§28. Study on Quick Similarity Retrieval in Massivesize Diagnostic Databases

Hochin, T., Nomiya, H. (Kyoto Inst. of Tech.),
Okumura, H. (Mie Univ.),
Nakanishi, H., Kojima, M., Nagayama, Y., Ohdachi, S., Emoto, M., Ohsuna, M.

Experiments of the fusion phenomena produce a lot of sequences of time-varying values which form waveforms. If the waveforms similar to a desired one can be obtained by using computer system, the burden of researchers in searching similar waveforms will extremely be decreased. We have addressed to the issue on this kind of retrieval. We have proposed an indexing method of the severely and quickly changing plasma waveforms for accelerating search and retrieval of their subsequences [1]. Dominant feature values are selected as representative ones, and are stored into a multi-dimensional index. As the feature values and multi-dimensional index are stored in the file, the more data increases, the more complicated the management becomes. Moreover, the method of the fast retrieval is only using a single index. Although some computers have two or more CPUs, and can use Redundant Arrays of Inexpensive Disks (RAID), no method using them has been examined yet.

This paper improves the performance of the similarity retrieval system by using the database management system (DBMS) for the efficient management of feature values. PostgreSQL is used in this paper because it supports a multi-dimensional index. Moreover, a table in the database is divided into several tables in order to improve the performance of retrieval through parallel processing.

The waveforms used in the experiments are of the magnetic field fluctuations. The number of the feature values is 133, while that of the representative ones is 15. The representative feature values as well as the shot number and the segment number are stored into a table. Rectangular index range retrieval is performed by using the R tree index. The SQL statement for the range retrieval, which uses the multi-dimensional index, is as follows:

```
SELECT * FROM table_name WHERE rep <@
'[(key1-thrs,..,key15-thrs), (key1+thrs,..,key15+thrs)]'
```

where table_name is the name of the table, rep, which is of the multi-dimensional cube type, holds the values of the representative feature values, key, which is of the double precision type, means the value of the nth ($1<=n<=15$) element of the query waveforms, and thrs, which is of the double precision type, is the value of the threshold.

Waveforms are divided into 256 segments, and their feature values are stored into several tables. Retrieval processes, each of which retrieves data from a table, are executed in parallel. The processing times are measured. 1000 waveforms are used in the experiment. The numbers of division are two, four, and eight. A table has the feature values of 500 (250, and 125, respectively) waveforms for division by two (four, and eight). The times of index retrieval concurrently performed to two (four, and eight) tables are measured. The experiments are conducted on two computers: PC1 (Intel Core2 Duo E7500 2.93GHz, 1.96GB memory, PostgreSQL 8.4.0, no RAID) and PC2 (Intel Xeon E5620 2.40GHz, 15.6GB memory, PostgreSQL version 8.1.23, RAID5).

The retrieval times on PC1 (PC2, respectively) for the different numbers of retrieval candidates are shown in Fig. 1 (Fig. 2). The times of two, four, and eight divisions as well as no division are shown. When PC1 is used, the retrieval performance with the division by two is the best as shown in Fig. 1. The performance with the division by four is almost the same as that without division. That with the division by eight becomes worse than that without division. On the other hand, when PC2 is used, the more a table is divided into, the better the retrieval performance becomes as shown in Fig. 2. Division by eight attains the best retrieval performance.

Too much division degrades the retrieval performance on PC1. This may be caused by that the waiting time within CPU of PC1, which is the dual core system, increases. On the other hand, PC2 can concurrently process the retrieval because it has 4 core 8 thread per one CPU. Even if a table is divided into eight, the effect of improvement in the speed by division is maintained as shown in Fig.2.

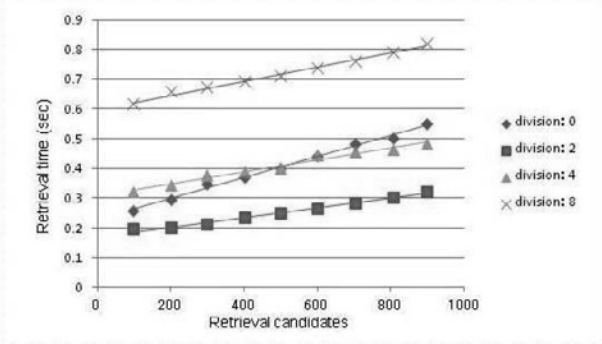

Fig. 1. Retrieval time in varying the number of division on PC1.

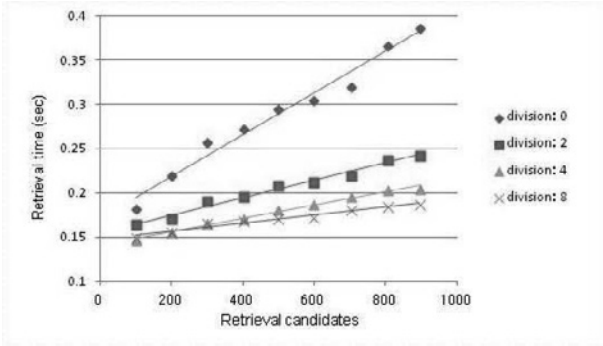

Fig. 2. Retrieval time in varying the number of division on PC2.

1) T. Hochin, Y. Yamauchi, H. Nakanishi, M. Kojima, H. Nomiya: Indexing of plasma waveforms for accelerating search and retrieval of their subsequences, Fus. Eng. Des., **85**(5) (2010) 649-654.