - Volume 2*, pages 53-60.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes,* 25, 259-284.

McCallum, A. & Nigam,K. (1998). A comparison of event models for naive Bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization.* Tech. Rep. WS-98-05, AAAI Press.

Nigam, K., Lafferty, J.,& McCallum,A. (1999). Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61-67.