BAB II

LANDASAN TEORI

2.1. Penelitian Terdahulu

Penelitian yang menggunakan metode *vector space model* telah dilakukan oleh beberapa peneliti diantaranya sebagai berikut :

1. Louisa K. Lake (2015) tentang *information retrieval system* dokumen teks abstrak skripsi dengan metode pembobotan tf-idf dan pengukuran similarity menggunakan *vector space model*. Dalam penelitian ini mengangkat masalah bagaimana membantu mahasiswa Teknik Informatika Universitas Katolik Widya Mandira Kupang dalam mencari dokumen skripsi yang diinginkan. Penelitian ini menghasilkan sebuah aplikasi *information retrieval system* yang mampu membantu mahasiswa dalam pencarian dokumen skripsi yang ada pada Program Studi Teknik Informatika berdasarkan *query* yang diinginkan.
2. Maarif (2015) tentang penerapan algoritma TF-IDF untuk pencarian karya ilmiah. Dalam penelitian ini mengangkat masalah bagaimana menerapkan algoritma TF-IDF yang dapat digunakan untuk mencari karya ilmiah. Penelitian ini menghasilkan sebuah aplikasi tentang karya ilmiah disertakan dengan hasil nilai bobot dari tiap karya ilmiah yang ditemukan.
3. DwijaWisnu dan Hetami (2015) tentang perancangan *information retrieval* (IR) untuk pencarian ide pokok teks artikel berbahasa inggris dengan pembobotan *vector space model*. Dalam penelitian ini mengangkat masalah bagaimana memanfaatkan *information retrieval* pada teks mining untuk menemukan ide pokok dalam teks pada artikel berbahasa inggris. Penelitian ini menghasilkan sebuah sistem pencarian ide pokok otomatis yang dapat membantu pembaca untuk lebih mudah dalam memahami isi artikel.
4. Mustain, Prasetyo, dan Rosyid (2015) tentang aplikasi pencarian jurnal skripsi menggunakan metode *vector space model* (model ruang vektor). Dalam penelitian ini mengangkat masalah bagaimana cara menerapkan

temu kembali informasi untuk menemukan abstrak naskah publikasi skripsi yang mirip *query* sesuai *input*an pengguna menggunakan metode *vector space model* (model ruang vektor). Penelitian ini menghasilkan *information retrieval system* dengan model ruang vektor yang berhasil diterapkan pada pencarian abstrak skripsi Program Studi Teknik Informatika UMG 2011-2012 dengan baik dan relatif dapat mempermudah dan mempercepat pencarian abstrak skripsi yang berbahasa indonesia.

Berdasarkan penelitian-penelitian sebelumnya tentang sistem temu kembali informasi, maka penelitian yang akan dibuat ini merujuk kepada penelitian yang dilakukan oleh Louisa K. Lake tentang *information retrieval system* dokumen teks abstrak skripsi dengan metode pembobotan tf-idf dan pengukuran similarity menggunakan *vector space model* karena pada penelitian ini menggunakan metode dan pembobotan yang sama. Pada penelitian Louisa K. Lake dibuat sebuah sistem temu kembali informasi yang mampu menampilkan dokumen skripsi yang ada pada Prodi Teknik Informatika. Dengan membandingkan penelitian Louisa K. Lake, maka pada penelitian ini akan dijelaskan secara rinci tentang metode vector space model dan pembobotan tf-idf dalam pencarian dokumen berdasarkan sebuah *query* yang kemudian akan divisualisasikan kedalam bentuk grafik.

2.2. *Information Retrieval*

Information retrieval adalah sebuah proses untuk menemukan kembali informasi yang dibutuhkan dari sebuah sistem penyimpanan dan penelusuran informasi. Sistem temu kembali informasi mensyaratkan ada kebutuhan informasi dari pengguna, ada dokumen atau data yang berisi informasi yang diorganisasikan dalam sebuah sistem yang memudahkan temu kembali informasi dan strategi penelusuran yang tepat sehingga dokumen yang sesuai dengan kebutuhan dapat ditemukan kembali. Sistem temu balik informasi terdiri dari tiga komponen utama, yaitu masukan (*input*), pemroses (*processor*), dan keluaran (*output*). Menurut Manning et al (2009) *Information Retrieval* (IR) adalah menemukan bahan (biasanya

dokumen) yang bersifat tidak terstruktur (biasanya teks) yang memenuhi kebutuhan informasi dari dalam koleksi besar (biasanya disimpan di komputer).

2.3. *Vector Space Model*

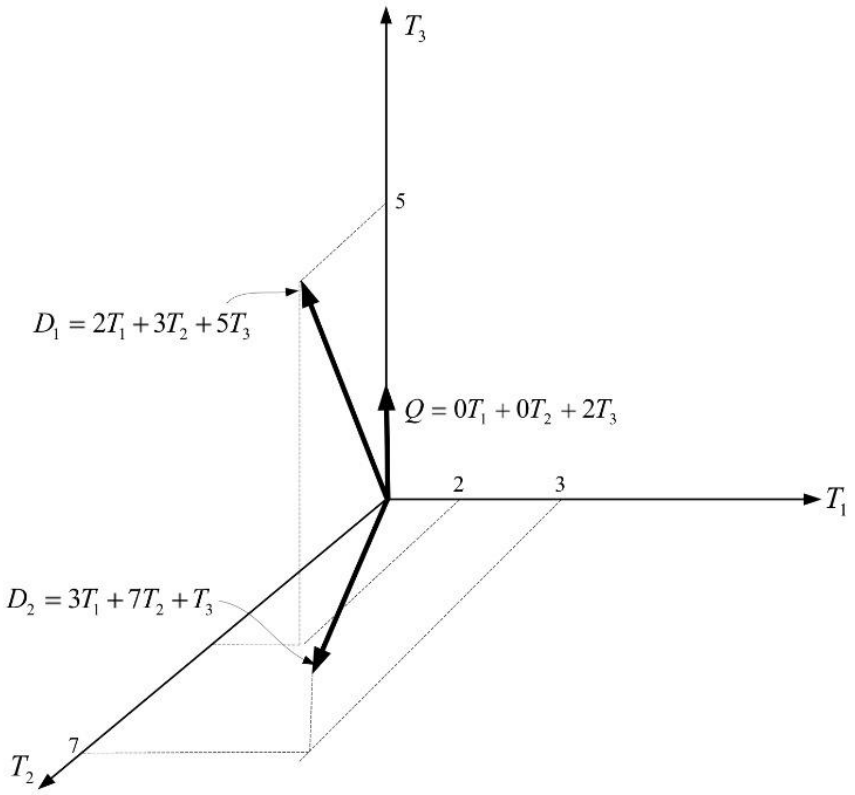
Vector Space Model merupakan model yang digunakan untuk mengukur kemiripan antara dokumen dan *query* yang mewakili setiap dokumen dalam sebuah koleksi sebagai sebuah titik dalam ruang [11]. Biasanya digunakan dalam penyaringan informasi (*information filtering*), penemuan informasi (*information retrieval*), *indexing* dan pemberian ranking yang saling relevan.

Misalkan terdapat jumlah n kata yang berbeda sebagai kamus kata (*vocabulary*) atau indeks kata (*terms index*). Kata-kata ini akan membentuk ruang vektor yang memiliki dimensi sebesar n. Setiap kata i dalam dokumen atau *query* diberikan bobot sebesar wi. Baik dokumen maupun *query* yang di representasikan sebagai vektor berdimensi n.

Sebagai contoh terdapat 3 buah kata (T1, T2 dan T3), 2 buah dokumen (D1 dan D2) serta sebuah *query* Q. Masing-masing bernilai:

$D1= 2T_1 + 3T_2 + 5T_3$; $D2 = 3T_1 + 7T_2$; $Q = 0T_1 + 0T_2 + 2T_3$

Maka representasi grafis dari ketiga vektor ini ditunjukkan pada gambar 2.1.



Gambar 2.1 Representasi Model Ruang Vektor

Model ruang vektor mengasumsikan bahwa sudah ada indeks *terms* yang telah merepresentasikan dokumen dan *query*. Sebuah dokumen D_i dan *query* Q_j direpresentasikan sebagai berikut :

$$D_i = [T_{i1}, T_{i2}, ..., T_{ik}, ..., T_{iN}]$$

$$Q_j = [Q_{j1}, Q_{j2}, ..., Q_{jk}, ..., Q_{jN}]$$

Dimana T_{ik} adalah nilai dari *term* k dalam dokumen i, Q_{jk} adalah nilai *query* k dalam *query* j, dan N adalah jumlah total istilah yang digunakan dalam dokumen dan *query*. Nilai dari T_{ik} dan Q_{jk} dapat diambil dari hasil pembobotan perhitungan TF-IDF.

Misalkan terdapat sekumpulan kata T sejumlah m, yaitu T = (T1, T2, …., Tm) dan sekumpulan dokumen D sejumlah n, yaitu D = (D1, D2, ….., Dn) serta w_{mn} adalah bobot kata m pada dokumen n.

Representasi matriks kata-dokumen digambarkan sebagai berikut :

$$\begin{array}{c} \\ T_1 \\ T_2 \\ \vdots \\ T_m \end{array} \begin{array}{c} \begin{array}{cccc} D_1 & D_2 & \dots & D_n \end{array} \\ \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{bmatrix} \end{array}$$

Gambar 2 Representasi nilai kata dalam dokumen

Untuk memperoleh kesamaan antara *query* dan dokumen dapat diperoleh dengan rumus sebagai berikut :

$$S(D_j, Q) = \frac{\overrightarrow{D_j} \cdot \vec{Q}}{|\overrightarrow{D_j}| \cdot |\vec{Q}|} = \frac{\sum_{i=1}^{N} w_{ij} \cdot w_{iq}}{\sum_{i=1}^{N} |w_{ij}| \cdot |w_{iq}|} \quad \dots\dots\dots\dots\dots (1)$$

Dimana

$S(D_j, Q)$ = nilai kesamaan antara sebuah dokumen i dengan *query*

w_{ij} = bobot term i pada dokumen j

w_{iq} = bobot term i pada *query*

Dengan rumusnya sebagai berikut :

$$|w_{ij}| = \sqrt{\sum_{i=1}^{N} {w_{ij}}^2} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

$$|w_{iq}| = \sqrt{\sum_{i=1}^{N} {w_{iq}}^2} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

Nilai cosinus yang cenderung besar menyatakan bahwa dokumen cenderung sesuai dengan *query* masukan. Sedangkan nilai cosinus yang cenderung kecil menyatakan bahwa dokumen cenderung tidak sesuai dengan *query* masukan.

2.4. *Indexing*

Indeks adalah bahasa yang digunakan di dalam sebuah buku konvensional untuk mencari informasi berdasarkan kata atau istilah yang mengacu ke dalam suatu halaman. Dengan menggunakan indeks si pencari informasi dapat dengan mudah menemukan informasi yang diinginkannya. Pada sistem temu-kembali informasi, indeks ini nantinya yang digunakan untuk merepresentasikan informasi di dalam sebuah dokumen.

Adapun tahapan dari pengindeksan adalah sebagai berikut :

1. *Pharsing*

   Pharsing merupakan proses pengubahan dokumen menjadi kumpulan *term* dengan cara menghapus semua karakter dalam tanda baca yang terdapat pada dokumen dan mengubah kumpulan term menjadi *lowercase*.

2. *Stopword Removal*

   Stopword Removal merupakan proses pembuangan kata buang seperti: tetapi, yaitu, sedangkan, dan sebagainya.

3. *Stemming*

   Stemming merupakan proses penghilangan/pemotongan dari suatu kata menjadi bentuk dasar.

4. *Term Weighting* dan *inverted file*

   Term Weighting dan *inverted file* merupakan proses pemberian bobot pada istilah.

2.5. Pembobotan tf-idf

Metode TF-IDF (*Term Frequency - Inverse Document Frequency*) merupakan suatu cara untuk memberikan bobot hubungan suatu kata (*term*) terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu, frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu dan invers frekuensi dokumen yang mengandung kata tersebut. Formula yang digunakan pada *Term Frequency* (tf) yaitu nilai tf diberikan berdasarkan jumlah kemunculan suatu kata di dokumen.

Misalkan jika muncul lima kali kata m, kata tersebut akan bernilai lima. *Inverse Document Frequency* (idf) dihitung dengan menggunakan formula sebagai berikut :

$$IDF_t = log\left(\frac{N}{df_t}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Dimana,

N = Jumlah total dokumen

df = Dokumen Frekuensi

t = *term*

Dengan demikian rumus umum untuk TF-IDF adalah penggabungan dari formula perhitungan tf dengan formula idf dengan cara mengalikan nilai *term frequency* (tf) dengan nilai *inverse document frequency* (idf).

$W_i = tf_i \; x \; IDF_i$ ..... (5)

$dimana$:

W_i = bobot kata indeks

tf_i = frekuensi term atau banyaknya term i yang ada pada sebuah dokumen

IDF_i = frekuensi dokumen atau banyak dokumen yang mengandung term i