IMPLEMENTASI METODE PROBABILISTIC LATENT SEMANTIC ANALYSIS UNTUK OPINION RETRIEVAL

Yusup Miftahuddin [1],Jasman Pardede [2] , Afdhalul Zikri [3]

Jurusan Teknik Informatika, Fakultas Teknik Industri, Itenas Bandung
Jln. PHH. Mustopha No.23 Bandung 40124 Telp. 022.772215
(1) yusufm@itenas.ac.id, (2) jasman@itenas.ac.id, (3) afdhalul.student91@gmail.com

ABSTRACT

Opinion retrieval is a search system by the user, where in the information needed is more of opinion than a fact. The method used for opinion retrieval system is probabilistic latent semantic analysis (PLSA). The search proses opinion sentences with opinion retrieval system have some steps, the documents processed in text processing, should be form matrix value term. The PLSA method gives matrix values to calculate e-step, m-step, decomposition matrix, and likelihood value. The similarities calculating process in opinion sentence and *query* used *cosine similarity formula.* So it has similarity value as identified opinion sentence. The result of testing in document with 5 query words, it has highest score for kappa statistic testing 0.152432875 with kappa slight interpretation.

Key words: *Pre-processing Text, Opinion Retrieval, PLSA, Query, Cosine Similarity.*

1. Pendahuluan

Opini adalah pendapat, ide atau pikiran untuk menjelaskan kecenderungan atau preferensi tertentu terhadap perspektif dan ideologi akan tetapi bersifat tidak objektif karena belum mendapatkan pemastian atau pengujian, dapat pula merupakan sebuah pernyataan tentang sesuatu yang berlaku pada masa depan dan kebenaran atau kesalahannya serta tidak dapat langsung ditentukan. Dengan opini, dapat dinilai tanggapan setiap orang terhadap produk atau peristiwa yang sedang hangat diperbincangkan. Kalimat opini banyak terdapat didalam dokumen atau artikel. Oleh karena itu, diperlukan sebuah cara untuk menata dan mengakses opini-opini tersebut secara efektif dan efisien, yaitu dengan sistem opinion retrieval. Opinion retrieval dapat dijadikan cara untuk mendapatkan opini-opini terkait dengan topik yang sesusai dengan keinginan pembaca didalam dokumen atau artikel.

Sistem opinion retrieval berfungsi untuk mengembalikan dokumen yang mengandung opini tentang sebuah query pencarian dari sebuah korpus data teks. Pada dasarnya, Opinion Retrieval sama seperti information retrieval, tetapi dengan target pengembalian yang berbeda. Opinion retrieval menggabungkan teknik information retrieval dengan opinion mining. Secara umum, proses opinion retrieval dapat dijelaskan dengan dua tahap pendekatan. Pertama, teknik information retrieval digunakan untuk mendapatkan dokumen yang relevan dengan query. Kedua, teknik opinion mining diaplikasikan pada dokumen yang telah ter-retrieve untuk kemudian diidentifikasi, opini-opini yang terdapat di dalamnya dan diperkirakan relevansi antara opini-opini tersebut dengan query yang diberikan. Langkah terakhir adalah perankingan ulang pada dokumen.

Sistem opinion retrieval ini menggunakan metode PLSA (Probabilistic Latent Semantic Analysis). PLSA dapat memproses banyak kata-kata didalam dokumen atau artikel, sehingga memudahkan user dalam proses pencarian dengan query yang diberikan. Sistem opinion retrieval dengan menggunakan metode PLSA ini diharapkan dapat mencari opini-opini dengan tepat dan akurat.

1.1 Rumusan Masalah

Berdasarkan latar belakang tersebut, dapat dirumuskan beberapa masalah sebagai berikut :

1. Bagaimana membangun sebuah sistem *opinion retrieval* dengan menggunakan metode PLSA.
2. Bagaimana perhitungan dan penerapan metode PLSA untuk sistem *opinion retrival*.
3. Bagaimana hasil dari sistem *opinion retrieval* dengan menggunakan metode PLSA.

1.2 Tujuan

Penelitian ini dilakukan bertujuan untuk membangun sebuah aplikasi untuk sistem *opinion retrieval* pada sebuah dokumen berdasarkan *query* dengan menggunakan metode PLSA.

1.3 Batasan Masalah

Didalam penelitian ini terdapat beberapa batasan masalah, antara lain sebagai berikut:

1. Dokumen atau artikel yang digunakan berbahasa Indonesia dan mengandung kalimat opini.
2. *Query* yang digunakan untuk mencari informasi adalah Bahasa Indonesia yang baik dan benar (EYD).
3. Aplikasi yang dibuat berjalan secara *offline*.
4. Jenis dokumen yang dapat dilakukan pencarian ialah dokumen yang memiliki format *.doc, *.docx dan *.pdf.
5. Algoritma yang digunakan untuk proses s*temming* adalah algoritma nazief dan Adriani.

1.4 Landasan Teori

Berikut ini adalah teori-teori yang mendukung untuk melakukan penelitian pada tugas akhir ini.

1.4.1 *Opinion Retrieval*

Opinion retrieval merupakan sebuah sistem pencarian oleh *user*, dimana informasi yang dibutuhkan lebih merupakan sebuah opini dibandingkan sebuah fakta. Sebuah dokumen dikatakan relevan apabila sesuai dengan topik *query* dan juga mengandung opini yang berkenaan dengan *query*. Pada umumnya, sistem *opinion retrieval* menggunakan dua tahap pendekatan: teknik *information retrieval* tradisional untuk mendapatkan dokumen yang relevan; dan teknik *opinion mining* yang diaplikasikan pada dokumen yang didapatkan sebelumnya untuk mengidentifikasi opini, mengestimasi relevansi opini terhadap *query*, kemudian meranking ulang dokumen.

1.4.2 *Teks Preprocessing*

Teks preprocessing merupakan tahapan untuk mengolah dokumen sebelum memasuki proses ekstraksi. Proses yang dilakukan dalam dokumen terdiri dari :

1. memecah dokumen menjadi kalimat-kalimat.
2. *Case folding* mengubah semua karakter huruf menjadi huruf non-kapital.
3. *Tokenizing* memecah setiap kalimat ke dalam kata-kata.
4. *Filtering* menghilangkan kata-kata yang tidak terlalu mengandung makna penting.
5. *Stemming* mengambil bentuk kata dasar dengan cara menghilangkan imbuhannya. Untuk algoritma *stemming* yang digunakan yaitu algoritma Nazief dan Adriani [3&4].

2. *Probabilistic Latent Semantic Analisys* (PLSA)[1&2]

PLSA adalah sebuah metode pendekatan probabilitas untuk dua model seperti kata dan dokumen. PLSA merupakan penyempurnaan dari metode *Latent Semantic Analysis* (LSA). Metode ini merupakan teknik *information retrieval* yang berfungsi untuk menganalisis dua keterhubungan kejadian data yang berdasarkan *model statistic* yang disebut *aspect model*. *Aspect model* didefinisikan sebagai sebuah variabel yang tidak terlihat (*latent variable*) dari sebuah dokumen. Berikut ini merupakan persamaan pada metode PLSA.

$$P(d_i, w_j) = P(d_i)P(w_j \mid d_i), P(w_j \mid d_i) = \sum_{z \in Z} P(w_j \mid z_k)P(z_k \mid d_i) \quad ..(1)$$

Keterangan :
P(d) : Probabilitas terhadap dokumen *d*.
P (z|d) : Probabilitas terhadap topik *z* yang disesuaikan dengan dokumen *d*.

P (w|z) : Probabilitas terhadap kata *w* yang disesuaikan dengan topik *z*.

Nilai *P(d), P (z|d)* dan *P (w|z)* dapat ditentukan dengan cara memaksimalkan fungsi *likelihood L* seperti yang terdapat pada persamaan berikut.

$$L = \sum_{i=1}^{I} \sum_{j=1}^{J} n(d_i, w_j) \log P(d_i, w_j) \quad .....(2)$$

Keterangan :
n(d,w) : merupakan bobot term pada dokumen

Didalam metode perhitungan PLSA, terdapat algoritma yang disebut Algoritma *Expectation Maximization* (EM), algoritma ini digunakan untuk memperkirakan nilai maksimum *likelihood* dalam model variabel latent.Terdapat dua langkah dalam algoritma ini yaitu : langkah *Expectation* (E-step) dan langkah *Maximization* (M-step). Proses E-Step berfungsi untuk menghitung probabilitas posterior untuk variabel *z* berdasarkan pada perkiraan parameter saat itu, dan proses M-Step berfungsi untuk meng-*update* parameter yang digunakan untuk menghitung nilai probabilitas posterior

variabel *z*, yang akan digunakan dalam perhitungan nilai *likelihood*. Berikut ini merupakan persamaan pada proses E-Step.

$$P(z \mid d, w) = \frac{p(w_j \mid z_k)\, p(z_k \mid d_i)}{\sum_{k=1}^{K} p(w_j \mid z_k)\, p(z_k \mid d_i)} \quad (3)$$

Sedangkan persamaan untuk proses M-Step adalah sebagai berikut.

$$P(w \mid z) = \frac{\sum_{i=1}^{I} n(di, wj) P(zk \mid di, wj)}{\sum_{i=1}^{I} \sum_{j=1}^{J} n(di, wj) P(zk \mid di, wj)} \quad \text{.....(4)}$$

Dan

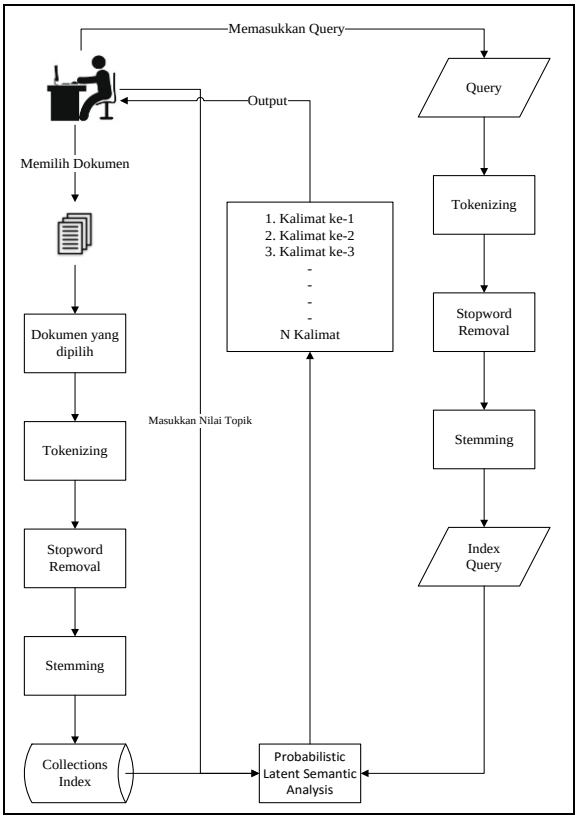
$$P(zk \mid di) = \frac{\sum_{j=1}^{J} n(di, wj)\, p(zk \mid di, wj)}{\sum_{j=1}^{J} \sum_{k=1}^{K} n(di, wj)\, p(zk \mid di, wj)}$$

3. Pembahasan

Pada sub bab pembahasan berisi tentang penjelasan penelitian yang dilakukan.

3.1 Analisis Sistem

Rancangan sistem aplikasi *opinion retrieval* dengan menggunakan metode PLSA terdapat pada Gambar 1.



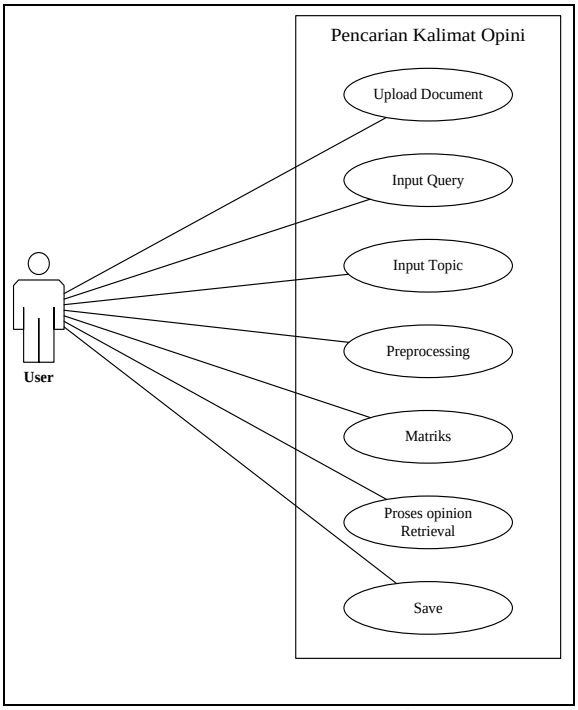
Gambar 1. Alur kerja apilikasi *opinion retrieval*

Tahapan dalam rancangan sistem pencarian kalimat opini adalah sebagai berikut.

1. *User* memasukkan dokumen dari *directory* penyimpanan dokumen.
2. *User* memasukkan *quey* yang diinginkan pada aplikasi pencarian kalimat opini.
3. *User* memasukkan jumlah topik yang diiginkan.
4. Sistem melakukan proses awal yaitu *teks preprocessing* untuk mengolah dokumen menjadi kalimat-kalimat, kemudian dilakukan proses *case folding*, *tokenizing*, *filtering* dan proses *stemming.*
5. Proses selanjutnya adalah sistem melakukan ekstraksi kalimat dimana hasil dari proses *stemming* akan dihitung bobot kemunculan kata pada tiap kalimat, sehingga menghasilkan matrik *term.*
6. Sistem akan menghitung nilai *e-step* dari matrik *random* dan nilai topik yang diberikan, selanjutnya menghitung nilai *m-step*, matrik *decomposition* dan nilai *likelihood.*
7. Selanjutnya sistem akan menghitung nilai kemiripan antara *query* dengan kalimat yang ada pada dokumen, kalimat yang memiliki nilai paling tinggi diidentifikasi sebagai kalimat opini.
8. Hasil pencarian kalimat opini ditampilkan sebagai *output* untuk *user*.

3.2 Use Case Diagram

Berdasarkan analisis sistem yang dilakukan, maka fungsionalitas yang dibutuhkan pada sistem pencarian kalimat opini pada sebuah dokumen menggunakan *Probabilistic Latent Semantic Analysis* (PLSA) yaitu fungsionalitas *upload*, *input query*, *input topic*, *preprocessing*, *matiks*, *process* dan *save.* Use case diagram dapat dilihat pada Gambar 2.



Gambar 2. Use case diagram

3.2 Pengujian Kappa *Statistics*

Pengujian kappa *statistics* merupakan pengujian hasil pencarian opini kunci dan pencarian opini pada aplikasi. Opini kunci didapat dari dari seseorang lulusan sastra dan Bahasa, sedangkan opini pada aplikasi didapat dari hasil pencarian kalimat opini pada aplikasi dengan menggunakan metode *probabilistic latent semantic analisys* (PLSA). Langkah untuk menghitung kappa *ststistic* yaitu menyusun kedua hasil pencarian opini terhadap objek penelitian kedalam tabel 2x2.

Tabel 1. Klasifikasi opini

		Opini Aplikasi		
		Ya	Tidak	Total
Opini Kunci	Ya	a	b	m_1
	Tidak	c	d	m_0
	Total	n_1	n_0	n

a dan *d* menyatakan jumlah kedua opini setuju, sedangkan *b* dan *c* menyatakan jumlah kedua opini tidak setuju. Ketika nilai *b* dan *c* bernilai 0 maka nilai *observed agreement* (*Po*) adalah 1 atau 100%, sebaliknya, jika *a* dan *d* bernilai 0 maka (*Po)* bernilai 0, selain itu n_1 menyatakan jumlah persetujuan opini kunci, sedangkan n_0 menyatakan jumlah total opini kunci tidak sesuai dengan hasil. Demikian halnya dengan m_1 dan m_0 secara berurutan keduanya menyatakan tingkat persetujuan dan ketidaksetujuan dari pencarian opini aplikasi. Rumus untuk menghitung Kappa *Statistics* seperti pada persamaan (5) , (6), dan (7).

$$K = \frac{(Po - Pe)}{(1 - Pe)} \quad \text{.....(5)}$$

Keterangan :
K = Menyatakan Nilai Kappa
Pe = *Expected Agreement*
Po = *Observed Agreement*

$$Pe = \left(\left(\frac{n1}{n}\right) * \left(\frac{m1}{n}\right)\right) + \left(\left(\frac{n0}{n}\right) * \left(\frac{m0}{n}\right)\right) \quad \text{.....(6)}$$

$$Po = \frac{a + d}{n} \quad \text{.....(7)}$$

Keterangan :
a, d : Menyatakan jumlah opini setuju
n_1 : Jumlah ya pada opini kunci
n_0 : Jumlah tidak pada opini kunci
m_1 : Jumlah ya pada opini aplikasi
m_0 : Jumlah tidak pada opini aplikasi

Tabel 2. Interpretasi Nilai Kappa

Interpretasi Nilai Kappa	
Nilai Kappa	Strength of Agreement
< 0	*Poor* (Lebih Rendah)
0 - 0.2	*Slight* (Rendah)
0.21 - 0.4	*Fair* (Cukup)
0.41 - 0.6	Moderate (Sedang)
0.61 - 0.8	Substansial (Baik)
0.81 - 1	Almost Perfect (Hampir sempurna)

Tabel 2 menunjukkan nilai interpretasi Kappa yang menandakan apabila nilai Kappa semakin tinggi, maka tingkat kesepakatan dari kedua hasil opini yang dibandingkan memiliki interpretasi yang tinggi.

Tabel 3. Rata-rata hasil pengujian kappa *statistics*

Tabel Hasil Pengujian Kappa Statistics

No	Dokumen	Observer	Pengujian dengan Query		
			Query 2 kata	Query 3 kata	Query 5 kata
1	1	1	-0.083333333	0	0.25
2	1	2	-0.583333333	-0.5	-0.25
3	1	3	0.083333333	0.166666667	0.416666667
4	2	1	-0.260869565	-0.260869565	-0.13043478
5	2	2	-0.043478261	-0.043478261	0.086956522
6	2	3	0.043478261	0.043478261	0.173913043
7	3	1	-0.166666667	0.055555556	0.166666667
8	3	2	-0.111111111	0.111111111	0.222222222
9	3	3	-0.055555556	0.166666667	0.277777778
10	4	1	-0.311111111	-0.088888889	0.133333333
11	4	2	-0.266666667	-0.044444444	0.177777778
12	4	3	0.155555556	0.377777778	0.6
13	5	1	-0.382978723	-0.276595745	-0.14893617
14	5	2	-0.404255319	-0.29787234	-0.17021277
15	5	3	-0.340425532	-0.234042553	-0.10638298
16	6	1	-0.2	-0.066666667	0.111111111
17	6	2	0.222222222	0.355555556	0.533333333
18	6	3	0.088888889	0.222222222	0.4
Rata-rata			**-0.145350384**	**-0.017434703**	**0.152432875**
Iterpretasi Kappa			**Poor**	**Poor**	**Slight**

Hasil pengujian kappa *statistics* pada sebuah dokumen dengan *query* dan jumlah topik yang berbeda, hasil dapat dilihat pada `tabel 3`.

4. Kesimpulan

Berdasarkan hasil penelitian dan pengujian sistem pencarian kalimat opini, maka dapat diambil kesimpulan yaitu :

Implementasi metode *probabilistic latent semantic analisys* (PLSA) telah berhasil diterapkan dengan format dokumen *.doc, *.docx, dan *.pdf. Aplikasi sistem pencarian kalimat opini mampu menemukan kalimat yang relevansi dengan *query* yang diinginkan oleh *user* seperti hasil pengujian pada `tabel 3`. Hasil pengujian pada aplikasi pencarian kalimat opini ini sangat dipengaruhi oleh *query* yang dimasukkan oleh *user* seperti pada hasil pengujian `tabel 3`. Hasil pengujian kappa *statistic* menunjukkan bahwa semakin banyak kata *query* yang dimasukkan oleh *user*, maka hasil yang dihasilkan akan semakin baik.

Daftar Pustaka

[1] Ratri Anggardani Prayitno, Warih Maharani, Adhe Romadhony, 2012, *Opinion Retrieval Dengan Menggunakan Probabilistic Latent Semantic Analysis.* Program studi S1 Teknik Informatika (Telkom University) 2012.

[2] Darwin Suhartono, 2014, Probabilistic Latent Semantic Analysis (PLSA) untuk Klasifikasi Dokumen Teks Berbahasa Indonesia. Technical Report Program Studi Doktor Ilmu Komputer Fakultas Ilmu Komputer Universitas Indonesia, Desember 2014.

[3] Agusta, L., 2009, *Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief dan Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia*. Konferensi Nasional Sistem dan Informatika, KNS&109-036.

[4] Nazief, B. A. A. and Adriani, M. (1996) Confix-stripping: Approach to *stemming* algorithm For Bahasa Indonesia. Internal publication, Faculty of Computer Science, University of Indonesia, Depok, Jakarta.