

Bab I

Pendahuluan

Bab ini memaparkan latar belakang yang melandasi penyusunan Tugas Akhir ini, diikuti dengan tujuan serta perumusan dan pembatasan masalah. Kemudian ditutup dengan metodologi yang digunakan dalam penyusunan Tugas Akhir serta sistematika pembahasan dalam laporan Tugas Akhir ini.

I.1 Latar Belakang

Sejak internet mulai marak digunakan sebagai sarana untuk distribusi informasi, banyak lembaga media yang memanfaatkannya untuk mempublikasikan *content*. Sebagai contoh, para penerbit surat kabar kini juga menyajikan beritanya dalam versi *online*—di samping dalam bentuk cetak konvensional yang telah kita kenal selama ini. Hal ini tentunya memudahkan seseorang untuk memperoleh berita dengan cepat dan mudah; cukup dengan memanfaatkan aneka peranti yang terhubung ke internet.

Setiap lembaga berita tentunya berusaha untuk menyajikan berita yang lengkap, adil, dan berimbang. Namun di sisi lain, kualitas *content* berita yang dipublikasikan oleh masing-masing lembaga dapat sangat beragam, tergantung pada sudut pandang, gaya bahasa, pilihan kata, maupun segmen pembaca. Tidak dapat disangkal, mungkin saja dapat terjadi penyimpangan maupun bias dalam penyajian berita akibat berbagai faktor, misalnya tarikan politik, wartawan yang kurang kompeten, maupun data yang kurang lengkap. Tidak mengherankan apabila seseorang seringkali mengakses berita dari berbagai sumber untuk dapat memperoleh gambaran menyeluruh tentang suatu peristiwa.

Teknologi internet, khususnya *World Wide Web*, memang telah mempermudah pengguna untuk melakukan hal tersebut. Namun tetap saja diperlukan usaha tertentu untuk mendapatkan berita dari berbagai sumber sekaligus. Tentunya akan lebih baik apabila seseorang dapat mengetahui semua berita dari berbagai sumber secara langsung.

Salah satu aplikasi yang umum digunakan untuk keperluan tersebut adalah agregator berita (*news aggregator*). Agregator berita adalah sebuah sistem atau aplikasi yang mengumpulkan berita dari berbagai sumber, kemudian menyuguhkannya dalam satu sajian. Sayangnya, agregator berita umumnya menyajikan berita dalam satu kumpulan besar dengan menggabungkan semua berita yang telah dikumpulkannya. Hal ini dapat menyulitkan pengguna untuk mengikuti perkembangan peristiwa yang terjadi. Selain

itu, hal ini juga dapat menimbulkan rasa kelebihan informasi (*information overload*) pada pengguna.

Clustering adalah metode pengelompokan objek secara otomatis menjadi sejumlah kelompok (*cluster*) yang bermakna. Tujuan dari proses *clustering* adalah menemukan pola pengelompokan yang alami dari data. Oleh karena itu, *clustering* sangat tepat jika diterapkan pada agregator berita untuk mengelompokkan berita dari berbagai sumber yang membahas suatu kejadian yang sama.

Terdapat banyak algoritma *clustering* yang telah dikembangkan untuk berbagai keperluan. Tentunya setiap algoritma tersebut memiliki kelebihan dan kekurangan masing-masing. Beberapa algoritma *clustering* bersifat generik, namun sebagian yang lain lebih cocok untuk digunakan pada domain atau keperluan tertentu, misalnya untuk bioinformatika, data spasial, maupun analisis ekonomi.

Berita yang diterbitkan oleh lembaga media *online* umumnya berupa teks polos (*plaintext*) yang disajikan dalam sebuah halaman web. Berita tersebut adalah data yang berupa dokumen. Analisis *cluster* dapat diterapkan pada dokumen selama terdapat representasi yang sesuai—lazim disebut *document clustering*. Salah satu bentuk representasi yang dapat digunakan adalah model ruang vektor [SWY75].

I.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, masalah utama yang akan dikaji pada Tugas Akhir ini adalah penentuan algoritma *clustering* yang sesuai untuk diterapkan pada dokumen berupa berita. Algoritma *clustering* tersebut setidaknya harus memiliki karakteristik sebagai berikut:

1. Hasil *clustering* intuitif

Proses *clustering* harus mampu “menangkap” struktur alami pada data sesuai dengan intuisi manusia. Diharapkan setiap kelompok berita dapat mencerminkan sebuah peristiwa. Misalnya jika sebuah *cluster C* membahas tentang suatu kejadian, maka semua berita yang berisi tentang kejadian itu harus berada dalam *cluster C* dan semua berita yang tidak terkait dengan kejadian tersebut seharusnya tidak berada dalam *cluster C*.

2. Dapat dilakukan secara *incremental*

Setiap lembaga berita akan mempublikasikan berita secara terus menerus seiring peristiwa yang terjadi. Oleh karena itu, berita dapat dianggap sebagai arus informasi yang selalu mengalir. Diharapkan berita yang baru datang dapat dikelompokkan tanpa harus melakukan *clustering* ulang terhadap semua berita yang sudah diproses sebelumnya.

3. Cepat, efisien, dan dapat menangani jumlah data yang besar

Algoritma harus dapat dijalankan dalam suatu batasan waktu yang dapat diterima (*reasonable*). Ukuran koleksi dokumen berupa berita yang akan diproses tentunya berjumlah sangat besar. Oleh karena itu, algoritma harus *scalable* terhadap banyaknya berita.

4. Dapat menangani *outlier*
Faktanya, tidak semua berita memiliki kesamaan bahasan dengan berita lainnya. Oleh karena itu, berita semacam ini dapat dianggap sebagai *outlier* dan tidak digabungkan dengan kelompok lain.
5. Dapat melakukan perbaikan berdasarkan umpan balik (*feedback*)
Pada dasarnya, *clustering* adalah proses pembelajaran yang bersifat tak terbimbing (*unsupervised learning*). Oleh karena itu, ada kemungkinan hasil *clustering* tidak sesuai dengan penilaian oleh manusia secara subjektif. Diharapkan algoritma dapat menerima masukan berupa umpan balik (*feedback*) dan berdasarkan umpan balik tersebut kemudian melakukan perbaikan pada kualitas *cluster* [WCRS01]. Dengan kata lain, algoritma menjadi bersifat *semi-supervised*.

I.3 Tujuan

Tujuan yang ingin dicapai dalam pelaksanaan Tugas Akhir ini adalah:

1. Memahami algoritma *clustering* yang sesuai untuk diterapkan pada dokumen, terutama dokumen yang berupa berita, sesuai dengan masalah-masalah yang telah dirumuskan sebelumnya pada bagian I.2.
2. Membangun agregator berita berbasis web dengan fungsionalitas utama mengumpulkan berita dari berbagai situs web berita serta menyajikannya dalam kelompok-kelompok berita yang memiliki topik yang sama dengan memanfaatkan *document clustering*.

I.4 Batasan Masalah

Batasan masalah pada pelaksanaan Tugas Akhir ini adalah:

1. Agregator berita yang akan dibangun dibatasi berupa prototipe.
2. Situs web yang akan digunakan sebagai sumber berita dibatasi hanya situs berita yang berbahasa Indonesia.

I.5 Metodologi

Tahapan yang akan dikerjakan selama pelaksanaan Tugas Akhir ini meliputi:

1. Studi literatur
Kajian terhadap literatur akan dilakukan pada seluruh proses pengerjaan Tugas Akhir. Studi literatur difokuskan pada berbagai algoritma *clustering* serta eksplorasi agregator berita yang sudah ada.
2. Eksplorasi algoritma *clustering*
Eksplorasi dilakukan dengan melakukan pengujian kinerja algoritma menggunakan data sampel, kemudian dipilih algoritma yang memiliki hasil yang terbaik.

3. Implementasi komponen *clustering engine*
Algoritma *clustering* yang terpilih akan diimplementasikan sebagai sebuah *clustering engine*.
4. Perancangan dan implementasi komponen pendukung
Meliputi komponen pendukung dari sistem agregator berita. Deskripsi lebih lanjut dari komponen-komponen tersebut dapat dilihat pada bagian VI.2.
5. Pengujian
Pengujian keseluruhan sistem meliputi semua komponen.
6. Perbaikan
Perbaikan dilakukan terhadap kesalahan-kesalahan yang mungkin terjadi pada perangkat lunak, laporan, maupun dokumentasi teknis.

I.6 Sistematika Pembahasan

Sistematika penulisan laporan Tugas Akhir ini adalah sebagai berikut:

Bab I Pendahuluan berisi penjelasan mengenai latar belakang pelaksanaan Tugas Akhir ini, perumusan masalah, tujuan, batasan masalah, metodologi pelaksanaan, serta sistematika pembahasan yang digunakan untuk menyusun laporan Tugas Akhir.

Bab II Analisis *Cluster* dan Representasi Dokumen berisi beberapa konsep dan teori yang mendasari Tugas Akhir ini, dengan menitikberatkan pada pembahasan mengenai analisis *cluster* dan representasi dokumen yang digunakan untuk keperluan analisis tersebut.

Bab III *Clustering* pada Aliran Informasi membahas algoritma *clustering* yang terkait dengan Tugas Akhir ini. Selain itu, pada bab ini dipaparkan pula kajian pengembangan algoritma *clustering* dengan pembatasan (*constrained clustering*).

Bab IV Analisis memaparkan analisis terhadap permasalahan utama Tugas Akhir ini, yaitu analisis *cluster* pada aliran berita yang berasal dari berbagai situs web layanan berita.

Bab V Eksperimen untuk Evaluasi Algoritma menjelaskan tahap eksperimen yang dilakukan dalam memilih algoritma *clustering* yang terbaik dan analisis atas hasil eksperimen tersebut.

Bab VI Perancangan & Implementasi menjelaskan proses perancangan dan implementasi agregator berita.

Bab VII Penutup menyimpulkan hal-hal yang diperoleh selama pelaksanaan Tugas Akhir ini serta saran untuk meningkatkan kualitas dan kegunaannya.