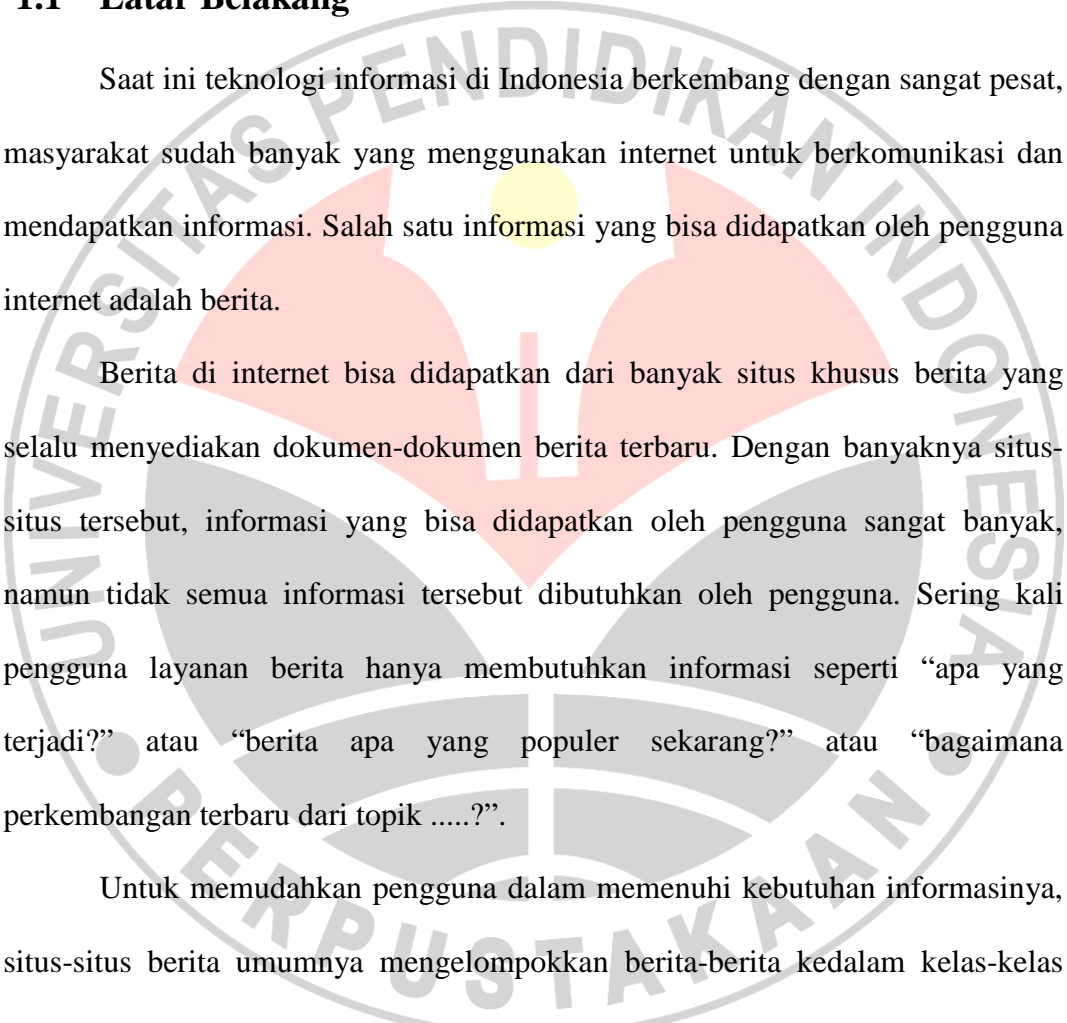


BAB I

PENDAHULUAN

1.1 Latar Belakang

Saat ini teknologi informasi di Indonesia berkembang dengan sangat pesat, masyarakat sudah banyak yang menggunakan internet untuk berkomunikasi dan mendapatkan informasi. Salah satu informasi yang bisa didapatkan oleh pengguna internet adalah berita.

Berita di internet bisa didapatkan dari banyak situs khusus berita yang selalu menyediakan dokumen-dokumen berita terbaru. Dengan banyaknya situs-situs tersebut, informasi yang bisa didapatkan oleh pengguna sangat banyak, namun tidak semua informasi tersebut dibutuhkan oleh pengguna. Sering kali pengguna layanan berita hanya membutuhkan informasi seperti “apa yang terjadi?” atau “berita apa yang populer sekarang?” atau “bagaimana perkembangan terbaru dari topik?”.


Untuk memudahkan pengguna dalam memenuhi kebutuhan informasinya, situs-situs berita umumnya mengelompokkan berita-berita kedalam kelas-kelas umum seperti regional, bisnis, internasional, kesehatan dan kelas-kelas umum lainnya. Namun jumlah dokumen berita pada tiap kelas-kelas tersebut masih sangat banyak dan berisi berita dengan berbagai topik yang lebih spesifik.

Pengelompokan dokumen-dokumen berita kedalam topik-topik yang lebih spesifik membutuhkan waktu dan sumber daya manusia yang tidak sedikit,

sebagai hasilnya hanya kejadian-kejadian besar seperti “Pemilihan Presiden 2009” atau “Piala Dunia 2010” yang dipilih oleh redaksi berita untuk diciptakan kelompok tersendiri. Untuk mengatasi permasalahan waktu dan sumber daya manusia tersebut, dibutuhkan pengelompokan dokumen-dokumen berita berdasarkan topik-topiknya secara otomatis, selain itu, untuk lebih memudahkan pengguna dalam menemukan informasi yang diinginkannya, dibutuhkan juga pemberian deskripsi dari kelompok berita tersebut.

Penelitian ini bertujuan untuk mengatasi permasalahan otomatisasi tersebut dengan mengaplikasikan algoritma *clustering* dalam pengelompokan dokumen-dokumen berita dengan topik yang sama, dilanjutkan dengan memproses kelompok-kelompok dokumen hasil *clustering* dengan metode *cluster labeling* untuk memberikan deskripsi pada tiap kelompok dokumen.

Algoritma *clustering* yang digunakan pada penelitian ini adalah algoritma *clustering* dengan metode *Group Average Hierarchical Agglomerative Clustering* (GA-HAC). Algoritma *clustering* GA-HAC dipilih karena sifatnya yang dapat memberikan hasil *clustering* dengan jumlah *cluster* yang tidak diketahui (Smet dan Moens, 2009, h.5).

Algoritma *cluster labeling* yang digunakan pada penelitian ini adalah algoritma *Mutual Information* (MI). Algoritma MI dipilih karena sifatnya yang memilih *cluster* label dengan membandingkan distribusi *term* pada suatu *cluster* dan membandingkannya dengan distribusi *term* pada *cluster* lain sehingga label yang dipilih tidak berulang dan bukan merupakan komplemen dari *cluster* lain (Manning, dkk, 2008, h.272, h. 396).

1.2 Rumusan Masalah

Rumusan masalah pada penelitian ini adalah bagaimana mengelompokkan dokumen-dokumen berita berbahasa Indonesia kedalam topik yang sama secara otomatis dengan algoritma GA-HAC, merepresentasikan kelompok tersebut agar dapat dimengerti pembaca dengan memilih frasa representasi menggunakan algoritma MI, membahas detail implementasi serta mengkaji kinerja implementasi kedua algoritma tersebut.

1.3 Batasan Masalah

Untuk mempertajam fokus penelitian, ditetapkan batasan-batasan masalah sebagai berikut:

1. Sumber data yang digunakan pada penelitian ini adalah aliran berita yang didapatkan selama 48 jam dari tanggal 25-04-2011 jam 14:00 sampai dengan tanggal 27-04-2011 jam 13:00 dari situs-situs berita:
 - <http://nasional.kompas.com/>
 - <http://nasional.vivanews.com>
 - <http://www.antaraneews.com/berita>
 - <http://www.republika.co.id/berita/nasional>
2. Dokumen-dokumen berita yang digunakan datanya telah dibersihkan dari artefak-artefak seperti navigasi, iklan, tautan berita lain, komentar, tag html dan lain-lain.
3. Bagian dokumen berita yang digunakan hanyalah judul dan isi berita saja, informasi-informasi lain seperti lokasi, keterkaitan berita berdasarkan

tautan-tautan artikel yang berhubungan, tanggal dan waktu berita tidak digunakan.

4. Data berita diproses secara *Batch*, yaitu pemrosesan dilakukan dengan terlebih dahulu mengumpulkan dokumen berita selama periode waktu 24 jam dengan metode *time window*.

1.4 Tujuan Penelitian

Tujuan dilaksanakannya penelitian ini adalah:

1. Meneliti sistem *clustering* dengan algoritma *Group Average Hierarchical Agglomerative Clustering* (GA-HAC).
2. Meneliti sistem *cluster labeling* dengan algoritma *Mutual Information*(MI)
3. Meneliti kinerja implementasi algoritma GA-HAC dan MI pada dokumen-dokumen berita berbahasa Indonesia.

1.5 Manfaat Penelitian

Diharapkan adanya penelitian ini, dapat memberikan manfaat sebagai berikut:

1. Mempersingkat waktu pencarian berita dengan topik yang berhubungan.
2. Memberikan kakas yang memudahkan penemuan berita-berita dengan topik yang sama untuk memenuhi kebutuhan informasi berita secara lebih singkat.
3. Dapat menjadi rujukan bagi penelitian selanjutnya yang memiliki keterkaitan dengan penelitian ini baik secara langsung maupun tidak langsung.

1.6 Sistematika Penulisan

Dalam penyusunan skripsi ini, sistematika penulisan dibagi menjadi lima bab dengan rincian sebagai berikut:

BAB I PENDAHULUAN

Bab ini meliputi pembahasan masalah secara umum meliputi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini memuat landasan teori yang berfungsi sebagai sumber atau alat dalam memahami permasalahan yang berkaitan dengan teori dan konsep pengelompokan dan perepresentasian topik dokumen-dokumen berita.

BAB III METODOLOGI PENELITIAN

Bab ini merupakan penjabaran dari desain penelitian, alat dan bahan yang digunakan pada penelitian, metode pengembangan perangkat lunak dan prosedur pengerjaan penelitian.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Bab ini membahas implementasi teknik yang digunakan untuk solusi dari rumusan masalah yang ditentukan. Penjabaran hasil implementasi teknik tersebut kemudian dibahas pada bagian evaluasi dan analisa hasil penelitian.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan yang dapat diambil dari penelitian dan saran yang dapat dijadikan pertimbangan dalam penelitian selanjutnya.