ABTRACT

The popularity of Bus Rapid Transit (BRT) which makes Trans Jogja as alternative of mass public transport for urban mobility. However, without supervision on temporal behavior patterns of passengers on Trans Jogja supply and demand will result decreases the number of BRT users and the increasing number of private vehicle users so that traffic jam remains difficult to avoid. Smart Card Automated Fare Collection System (SCAFCS) which is used as public transport e-ticketing in Trans Jogja can be used to analyze the pattern of the passengers with approaches data clustering in data mining techniques. In this study applies the algorithm *k-means++* clustering as optimization for *simple k-means* random initiation, and the implementation of *Hadoop Platform* as distributed computing and data preprocessing with data warehouse to improve *k-means++* data clustering performance on the scalability of large datasets, in this case SCAFCS Trans Jogja has a large dataset (volume) and rapid growth data (velocity).

Scalable algorithms scalable *k-means++* generating five clusters with characteristic number of clusters, namely Very Low, Average, High, Very High. 5 clusters used to extract patterns of passengers based on the dimensions of time (temporal) segmentation of passengers (structure) to determine the variability of passengers bases on card they used and transaction peak on boarding location (spatio). Experimental setup is done by comparing three algorithms, *simple k-means* and *k-means++* and implementation of *Hadoop Platform* as parallel and distributed computing, by comparing Sum of Square Error (*SSE*), which is total square error k *cluster* at the centroid, and *Silhouette* Coefficient (*SC*) to validate the strength of the *cluster* and data quality that placed in clusters. *k-means++* with *Hadoop Platform* implementation generates smaller SSE value than *simple k-means* and *k-means++* algorithms, shows it has good SSE value with increasing the number of clusters. The result of *Silhouette* Coefficient shows the comparison value is significant enough, with simple *k-means* algorithm average value is in range 0:25 <SC <= 0.5 or in weak structure category and *k-means++* 0.5 <SC <= 0.7 or in medium structure category.

Keywords– *Data mining, Data Warehouse, Smart Card Data, BRT Trans Jogja, k-means++ Clustering, Hadoop Platform.*

INTISARI

Semakin popularnya *Bus Rapid Transit* (BRT) yang menjadikan Trans Jogja sebagai moda transportasi publik massal alternatif berbasis sistem transit yang cepat dan murah untuk mobilitas perkotaan. Akan tetapi tanpa pengawasan masalah pola perilaku *temporal* penumpang terhadap *supply* dan *demand* bus Trans Jogja berdasarkan waktu akan mengakibatkan penurunan jumlah penumpang BRT dan naiknya jumlah pemakaian kendaraan pribadi sehingga kemacetan tetap sulit terhindarkan. *Smart Card Automated Fare Collection System* (SCAFCS) yang digunakan sebagai *e-ticketing* transportasi publik Trans Jogja dapat dimanfaatkan untuk menganalisis pola penumpang dengan pendekatan pengelompokan data teknik *data mining.* Dalam penelitian ini menerapkan algoritme *k-means++ clustering* sebagai optimasi dan penyelesaian masalah inisiasi *random* pada *simple k-means*, dan penerapan *Hadoop Platform* juga pendekatan *preprocessing* dengan *data warehouse* untuk meningkatkan performa pengelompokan data algoritme *k-means++* pada skalabilitas *dataset* besar, dalam hal ini SCAFCS Trans Jogja memiliki *dataset* besar (*volume)* dan pertumbuhan *data* yang cepat (*velocity).*

Algoritme *scalable k-means++* menghasilkan lima *cluster* dengan karakteristik jumlah kelompok, yaitu: *Very Low*, *Average*, *High, Very High* yang digunakan untuk mengekstraksi pola penumpang berdasarkan dimensi waktu (*temporal*) segmentasi jenis penumpang (*structure*) untuk mengetahui keberagaman penumpang berdasarkan kartu yang digunakan dan *peak* transaksi penumpang pada lokasi keberangkatan (*spatio*). Pengujian dilakukan dengan perbandingan tiga algoritme *simple k-means* dan *k-means++* dan penerapan *Hadoop Platform* sebagai komputasi terdistribusi. Pengujian *Sum of Square Error* (SSE) yang menyatakan total kesalahan kuadrat *k cluster* pada *centroid,* dan *Silhouette Coefficient* (SC) untuk memvalidasi kekuatan *cluster* dan kualitas *data* yang ditempatkan dalam suatu *cluster*. *k-means++* dengan penerapan *Hadoop Platform* menghasilkan nilai SSE yang lebih kecil daripada algoritme *simple k-means* dan *k-means++*, menunjukkan nilai SSE yang baik dengan bertambahnya jumlah *cluster*. Hasil pengujian *Silhouette Coefficient* menunjukkan perbandingan nilai yang cukup signifikan, yaitu rata-rata nilai algoritme *simple k-means* pada 0,25 < SC <= 0,5 atau pada kategori *Weak Structure* dan *k-means++* 0,5 < SC <= 0,7 atau pada kategori *Medium Structure.*

Kata kunci – *Data mining, Data Warehouse, Smart Card Data, BRT Trans Jogja, Scalable k-means++ Clustering, Hadoop Platform.*