INTISARI

ANALISIS SELF ORGANIZING MAP (SOM) UNTUK IMPLEMENTASI DATA ANALYSIS PADA BASIS DATA TERDISTRIBUSI DI SISTEM BIG DATA

Oleh

Arif Ardiasmono

13/348719/PA/15474

Banyak perusahaan beralih dari penggunaan basis data relasional menuju basis data terdistribusi karena memuat lebih banyak data serta dapat tersebar di berbagai lokasi. Cepatnya pertumbuhan data menuntut adanya peningkatan waktu pemrosesannya. Analisis perlu dilakukan untuk mengetahui adanya pola atau kemiripan antardata multi-dimensi. Beberapa penelitian menunjukkan *Self Organizing Map* (SOM) dapat digunakan untuk mengklaster data berdimensi tinggi. SOM melatih data secara *unsupervised* sehingga tidak diperlukan variabel yang ditargetkan. Dengan metode ini, sekumpulan data multi-dimensi dapat dikelompokkan berdasarkan kemiripannya dan tersebar merata dalam klaster.

Pada penelitian ini diambil *dataset* multi-dimensi seperti abstrak penelitian yang diimpor dalam basis data terdistribusi HBase secara *map-reduce*. Ekstraksi fitur kemudian dilakukan agar *dataset* dapat dilatih dengan metode SOM. Selama pelatihan dilakukan pencatatan waktu untuk mengetahui kecepatan kinerjanya. Setelah pelatihan, *record* dalam basis data diuji dan dicatat jumlah data di tiap klasternya untuk mengetahui pesebaran hasil pengelompokkan. Terakhir kedua parameter uji tersebut dihitung korelasinya untuk mengetahui tingkat keterkaitannya.

Hasil analisis didapatkan waktu kompleksitas SOM sebesar $T(k,r) = O(k^2 r) + O(k^2) + O(r)$ dan dari hasil pengujian didapatkan bahwa algoritma SOM dapat diimplementasikan pada basis data terdistribusi untuk *dataset* internet dan ETD dengan perbandingan ukuran 1,85 : 180,77 dan jumlah *record* 1014 : 84354 dihasilkan perbandingan waktu komputasi dengan rata-rata 71,98 : 5474,72 dan simpangan baku sebaran jumlah data sebesar 73,12 : 8556,51. Pengujian korelasi menghasilkan adanya keterhubungan antara kedua parameter yang dinalisis, yaitu sebesar -0,52 untuk *dataset* internet dan -0,63 untuk *dataset* ETD.

Kata kunci: self organizing map, map reduce, hbase, hadoop, distributed database

ABSTRACT

SELF ORGANIZING MAP (SOM) ANALYSIS FOR DATA ANALYSIS IMPLEMENTATION ON DISTRIBUTED DATABASE IN BIG DATA SYSTEM

By

Arif Ardiasmono

13/348719/PA/15474

Many companies shift from using relational databases to distributed databases because they contain more data and can be scattered in multiple locations. Rapid data growth requires an increase in processing time. Analysis needs to be done to determine the pattern or similarity of multi-dimensional data. Several studies show Self Organizing Map (SOM) can be used to cluster high-dimensional data. SOM trains data unsupervisely so that no targeted variables are required. With this method, a set of multi-dimensional data can be reduced and grouped by their similarity and spread evenly in clusters.

In this study, HBase distributed databases import high-dimensional dataset of research abstract using map-reduce. Feature extraction then performed on the dataset in order to be trained with the SOM method. Recording time during the training is done to determine the speed of its performance. After the training, records in the database are tested and recorded the amount of data in each cluster to know the spread of grouping results. Finally, the computation time parameter and standard deviation between clusters correlation coefficient are calculated to know the level of connectedness.

The result shows that the SOM complexity is $T(k,r) = O(k^2 r) + O(k^2) + O(r)$ and from the test result shows that SOM algorithm can be implemented in distributed database for internet dataset and ETD with the ratio size of 1.85 : 180.77 and the number of records 1014 : 84354 computational time is generated with an average of 71.98 : 5474,72 and standard deviation of the distribution of data amounted to 73.12 : 8556.51. The correlation test resulted in a correlation between the two parameters analyzed, ie -0.52 for the internet dataset and -0.63 for the ETD dataset.

Keywords: self organizing map, map reduce, hbase, hadoop, distributed database