ANALISIS DAN IMPLEMENTASI FUZZY HIERARCHICAL AGGLOMERATIVE CLUSTERING UNTUK PELABELAN BERITA BERBAHASA INDONESIA BERHIRARKI

Nora Novita Sianturi[1], Imelda Atastina[2], Intan Nurma Yulita[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Clustering dokumen berita diperlukan untuk dapat mengelola informasi menjadi sebuah pengetahuan yang berguna. Salah satu metode yang dapat digunakan untuk klasterisasi dokumen berita ini adalah metode Fuzzy Hierarchical Agglomerative Clustering (FHAC). Metode FHAC mempunyai kemampuan untuk menemukan, menganalisa, dan menggabungkan daerah data dengan cepat. Dengan menggunakan fungsi evaluasi fuzzy, metode FHAC dapat menemukan skema clustering yang paling optimal. Kualitas cluster yang dihasilkan bergantung dengan nilai beberapa parameter inputan, yaitu A (threshold merging cluster) dan K (jumlah tetangga). Untuk data yang ketidakmiripannya kecil/ jarak antar titik kecil/ jumlah label dokumen sedikit lebih tepat menggunakan nilai K yang besar dibandingkan dengan penggunaan nilai K yang kecil, begitu pula sebaliknya pada data dengan kemiripan yang besar/ jarak antar titik besar/ jumlah label dokumen banyak menggunakan nilai K yang kecil. Setiap cluster yang telah terbentuk diberi label yang paling merepresentasikan cluster, yaitu dengan melihat label dominan dalam cluster. Kualitas cluster dan akurasi pelabelan cluster diukur dengan silhouette coefficient dan precision.

Kata Kunci : Klasterisasi, FHAC, Pelabelan, Silhouette Coefficient, Precision

Abstract

Clustering news document is necessary to carry out information become an important knowledge. One of method that can be used for news document clustering is Fuzzy Hierarchical Agglomerative Clustering (FHAC). FHAC have ability to find, analyze, and merge data area fast. Accompanying fuzzy evaluation function, FHAC can find the most optimum scheme. Quality of cluster depend on the value of some parameter input, A (threshold for merging clusters) and K (number of neighbour). Data with a small dissimiliarity/small distance/few label will be more appropriate using big K compared with the small ones, and vice versa data with a big dissimiliarity/big distance/many label using small K. Every cluster that has been created will be labeled by the most representatif label that can be identified the most dominant label on the clusters. The qualiaty of cluster and accuracy of labelling cluster measured by silhouette coefficient and precision.

Keywords : Clustering, FHAC, Labelling, Silhouette Coefficient, Precision

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam era yang berkembang saat ini, jumlah informasi yang ada dan tersebar sangatlah banyak. Namun berlimpahnya informasi ini justru menjadi sulit karena umumnya informasi yang yang ada beraneka macam dan tidak terstruktur sehingga mempersulit para pencari informasi untuk menemukan informasi sesuai dengan yang mereka cari. Hal inilah yang kemudian mendorong meningkatnya kebutuhan untuk menemukan dan mengelola informasi tersebut dengan baik sehingga dapat dihasilkan pengetahuan yang berguna.

Salah satu komponen yang penting dalam pengelolaan informasi untuk pengelompokan berita adalah dengan *clustering*. *Clustering* ada dua macam, yaitu *clustering* tipe hierarki (*hierarchical clustering*) dan clustering tipe non hierarki (*partitioned clustering*). Pada tugas akhir ini akan digunakan *hierarchical clustering*. Ide pada *hierarchical clustering* adalah dengan adanya pengelompokkan data dengan membuat suatu hirarki, dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan[7].

Salah satu metode dalam *hierarchical clustering* adalah *Hierarchical Agglomerative Clustering*. Pada *Hierarchical Agglomerative Clustering*, setiap obyek atau observasi dianggap sebagai sebuah cluster tersendiri. Dalam tahap selanjutnya, dua cluster yang mempunyai kemiripan digabungkan menjadi sebuah cluster baru demikian seterusnya.

Untuk peningkatan ke-efektifitasan dari *Hierarchical Agglomerative Clustering* pada klasterisasi berita berbahasa Indonesia akan digabungkan dengan pengimplementasian fungsi evaluasi *fuzzy*. Hasil penelitian menunjukkan bahwa *Fuzzy Hierarchical Clutering* mempunyai efisiensi *clustering* yang lebih tinggi dan presisi[10]. *Fuzzy hierarchical clustering* mempunyai kemampuan dalam menemukan daerah-daerah berkonsentrasi tinggi dengan metode *Hierarchical Agglomerative Clustering* yang dapat menganalisa dan menggabungkan daerah data dengan cepat. Kemudian, menggunakan fungsi evaluasi untuk menemukan skema *clustering* optimal [10] .

1.2 Perumusan Masalah

Pada tugas akhir ini, masalah yang akan diselesaikan adalah :

1. Bagaimana penerapan fungsi evaluasi *fuzzy* pada *Hierarchical Agglomerative Clustering* ?
2. Bagaimana proses pembentukan kluster dengan metode *Fuzzy Hierarchical Agglomerative Clustering* dalam proses klasterisasi serta pelabelan *cluster* dari dokumen berita berbahasa Indonesia berhirarki?
3. Bagaimana kualitas hasil klasterisasi dari pengimplementasian metode *Fuzzy Hierarchical Agglomerative Clustering* dalam proses klasterisasi serta pelabelan *cluster* dari dokumen berita berbahasa Indonesia berhirarki?

Batasan masalah pada tugas akhir ini, yaitu :

1. Dataset yang akan digunakan merupakan artikel berita berbahasa Indonesia yang bersifat *offline*.
2. Tidak menangani *preprocessing*

1.3 Tujuan

Tujuan dari penelitian tugas akhir ini adalah :

1. Menerapkan fungsi evaluasi *fuzzy* pada metode *Hierarchical Agglomerative Clustering.*
2. Mengetahui serta menerapkan metode *Fuzzy Hierarchical Clustering* dalam proses klasterisasi serta pelabelan *cluster* dari dokumen berita berbahasa Indonesia berhirarki.
3. Mengetahui kualitas hasil klasterisasi dari pengimplementasian metode *Fuzzy Hierarchical Agglomerative Clustering* dalam proses klasterisasi serta pelabelan *cluster* dari dokumen berita berita berbahasa Indonesia berhirarki.

1.4 Metodologi Penyelesaian Masalah

Metodologi yang digunakan untuk penyelesaian masalah pada tugas akhir ini adalah :

1. Studi Literatur

   Pencarian yang layak mengenai text mining, *Hierarchical Agglomerative Clustering*, dan *Fuzzy Hierarchical Agglomerative Clustering* serta pelabelan *cluster.*

2. Desain Penelitian

   Untuk mencapai tujuannya, metode *Hierarchical Agglomerative Clustering* dengan *fuzzy* akan diimplementasikan secara langsung oleh peneliti. Dari hasil desain penelitian ini juga nantinya dapat diketahui kinerja *Fuzzy Hierarchical Agglomerative Clustering* ini pada proses klasterisasi berita berbahasa Indonesia serta pelabelan *cluster*.

3. Pengujian dan Analisa Hasil

   Melakukan proses pengujian terhadap kualitas hasil *Hierarchical Agglomerative Clustering* dengan *fuzzy,* serta melakukan analisa terhadap kelebihan dan keterbatasan *Hierarchical Agglomerative Clustering* dengan *fuzzy* untuk proses klasterisasi berita berbahasa Indonesia serta pelabelan *cluster*.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. Parameter *K* dan *A* mempengaruhi pembentukan dan hasil *cluster*.
2. Untuk menghasilkan kualitas *cluster* yang baik, pemilihan nilai *K* pada *FHAC* harus disesuaikan dengan kondisi data yang ada. Untuk data yang kemiripan antar titiknya tidak terlalu jauh (jumlah label dokumen sedikit), penggunakan *K* yang besar, lebih baik dibandingkan penggunaan nilai *K* yang kecil. Sebaliknya jika jaraknya dekat penggunan *K* dan *A* yang kecil lebih baik.
3. *FHAC* menghasilkan kualitas *cluster* yang baik tetapi dengan jumlah *cluster* yang tidak sesuai dengan *cluster* dokumen sebenarnya.
4. Pengimplementasian fungsi evaluasi *fuzzy* pada *Hierarchical Agglomerative Clustering* adalah sebuah pilihan yang baik untuk meningkatkan kualitas hasil klasterisasi.

5.2 Saran

1. Diperlukan adanya penanganan *feature selection* untuk memberikan kualitas *cluster* yang semakin baik pada algoritma *FHAC* ini.
2. Diperlukan adanya suatu metode untuk menangani ketidakmampuan realokasi *cluster* pada metode *FHAC* sehingga dapat memberikan hasil *cluster* yang lebih baik.

Daftar Pustaka

[1] Adiwijaya. 2006. *Text Mining and Knowledge Discovery*. Kolokium bersama Komunitas Data Mining Indonesia & Softcomputing Indonesia

[2] Berkhin, Pavel. 2002. *Survey of Clustering Data Mining Techniques*. Accure Software. Inc.

[3] Dubes, R, C, and A,K, Jain. 1988. *Algorithms for Clustering Data*. Prentice Hall

[4] *Fan Weiguo, Wallace Linda, Rich Stephanie, and Zhang Zhongju, Tapping into the Power of Text Mining. 2005.* accepted for publication at the Communications of ACM. 2005

[5] Feldman, Ronen dan James Sanger. 2007. *The Text Mining Handbook,* New York: Cambridge University Press

[6] GhasemiGol, Mohammad, Yazdi, Hadi Sadoghi, Monsefi, Reza. 2010. *A New Hierarchical Clustering Algorithm on Fuzzy Data* (FHCA). International Journal of Computer and Electrical. Vol.2, No, 1, February, 2010

[7] Jiawei, Han and Kamber, Micheline. 2006. *Data Mining : Concept and Techniques. 2nd ed.* San Fransisco: Morgan Kaufman Publisher

[8] Kantardzic, Mehmed. 2003. *Data Mining Concepts Models, Methods, and Algorithms,* New Jersey: IEEE

[9] Kariyam dan Subanar. *Comparison of Some Criterion Indexes for Determining the Optimal Number of Clusters*. Yogjakarta : Universitas Gadjah Mada

[10] Li, Ling-Juan and Liang, Yu-Long. 2010. *A Hierarchical Fuzzy Clustering Algorithm.* Nanjing-China: Coll, of Comput, Nanjing Univ, of Posts & Telecommun

[11] Liu, Liu, Chen Ma. *An Evaluation of feature selection for clustering*. ICML Conference 2003

[12] Steinbach Michael, Karypis George, Kumar Vipin. *A Comparison of Document Clustering Techniques.* Minneapolis: Department of Computer Science / Army HPC Research Center. University of Minnesota

[13] Manning, Christopher D, Raghavan Prabhakar and Schütze Hinrich. 2008 *Introduction to Information Retrieval*. Cambridge University Press

[14] Miyamoto, Sadaaki, Ichihashi Hidetomo, Honda Katsuhiro. 2008. *Algortihms for Fuzzy Clustering*: *Methods in C-Means Clustering with Applications*. Heidenberg: Springer

[15] Oliveira, J, Valente de. 2007. *Advances in Fuzzy Clustering and its Applications*. England: John Wiley & Sons, Ltd

[16] Sato-Ilic, Mica and Jain, Lakhmi C. 2006. *Innovations in Fuzzy Clustering*. New York: Springer

[17] Treeratpituk, Pucktada and Callan, Jamie. 2006. *Automatically labeling hierarchical clusters* In Proc of the Sixth National Conference on Digital Government Research