ABSTRACT

Text mining has been developed in distributed system due to increasing data. One of text mining problems is document or phrase clustering based on similarity without knowing their categories. The popular clustering Algorithms are K-Means (KM) and Fuzzy C-Means (FCM). By combining with MapReduce Algorithm both KM and FCM can become Distributed K-Means (DKM) and Distributed FCM (DFCM). Comparison of both algorithms in Hadoop environment is already performed by many researchers. However, performance comparison between DKM and DFCM that use Cosine Distance Measure (CDM) and Jaccard Distance Measure (JDM) has not been studied yet. Therefore this work compares DKM and DFCM that use CDM and JDM to obtain the best combination of algorithm and distance measure..

Work flow of comparing both algorithms is started by converting the Yahoo Answers dataset in CSV format to 1,400,000 text files and then convert the text files to sequential format. Sequencing files format is mandatory if we want to process a dataset in Hadoop environment. After the file is in sequential format, it is converted to vector and performed preprocessing using tokenization, TF, pruning, dan TF-IDF.. After that the vector files are processed in two testing scenarios which are first scenario and second scenario. In first scenario, all dataset is clustered using ten clusters (c = 10) in local, 2 nodes, 4 nodes, and 8 nodes. Then the computational time is recorded and all clusters are evaluated with inter-cluster density and intra-cluster density. In second scenario, all dataset, half of dataset, a quarter of dataset, and 10% of dataset is clustered using ten clusters (c = 10) in 8 nodes.

The result shows that combination of DKM and CDM called DKM-C has same cluster quality with combination of DKM and JDM called DKM-J. As well as combination of DFCM and CDM called DFCM-C has same cluster quality with combination of DFCM and JDM called DFCM-J. DFCM have smaller inter-cluster density and greater intra-cluster density than DKM’s inter-cluster and

intra-cluster density. Based on that, DFCM has better cluster quality than DKM. More over, DFCM-J has smallest computation time. So DFCM-J becomes best combination of all because it has the best quality cluster and the smallest computation time.

Keywords: Distributed Clustering, Hadoop, FCM, K-Means, MapReduce, Jaccard, Cosine

INTISARI

Penambangan Teks telah dikembangkan di sistem terdistribusi karena peningkatan jumlah data yang perlu diproses. Salah satu masalah pada Penambangan Teks adalah penklasteran dokumen atau frase berdasarkan kemiripan tanpa mengetahui kategorinya. Dua algoritme penklasteran yang populer adalah K-Means (KM) dan Fuzzy C-Means (FCM). Jika dikombinasikan dengan MapReduce maka kedua algoritme tersebut dapat menjadi Algoritme Distributed KM (DKM) dan Distributed FCM (DFCM). Perbandingan antara kedua algoritme di lingkungan Hadoop sudah dilakukan oleh berbagai peneliti. Namun studi tentang perbandingan algoritme antara DKM dan DFCM yang menggunakan Pengukuran Jarak Jaccard (PJJ) dan Pengukuran Jarak Cosine (PJC) masih belum dilakukan. Untuk itu dilakukan perbandingan DKM dan DFCM yang menggunakan PJT dan PJC untuk memperoleh kombinasi algoritme dan pengukuran jarak terbaik.

Alur pengujian yang dilakukan dimulai dari *dataset* Yahoo Answers yang berformat CSV dirubah menjadi 1.400.000 file teks dan kemudian semua file teks tersebut dirubah menjadi format sequential. Merubah file ke format sequential adalah langkah yang wajib dilakukan karena hanya file yang berformat sequential saja yang dapat diproses di Hadoop. Selanjutnya file sequential tersebut dirubah menjadi vector dan dilakukan prapengolahan dengan menggunakan tokenization, TF, pruning, dan TF-IDF. Selanjutnya dilakukan pengujian dengan menklaster *dataset* tersebut menjadi 10 klaster. Percobaan yang dirancang menjadi dua skenario yaitu skenario pertama dan skenario kedua. Pada skenario pertama, *dataset* yang utuh diklaster dengan algoritme DKM dan DFCM di komputer tunggal atau lokal, klaster Hadoop yang nodenya berjumlah 2, 4 dan 8. Waktu komputasi diukur untuk tiap percobaan dan kepadatan *intra-cluster* dan *inter-cluster* dihitung sebagai parameter evaluasi. Sedangkan pada skenario kedua, *dataset* yang utuh, 50 % *dataset*, 25 % *dataset*, dan 10% *dataset* diklaster dengan algoritme DKM dan DFCM di klaster Hadoop yang nodenya berjumlah 8. Waktu komputasinya diukur untuk setiap pengujian dataset.

Hasil pengujian menunjukan bahwa kombinasi dari DKM dan PJC yang disebut dengan DKM-C memiliki kualitas klaster yang sama dengan kombinasi dari DKM dan PJJ yang disebut dengan DKM-J. Seperti halnya pada kombinasi dari DFCM dan PJC yang disebut dengan DFCM-C memiliki kualitas klaster yang sama dengan kombinasi dari DFCM dan PJJ yang disebut dengan DFCM-J. DFCM memiliki nilai *inter-cluster* lebih kecil dan nilai *intra-cluster* lebih besar dibandingkan dengan DKM. Sehingga DFCM memiliki kulitas klaster yang lebih baik jika dibandingkan dengan DKM. Kemudian berdasarkan waktu komputasi, DFCM-J membutuhkan waktu pemrosesan yang paling rendah. Sehingga DFCM-J menjadi kombinasi yang paling baik karena memiliki kualitas klaster yang terbaik dan waktu komputasi yang paling rendah.

Kata kunci: Penklasteran Terdistribusi, Hadoop, FCM, K-Means, MapReduce, Jaccard, Cosine