PENGEMBANGAN SISTEM AGREGATOR BERITA BAHASA INDONESIA MENGGUNAKAN *CONTENT EXTRACTION* DAN *HIERARCHICAL AGGLOMERATIVE CLUSTERING*

Stenly Tirta Wijaya [1] **Viny Christanti M.** [2] **Janson Hendryli** [3]

[1] Teknik Informatika Universitas Tarumanagara
Jl. Let. Jend. S. Parman No. 1, Jakarta 11440 Indonesia
email : 535120063@fti.untar.ac.id [1] viny@untar.ac.id [2] jansonh@fti.untar.ac.id [3]

ABSTRACT

The main focus of this study is to develop system to aggregate Indonesian online newspaper and cluster it according to its topic automatically. The system use content extraction to get the main content of articles and Hierarchical Agglomerative Clustering to group articles by its topic with Dice Similarity Coefficient for similarity measure. To determine the cutting point, we cut dendrogram where the gap between two successive combination similarities is largest. Additionally, we add threshold to limit cutting area to improve cluster result. We use Standard Boolean Model for searching feature and Silhouette to evaluate cluster results. Test results using 998 articles shows that limiting cutting area with 0.1 and 0.5 can produce highest average silhouette value 0.264.

Key words

Content Extraction, Dice Similarity Coefficient, Hierarchical Agglomerative Clustering, Silhouette, Standard Boolean Model