INTISARI

PERBANDINGAN PENGARUH ALGORITMA CLUSTERING PADA SISTEM PENCARIAN DOKUMEN

Oleh

Rochana Prih Hastuti

12/331363/PA/14626

Sistem pencarian dokumen dalam konteks *information retrieval* secara cepat menjadi dominan dalam hal akses informasi, menyaingi pencarian database secara tradisional. Sistem pencarian dokumen memberikan hasil berupa dokumen yang memiliki rangking relevansi terhadap kueri yang diberikan pengguna. Beberapa model representasi dalam sistem ini telah dikembangkan misalnya dengan *Vector Space Model.* Pada umumnya, kendala yang muncul pada sistem pencarian dokumen adalah waktu pencarian dokumen yang lama.

Clustering merupakan salah satu pendekatan yang bisa digunakan untuk menangani masalah waktu pencarian. Proses *clustering* pada penelitian ini digunakan untuk mengelompokan dokumen menjadi gugusan yang lebih kecil yang kemudian digunakan sebagai basis data dalam pencarian. Algoritma *k-means, bisecting k-means* dan *hierarchical agglomerative centroid clustering* diterapkan pada sistem pencarian Hadits Shahih Bukhari. Kualitas masing-masing algoritma dalam menjalankan *clustering* dievaluasi dari nilai *purity* dan nilai *silhouette*. Sementara kinerja algoritma *clustering* pada sistem pencarian dokumen dievaluasi dari waktu pencarian, presisi, dan recall dari hasil pencarian.

Berdasarkan pengujian, algoritma *k-means* memiliki kualitas terbaik dengan nilai *purity* rata-rata 0,49 dan nilai *silhouette* 0,03. Selain itu, implementasi algoritma tersebut dengan k = 93 terbukti dapat mempercepat waktu pencarian yakni rata-rata 12 kali dibandingkan sistem pencarian biasa. Kualitas hasil pencarian yang dihasilkan oleh *k-means* pada dua macam uji coba juga dapat mengungguli algoritma lain dengan nilai presisi 0,315 dan 0,482, begitu juga recall dengan nilai yakni 0,131.

Kata kunci: *information retrieval, vector space model, k-means, bisecting k-means.*

ABSTRACT

A COMPARATIVE STUDY OF CLUSTERING ALGORITHM IMPACT IN DOCUMENT RETRIEVAL

By

Rochana Prih Hastuti

12/331363/PA/14626

A document searching system in the context of information retrieval fast become dominant in information access, overtaking traditional database searching. Information retrieval results list of ranked documents to user in relevance of the query given. A number of representation models used in information retrieval system have been developed, one of them is *Vector Space Model.* Generally, the problem that arise in information retrieval system is runtime of the system to retrieve the documents that is slow.

Clustering is an approach that can resolve runtime problem. Clustering in this research is used to cluster the document collections to some smaller groups that further used as the main collection in an information retrieval system. K-means, bisecting k-means, and hierarchical agglomerative clustering are implemented in the searching system of Hadith Shahih Bukhari. The quality of the clusters made by each algorithm evaluated based on the purity score and silhouette value. While the performance of the implementation of these algorithms in the information retrieval system is evaluated through runtime, precision, and recall of the search results.

Based on the experiments, *k-means* get the best average purity score at 0,49 and best silhouette value at 0,03. Besides, the implementation of these algorithms with k = 93 proven can accelerate searching time 12 times faster than the ordinary information retrieval system. The quality of searching results using k-means achieved the best value in precision of both experiments compared to other algorithms which are 0,315 and 0,482 respectively, as well as the recall value is 0,131.

Keywords: *information retrieval, vector space model, k-means, bisecting k-means.*