BAB V

PENUTUP

1.1 Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan beberapa hal berikut:

1. Algoritma GA-HAC dengan TF-IDF, *Vector Space Model*, *Cosine Similarity*, penggunaan 20% *stopwords* dan *Natural Clustering* dapat menghasilkan sekumpulan *cluster* yang berisi dokumen dengan topik berita yang sama dengan rata-rata nilai evaluasi F_1 = 0,691 dengan *precision* = 0,844 dan *recall* 0,629. Ini mengindikasikan bahwa hasil *clustering* yang dihasilkan lebih memenuhi kebutuhan informasi *web surfer* umum daripada kebutuhan informasi *intelegent analyst.*
2. Pemilihan frasa-frasa calon *cluster label* dengan frasa kata benda hasil *Part-of-Speech Tagger* dan algoritma *Mutual Information* yang dikombinasikan dengan *Laplace Correction* dapat digunakan untuk menghasilkan *cluster label* yang merepresentasikan dokumen-dokumen berita dari tiap *cluster* dan bekerja paling baik saat jumlah dokumen berita yang terkandung didalamnya cukup banyak.
3. Jika topik-topik terpopuler ditentukan dengan jumlah dokumen yang terkandung pada tiap *cluster*, maka kombinasi algoritma-algoritma yang digunakan pada penelitian ini dapat digunakan untuk mendefinisikan topik-topik berita terpopuler saat itu, selain itu, dengan digunakannya *time*

window dan segmentasi koleksi dokumen, kebutuhan dalam pendefinisian topik-topik terpopuler dapat dilakukan secara dinamis baik untuk per-hari, per-minggu, per-bulan dan lain-lain.

1.2 Saran

Sistem yang dikembangkan oleh penulis dalam skripsi masih belum sempurna dan dapat dikembangkan dengan penelitian-penelitian lanjutan. Adapun beberapa hal yang dapat penulis sarankan untuk penelitian lanjutan mengenai hal ini adalah:

1. Karena jumlah dokumen berpengaruh dalam menentukan hasil *cluster labeling*, maka disarankan untuk menggunakan jumlah dokumen yang lebih banyak pada penelitian lanjutan mengenai *cluster labeling*.
2. Penentuan frasa-frasa calon *cluster label* pada penelitian ini dilakukan secara sederhana, yaitu dengan mengikut sertakan *token* yang memiliki kelas kata NN atau NNP dari hasil *Part-Of-Speech Tagger*. Penentuan frasa-frasa calon *cluster label* yang sederhana ini dapat menyebabkan frasa-frasa kata benda yang bukan merupakan frasa penting disertakan sebagai calon *cluster label*. Salah satu pengembangan yang dapat dilakukan untuk menentukan frasa-frasa calon *cluster label* adalah dengan menggunakan algoritma *Named Entity Recognizer*.
3. Penggunaan frasa-frasa untuk mewakili suatu topik kadang tidak cukup untuk menjelaskan isi dokumen-dokumen berita pada suatu *cluster* kepada pembaca berita, dibutuhkan deskripsi lebih mendetail berupa sebuah kalimat singkat sehingga penjelasan isi dokumen-dokumen berita dapat

dilakukan dengan lebih baik. Salah satu solusi yang dapat dikembangkan untuk permasalahan ini adalah dengan mengaplikasikan algoritma *document summarizer* pada tiap *cluster* dokumen.

4. Penggunaan *time window* yang tidak menyertakan dokumen-dokumen yang melebihi batas waktu kadaluarsa menyebabkan tidak disertakannya dokumen-dokumen tersebut dalam hasil *clustering*. Ini menyebabkan hilangnya informasi masa lampau mengenai topik tersebut sehingga penelusuran topik pada dokumen masa lampau tidak dapat dilakukan. Salah satu solusi yang dapat dikembangkan untuk mengatasi permasalahan ini adalah dengan menggunakan pembobotan *term* yang berkurang secara berkala, dimana dokumen-dokumen pada masa lampau diberikan faktor pembobotan yang lebih kecil dibandingkan dengan dokumen-dokumen yang lebih baru.