BAB II

TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Document summarization adalah proses pengambilan teks dari sebuah dokumen dan membuat sebuah ringkasan yang mempunyai informasi yang lebih berguna bagi *user* (Radev *et al*, 2002). Secara umum terdapat dua model ringkasan dokumen yaitu model esktraktif dan model abstraktif (Long *et al*, 2010). Model berbasis ekstraktif mengambil kalimat yang sudah ada di dalam dokumen dan kemudian menggabungkannya untuk membentuk sebuah ringkasan (Turner *et al*, 2005), sehingga memungkinkan ringkasan untuk meningkatkan informasi secara keseluruhan tanpa menambah panjang ringkasan (Zajic *et al*, 2007).

Long *et al* (2010), membuat sebuah framework dimana *multi-dokumen summarization* dapat dimodelkan menggunakan teori informasi jarak. Ringkasan terbaik didefinisikan sebagai yang mempunyai informasi jarak yang minimal (kondisi informasi jarak) ke seluruh dokumen. Carbonell *et al* (1998) mengusulkan metode Maximal Marginal Relevance (MMR) yang digunakan untuk memilih kalimat ringkasan berdasarkan dari query pengguna dan memiliki kemiripan dengan yang dipilih sebelumnya. Radev *et al* (2004) menjelaskan sebuah sistem peringkas dokumen dengan model ekstraktif yang mengambil kalimat dari banyak dokumen sesuai dengan *cluster centroids* dari dokumen.

Dengan menggabungkan informasi yang sama dari sebuah dokumen, *multi-document summarization* dapat membantu pengguna dalam memperoleh informasi

yang relevan dengan membaca informasi yang minimal menggunakan sebuah sistem pencari informasi. Perangkat lunak peringkas dokumen dari banyak sumber ini dapat digunakan untuk membuat ringkasan yang bersumber dari berbagai dokumen yang terdapat di media online. Dokumen yang memiliki topik yang sama akan dibuat menjadi sebuah ringkasan yang memiliki informasi yang lebih relevan.

Perangkat lunak ini menggunakan metode TF-IDF untuk membuat ringkasan dari banyak sumber dokumen. TF-IDF (*Term Frequency-Inverse Document Frequency)* merupakan salah satu metode yang biasa dipakai dalam pembobotan sebuah kata di dalam sistem pencarian informasi (Aizawa, 2002). Metode TF-IDF menghitung nilai dari masing-masing kata di dalam dokumen menggunakan frekuensi kata tersebut muncul. Kata dengan nilai TF-IDF yang tinggi berarti bahwa kata tersebut mempunyai hubungan yang kuat dengan dokumen dimana kata tersebut muncul, diasumsikan bahwa jika kata tersebut muncul di dalam query maka akan memiliki ketertarikan bagi pengguna (Ramos, 2004). TF-IDF juga dapat digunakan dalam pembobotan kata untuk mencari keputusan yang relevan (Wu *et al*, 2008).

Dalam tesis ini metode TF-IDF digunakan untuk melakukan pembobotan kata di dalam sebuah dokumen dan kemudian dilakukan *sentence scoring* menggunakan nilai dari kata di dalam masing-masing kalimat dan mengambil beberapa kalimat dengan nilai tertinggi untuk kemudian digabungkan menjadi sebuah ringkasan. Dokumen yang akan diringkas didapat dari media online yang tersebar banyak di dunia maya.

Setelah didapat ringkasan dari masing-masing dokumen, ringkasan tersebut kemudian digabungkan dan dilakukan *sentence scoring* menggunakan metode TF-IDF. Beberapa kalimat yang memiliki nilai tertinggi akan dipilih dan kemudian digunakan sebagai ringkasan akhir.

Perbandingan perangkat lunak peringkas dokumen dari banyak sumber berbasis web menggunakan *sentence scoring* dengan metode TF-IDF dengan aplikasi – aplikasi yang lain dapat dilihat pada tabel 2.1.

Tabel 2.1 Tinjauan Pustaka.

Penulis	Judul	Metode	Pokok Bahasan
(Radev *et al*, 2004)	*Centroid-based summarization of multiple documents. Information Processing and Management*	*Centroid-based*	1. Sistem peringkas dokumen dengan model ekstraktif yang mengambil kalimat dari banyak dokumen sesuai dengan *cluster centroids* dari dokumen.
(Carbonell *et al*, 1998)	*The use of MMR, diversity-based reranking for reordering documents and producing summaries*	*MMR*	1. Menggunakan metode Maximal Marginal Relevance (MMR) untuk memilih kalimat ringkasan berdasarkan dari query pengguna dan memiliki kemiripan dengan yang di pilih sebelumnya.
(Long *et al*, 2010)	*A New Approach for Multi-Document Update Summarization*	*Information distance*	1. *Multi-dokumen summarization* dapat di modelkan menggunakan teori informasi jarak.
(Evan, 2014)	*Pembangunan Perangkat Lunak Peringkas Dokumen dari Banyak Sumber Berbasis Web Menggunakan Sentence Scoring dengan Metode TF-IDF*	*TF-IDF*	1. Sistem peringkas dokumen dengan model ekstraktif yang mengambil kalimat dari banyak dokumen sesuai dengan *sentence score* menggunakan metode TF-IDF dari dokumen. 2. Berbasis Web

2.2 Landasan Teori

2.2.1 *Document Summarization*

Document summarization adalah proses pengambilan teks dari sebuah dokumen dan membuat sebuah ringkasan yang mempunyai informasi yang lebih berguna bagi *user* (Radev *et al*, 2002). Perbedaan antara *single-document summarization* dan *multi-document summarization* adalah, *multi-document summarization* lebih kompleks daripada *single-document summarization*.

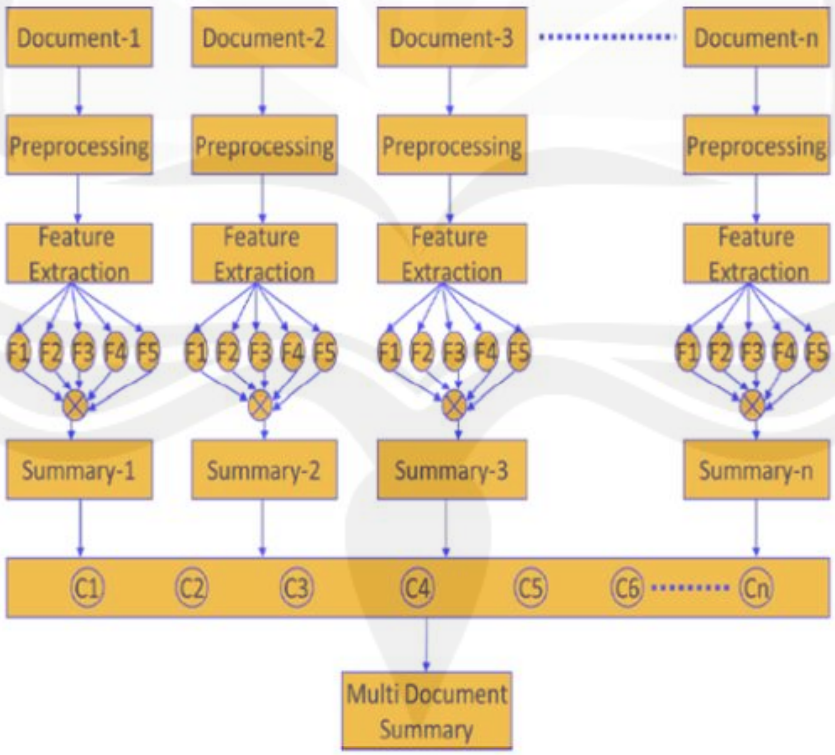
Ringkasan *Multi-Document* merupakan proses penyaringan informasi penting dari beberapa dokumen untuk menghasilkan ringkasan yang singkat untuk pengguna dan aplikasi. Ringkasan *Multi-Document* dapat dilihat sebagai perpanjangan dari ringkasan *Single-Document* (Gupta *et al*, 2012). Faktor yang membuat *multi-document summarization* lebih rumit antara lain :

a. Setiap artikel yang di gunakan ditulis oleh penulis yang berbeda yang mempunyai perbedaan gaya menulis dan struktur dokumen.
b. Beberapa artikel mungkin mempunyai pandangan yang berbeda tentang suatu topik yang sama.
c. Fakta dan sudut pandang dapat berubah kapanpun setiap dokumen dibuat sehingga dapat menimbulkan konflik dalam penyediaan informasi.

Ringkasan biasanya dihasilkan menggunakan dua teknik yang biasa disebut dengan ekstraksi dan abstraksi (Jing 2000, Knight 2002, Mani 2001). Ringkasan ekstraktif hanya mengekstrak informasi penting, seperti kalimat, dari dokumen input dan "menempatkan mereka bersama-sama" untuk membentuk sebuah ringkasan. Meskipun ringkasan yang dihasilkan dengan cara ini mungkin kurang

koherensi, pendekatan ekstraktif masih populer saat ini karena murah dan mudah untuk diterapkan ke domain umum.

Bertujuan untuk menghasilkan ringkasan koheren gramatikal, ringkasan abstraktif menciptakan ringkasan dengan cara sintesis dan menulis ulang kalimat berdasarkan pemahaman kontekstual dan linguistik, yang sangat tergantung pada analisis yang mendalam dan teknik bahasa generasi. Ratherthan hanya menggunakan teknik tunggal saja, beberapa peneliti tertarik regenerasi sebagai pasca-proses untuk menghasilkan ringkasan ekstraktif, yaitu, membuat pemangkasan atau revisi berdasarkan ringkasan ekstraktif (Mani 2001, Otterbacher 2002).



Gambar 2.2.1 ***Multi Document Summarization***

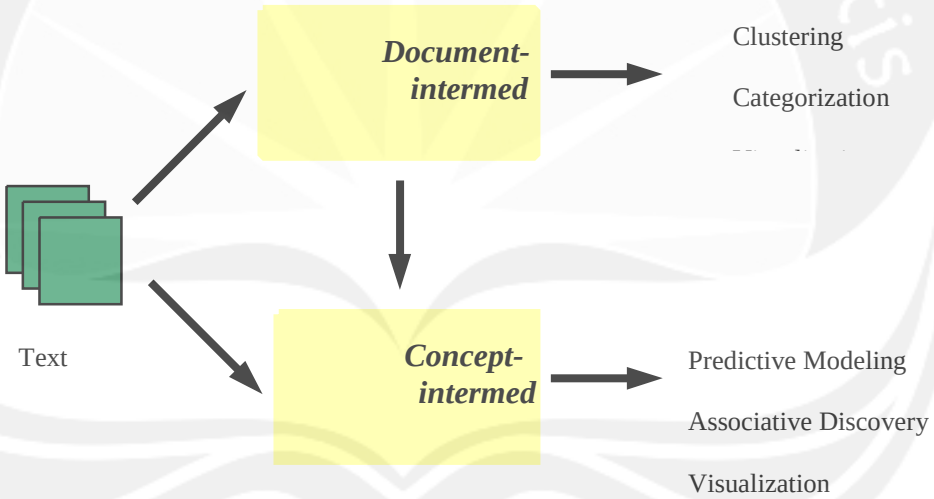
2.2.2 *Text Mining*

Text mining, juga dikenal sebagai *teks data mining* (Hearst, 1997) atau penemuan pengetahuan dari database tekstual (Feldman, 1995), umumnya mengacu pada proses ekstraksi pola atau pengetahuan yang menarik dari dokumen teks yang tidak terstruktur. Hal ini dapat dilihat sebagai perpanjangan dari *data mining* atau *knowledge discovery* dari database (Simoudis,1996).

Sebagai bentuk paling alami dalam menyimpan sebuah informasi adalah berupa teks. *Text mining* diyakini memiliki potensi komersial lebih tinggi dari *data mining*. Bahkan, penelitian terbaru menunjukkan bahwa 80% dari informasi perusahaan terkandung dalam sebuah dokumen teks. *Text mining*, bagaimanapun, juga merupakan proses yang lebih kompleks (dibanding *data mining*) karena melibatkan data teks yang secara inheren tidak teratur dan kabur. *Text mining* adalah bidang multidisiplin, yang melibatkan pencarian informasi, analisis teks, ekstraksi informasi, *clustering*, kategorisasi, visualisasi, basis data teknologi, pembelajaran mesin, dan *data mining* (Tan, 1999).

Text mining dapat divisualisasikan mempunyai dua tahap: Teks penyulingan yang mengubah dokumen teks dalam bentuk-bebas ke dalam bentuk peralihan yang dipilih, dan pengetahuan distilasi yang menyimpulkan pola atau pengetahuan dari bentuk peralihan. Bentuk peralihan (IF) bisa semi-terstruktur seperti representasi grafik konseptual, atau terstruktur seperti representasi data relasional. Bentuk peralihan dapat berupa dokumen dimana setiap entitas merupakan sebuah dokumen, atau berupa konsep dimana setiap entitas merupakan objek atau konsep yang memiliki kepentingan dalam domain tertentu. Pertambangan berdasar

dokumen dengan cara bentuk peralihan menyimpulkan pola dan hubungan antar dokumen. Dokumen pengelompokan / visualisasi dan kategorisasi adalah contoh dari pertambangan dari bentuk peralihan berdasar dokumen. Pertambangan berdasar konsep dengan cara bentuk peralihan memiliki pola dan hubungan di seluruh objek atau konsep. Operasi *data mining*, seperti model prediksi dan penemuan asosiatif, termasuk dalam kategori ini. Sebuah bentuk peralihan berdasar dokumen dapat diubah menjadi sebuah bentuk peralihan berdasar konsep dengan meluruskan kembali atau mengekstraksi informasi yang relevan sesuai dengan objek kepentingan dalam domain tertentu.



Gambar 2.2.2 *Text Minning*

2.2.3 TF-IDF (*Term Frequency – Inverse Document Frequency*)

TF-IDF digunakan untuk menentukan nilai frekuensi sebuah kata di dalam sebuah dokumen atau artikel dan juga frekuensi di dalam banyak dokumen. Perhitungan ini menentukan seberapa relevan sebuah kata di dalam sebuah dokumen.

Penggunaan *term frequency* dan *document frequency* untuk memberikan peringkat pada dokumen dipelajari secara ekstensif, khususnya oleh Salton *et al*., untuk model ruang vektor (Salton *et al*, 1988). Mengikuti pertimbangan dari model diskriminasi kata (Salton *et al*, 1973), mereka berpendapat bahwa kata yang muncul dalam dokumen harus diberi nilai sebanding dengan frekuensi kata dan terbalik sebanding dengan frekuensi dokumen. Pembobotan skema yang mengikuti pendekatan ini disebut *TF* × *IDF* (*term frequency* × *inverse document frequency*).

Prosedur dalam implementasi TF-IDF terdapat perbedaan kecil di dalam semua aplikasinya, tetapi pendekatannya kurang lebih sama. TF (*Term Frequency)* merupakan frekuensi sebuah kata terdapat di dalam sebuah dokumen. Nilai dari TF didapat menggunakan persamaan:

$$TF(t) = f_{t,d} / \sum t, d \quad (1)$$

dimana $f_{t,d}$ merupakan frekuensi sebuah kata (*t*) muncul di dalam dokumen *d*, sedangkan $\sum t, d$ merupakan total keseluruhan kata yang terdapat di dalam dokumen *d* (Aizawa, 2002).

Kemudian untuk menghitung nilai IDF (*Inverse Document Frequency*) dari sebuah kata di dalam kumpulan dokumen menggunakan persamaan:

$$IDF(t) = \log(|D| / f_{t,D}) \quad (2)$$

|*D*| merupakan jumlah dokumen yang ada di dalam koleksi, sedangkan $f_{t,D}$ merupakan jumlah dokumen dimana *t* muncul di dalam *D* (Salton & Buckley, 1988,

Berger, *et al*, 2000). Dalam koleksi dokumen *D*, sebuah kata *t* dan dokumen individu *d* ϵ *D*, dapat dihitung nilai TF-IDF menggunakan rumus:

$$TF - IDF(t) = TF(t) * IDF(t) \quad (3)$$

Diasumsikan bahwa $|D| \sim f_{t,D}$, ukuran dari kumpulan dokumen hampir sama dengan frekuensi *t* di dalam *D*. Jika $1 < \log(|D|/f_{t,D}) < c$ untuk sebuah konstanta c dengan nilai yang kecil, maka w_d akan lebih kecil daripada $f_{t,d}$ tetapi tetap bernilai positif. Hal berarti bahwa w mempunyai relasi yang biasa dengan seluruh dokumen tetapi tetap menyimpan beberapa informasi yang penting (Ramos, 2000).

Sebagai contoh adalah ketika metode TF-IDF menemukan kata android di dalam kumpulan dokumen yang berisi artikel tentang android. Hal ini juga berlaku untuk beberapa kata seperti kata sambung yang tidak mempunyai arti yang releven terhadap sebuah dokumen. Kata sambung tersebut diberi nilai yang rendah menggunakan TF-IDF.

Jika $f_{t,d}$ bernilai besar dan $f_{t,D}$ bernilai kecil, maka nilai dari $\log(|D|/f_{t,D})$ cenderung tinggi dan juga TF-IDF(*t*) akan bernilai tinggi. Kata dengan nilai TF-IDF(*t*) yang tinggi berarti bahwa kata tersebut merupakan sebuah kata yang sangat penting di dalam dokumen *d* tetapi tidak di dalam kumpulan dokumen *D*.