IPSJ Transactions on Databases Vol. 3 No. 3 88–98 (Sep. 2010)

Regular Paper

mm-GNAT: Index Structure for Arbitrary L_p Norm

Kensuke Onishi,[†1] Michihiro Kobayakawa[†2,∗1]
and Mamoru Hoshi[†2]

For fast ε-similarity search, various index structures have been proposed. Yi, et al. proposed a concept *multi-modality support* and suggested inequalities by which ε-similarity search by L_1, L_2 and L_∞ norm can be realized. We proposed an extended inequality which allows us to realize ε-similarity search by arbitrary L_p norm using an index based on L_q norm. In these investigations a search radius of a norm is converted into that of other norm. In this paper, we propose an index structure which allows search by arbitrary L_p norm, called *mm-GNAT* (*multi-modality support GNAT*), with the extention of ranges of GNAT, instead of extending the search radius. The index structure is based on GNAT (Geometric Near-neighbor Access Tree). We show that ε-similarity search by arbitrary L_p norm is realized on mm-GNAT. In addition, we performed search experiments on mm-GNAT with artificial data and music data. The results show that the search by arbitrary L_p norm is realized and the index structure has better search performance than Yi's method except for search by L_2 norm.

1. Introduction

To search multimedia data and/or time series data, we extract various features for retrieval from original data and search objects in the feature space. In most cases, the feature space is represented as a vector space. In this paper, we focus attention on index structures for ε-similarity search on vector space.

Index structures for fast ε-similarity search have been studied, for example, R-tree [1)], SS-tree [2)], SR-tree [3)], VP-tree [4)], M-tree [5)] and GNAT [6)]. For the other researches than those above see Böhm's survey [7)] and Chávez's survey [8)]. In the index structures, data set is divided into subsets. A retrieval speeds up based on the subdivision of data set. Each subset, called a *cluster*, is constructed based on distance between points. Clusters vary depending on the norm used when the clusters are constructed. Therefore, the index structures depend upon norms used for constructing the clusters. When we execute ε-similarity search, some clusters may not be searched. If a cluster and an ε-ball (which is a region to be searched) in the space have no intersection, the cluster does not contain any correct point of ε-similarity search and then need not to be searched. Intersection check is realized using a distance between the query point and the cluster.

Yi and Faloutsos proposed a concept: *multi-modality support* [9)]. The concept is that a user would search by various similarity models and the index structure must support all similarity models. They considered L_p norm (Minkowski norm) as similarity models and proposed a method which realizes ε-similarity search by arbitrary L_p norm [9)]. They showed an inequality by which a query of L_p norm is converted into that of Euclidean norm (L_2 norm) and performed experiments for L_1 norm and L_∞ norm. Lee, et al. applied this method to minimum distance [10)]. Ciaccia and Patella consider a class of norm which is lower bounded by other norm. They proposed a retrieval method using the lower bound norm and analyzed distance distribution [11)]. The key idea in the methods above is an extension of search radius. Therefore search region becomes larger. In this paper, we propose an index structure for ε-similarity search by arbitrary L_p norm with the extention of ranges of GNAT, instead of extending the search radius.

In Section 2, we explain Yi's method [9)], QIC-m-tree [11)] and GNAT [6)] as related works. In Section 3, we propose an index structure *mm-GNAT* for ε-similarity search and show that ε-similarity search of *arbitrary* L_p norm can be realized by mm-GNAT. In Section 4, we show experimental results of ε-similarity search with mm-GNAT. In Section 5, we discuss the results. In Section 6, we show results of ε-similarity search experiment on music data and discuss the results.

2. Related Works

We explain the framework of the ε-similarity search based on subdivision. The ε-similarity search is executed as follows:

Step.1 determine *unnecessary* clusters which do not contain any correct point for the search (we need not to check the points in the clusters).

Step.2 calculate distances between a query point and the points in the *necessary*

†1 Department of Mathematical Sciences, Tokai University
†2 Graduate School of Information Systems, University of Electro-Communications
∗1 Presently with Tokyo Metropolitan College of Industrial Technology

clusters which may contain correct points of the search.
The more unnecessary clusters are found in Step.1, the fewer the number of distance calculations is, in other words, the cost of search decreases.

2.1 Yi's Method

Yi and Faloutsos showed the following inequalities among L_p norms ($p = 1, 2, \infty$):

$$\mathrm{dist}_2(\boldsymbol{x}, \boldsymbol{y}) \leq \mathrm{dist}_1(\boldsymbol{x}, \boldsymbol{y}),\ \mathrm{dist}_2(\boldsymbol{x}, \boldsymbol{y}) \leq d^{\frac{1}{2}} \cdot \mathrm{dist}_\infty(\boldsymbol{x}, \boldsymbol{y}),$$

where $\mathrm{dist}_p(\boldsymbol{x}, \boldsymbol{y})$ is the L_p distance function for d dimensional vectors $\boldsymbol{x} := (x_1, x_2, \ldots, x_d), \boldsymbol{y} := (y_1, y_2, \ldots, y_d)$:

$$\mathrm{dist}_p(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \left\{ \sum_{i=1}^{d} |x_i - y_i|^p \right\}^{1/p} & (p = 1, 2, \cdots) \\ \max_{i=1}^{d} |x_i - y_i| & (p = \infty) \end{cases}.$$

The following inequality can be easily shown from their result:

$$\mathrm{dist}_2(\boldsymbol{x}, \boldsymbol{y}) \leq d^{\frac{1}{2}} \cdot \mathrm{dist}_p(\boldsymbol{x}, \boldsymbol{y}) \quad (p = 3, 4, \cdots, \infty).$$

By the inequality, ε-similarity search of L_p norm is replaced by $d^{1/2} \cdot \varepsilon$-similarity search of L_2 norm. When we execute an ε-similarity search of L_p norm ($\mathrm{dist}_p(\boldsymbol{x}, \boldsymbol{y}) \leq \varepsilon$) on the index structure of L_2 norm, we execute the Step.1 of the framework of search by using inequalities

$$\begin{aligned} &\mathrm{dist}_2(\boldsymbol{x}, \boldsymbol{y}) \leq \varepsilon && (p = 1, 2), \\ &\mathrm{dist}_2(\boldsymbol{x}, \boldsymbol{y}) \leq d^{\frac{1}{2}} \cdot \varepsilon && (p = 3, 4, \ldots, \infty). \end{aligned}$$

Then the Step.2 of the framework is done.

2.2 QIC-m-tree

Ciaccia and Patella proposed *QIC-m-tree* [11]. They showed *Lower-Bounding property*. They proposed multi-modality support retrieval for a class of norms by scaling of ε and the following property :

$$\begin{aligned} &\mathrm{dist}_q(\boldsymbol{x}, \boldsymbol{y}) \leq \mathrm{dist}_p(\boldsymbol{x}, \boldsymbol{y}) && (p = 1, 2, \ldots, q), \\ &\mathrm{dist}_q(\boldsymbol{x}, \boldsymbol{y}) \leq d^{\frac{1}{q} - \frac{1}{p}} \cdot \mathrm{dist}_p(\boldsymbol{x}, \boldsymbol{y}) && (p = q+1, q+2, \ldots, \infty), \end{aligned}$$

where p, q are positive integers. By the inequality above, we execute ε-similarity search of L_p norm on the index structure of L_q norm, we execute the Step.1 of the framework of search by using inequalities

$$\begin{aligned} &\mathrm{dist}_q(\boldsymbol{x}, \boldsymbol{y}) \leq \varepsilon && (p = 1, 2, \ldots, q), \\ &\mathrm{dist}_q(\boldsymbol{x}, \boldsymbol{y}) \leq d^{\frac{1}{q} - \frac{1}{p}} \cdot \varepsilon && (p = q+1, q+2, \ldots, \infty). \end{aligned}$$

Then the query by L_p norm is executed on L_q based index structure. When $p = 1,\ q = 2$ or $p = \infty,\ q = 2$ in the inequality, we have the Yi's inequalities. The coefficient of the right-hand is tight on ε-similarity search. Kimura, et al. independently proved the same property mentioned above [12].

2.3 GNAT

Brin proposed an index structure GNAT (Geometric Near-neighbor Access Tree)[6]. A set of separate points is selected from data set and is used for subdivision. Points in the space are divided into clusters such that every point in the same cluster D_{SP_i} is closer to the separate point SP_i than to all other separate points.

The algorithm for building GNAT is shown in **Algorithm 1**.

Algorithm 1: Construction algorithm of GNAT	
Input:	original data set, its cardinality is n (the number of data points); k (the number of separate points);
Output:	k separate points (SP_i); k clusters (D_{SP_j}); existence ranges between SP_i and D_{SP_j};
Step.1	Choose k separate points from the data set.
Step.2	Divide the original data set into k clusters D_{SP_j}. Each point in D_{SP_j} is nearer to SP_j than other separate points.
Step.3	For each cluster D_{SP_j}, compute the minimum and the maximum distance between D_{SP_j} and separate points SP_i ($i = 1, \ldots, k, i \neq j$).

In Step.1, separate points are selected. In Step.2, clusters are computed with the separate points. In this step, we compute distance between separate points and all points in the data set and determine the nearest separate point SP_i for each point. The norm used for calculating the distance is called *construction norm*. A cluster D_{SP_j} is the set of points which are nearer to the separate point SP_j than the other separate points. This step corresponds to the computation of

Voronoi diagram for separate points and each cluster corresponds to the Voronoi region of a separate point. In Step.3, for each pair of D_{SP_i} and SP_j $(i \neq j)$, we compute the minimum and the maximum distance between the cluster and the separate point *i.e.*, $\min_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}(SP_i, \boldsymbol{x})$ and $\max_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}(SP_i, \boldsymbol{x})$. We define the *existence range* of SP_i and D_{SP_j} as :

$$\mathrm{range}(SP_i, D_{SP_j}) = \left[\min_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}(SP_i, \boldsymbol{x}), \max_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}(SP_i, \boldsymbol{x}) \right].$$

GNAT has the following records:

- k separate points SP_i $(i = 1, \ldots, k)$;
- cluster D_{SP_j} for a separate point SP_j $(j = 1, \ldots, k)$;
- existence ranges by the construction norm.

When the number of points in a cluster is large, we might apply the construction algorithm to the cluster recursively, that is, the cluster is divided into subclusters recursively. In such case, GNAT has tree structure of inclusion relation.

When we execute ε-similarity search at a query point $\boldsymbol{q}$, we compute the following range

$$[\mathrm{dist}(SP_i, \boldsymbol{q}) - \varepsilon, \mathrm{dist}(SP_i, \boldsymbol{q}) + \varepsilon],$$

called *query range*. *Intersection check* is defined as whether the query range and the existence range $\mathrm{range}(SP_i, D_{SP_j})$ have intersection or not. When the ranges have intersection, the check is *true*, otherwise is *false*. If the intersection check is false, then the cluster D_{SP_j} is *unnecessary* for the search. We apply the intersection check above to all pairs of separate point SP_i and cluster D_{SP_j} (this process corresponds to the Step.1 of the framework of search). For the necessary clusters, we apply Step.2 of the framework, i.e., we compute distance between the query point and each point in the necessary clusters and check whether the distance is less than ε or not.

Suppose a GNAT based on L_1 norm and search by L_2 norm. **Figure 1** shows two ranges: one is based on L_1 norm (dotted line segment) and another is on L_2 norm (solid line segment), called L_1-based range and L_2-based range, respectively. Since L_2 norm is smaller than or equal to L_1 norm for any two points, the L_2-based range exists to the left side of the L_1-based range. So, we can select a query point $\boldsymbol{q}$ and a search radius ε such that the query range intersects with

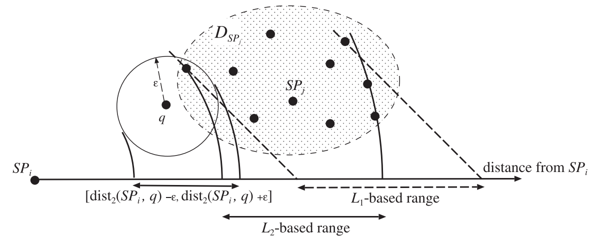

Fig. 1 A case of a necessary cluster being regarded as unnecessary.

the L_2-based range and does not with the L_1-based range. When we execute ε-similarity search by L_2 norm at the $\boldsymbol{q}$, the intersection check is executed. If L_2-based range is used for the check, L_2-based range intersects with the query range and then the cluster D_{SP_j} is considered *necessary*. If L_1-based range is used for the check, L_1-based range does not intersect with the query range and then the cluster D_{SP_j} is considered *unnecessary* (Fig. 1). In the next section, we resolve this problem by extending the existence range.

3. mm-GNAT

In this section we propose an index structure mm-GNAT (multi-modality support GNAT) for ε-similarity search by arbitrary L_p norm.

The following inequalities hold among L_p norms.

Lemma 1 Let $\boldsymbol{x}, \boldsymbol{y}$ be vectors. Then

$$\mathrm{dist}_\infty(\boldsymbol{x}, \boldsymbol{y}) \leq \mathrm{dist}_p(\boldsymbol{x}, \boldsymbol{y}) \leq \mathrm{dist}_1(\boldsymbol{x}, \boldsymbol{y}) \quad (p = 1, 2, \ldots, \infty)$$

hold for any L_p norm.

Proof: This inequality is directly proved from Hölder's inequality. □

From Lemma 1, the existence range of GNAT can be extended well by replacing the lower bound and the upper bound of the existence range with the lower bound measured by L_∞ norm and the upper bound measured by L_1 norm, respectively. We define the *mm-range* as follows:

$$\mathrm{mm\text{-}range}(SP_i, D_{SP_j}) = \left[\min_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}_\infty(SP_i, \boldsymbol{x}), \max_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}_1(SP_i, \boldsymbol{x}) \right]. \quad (1)$$

Figure 2 shows L_1-based range, L_2-based range, L_∞-based range and mm-range.

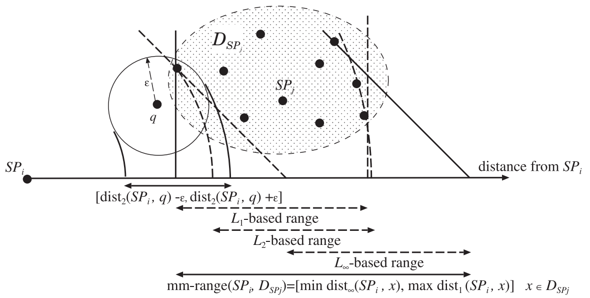

Fig. 2 Intersection check on mm-GNAT.

The mm-range contains these three ranges.

The records other than existence range of GNAT are not changed. Therefore, mm-GNAT has the following records:

- k separate points SP_i $(i = 1, \ldots, k)$;
- cluster D_{SP_j} for a separate point SP_j $(j = 1, \ldots, k)$;
- mm-range(SP_i, D_{SP_j}) $(i, j = 1, \ldots, k, i \neq j)$.

Tree structure of mm-GNAT may be constructed in the similar way to that of GNAT.

We show that ε-similarity search by arbitrary L_p norm can be executed on mm-GNAT. It is sufficient to show that any necessary cluster cannot be regarded as unnecessary. We show that if a query range has intersection with the existence range of L_p norm, then the query range has intersection with the mm-range (1). It is sufficient to show the following two inequalities hold for any p:

$$\min_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}_\infty(SP_i, \boldsymbol{x}) \leq \min_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}_p(SP_i, \boldsymbol{x}),$$

$$\max_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}_1(SP_i, \boldsymbol{x}) \geq \max_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}_p(SP_i, \boldsymbol{x}).$$

We prove the former inequality. The latter is proved similarly. Let $\boldsymbol{y}^\infty$ be a point such that $\mathrm{dist}_\infty(SP_i, \boldsymbol{y}^\infty) = \min_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}_\infty(SP_i, \boldsymbol{x})$ and $\boldsymbol{y}^p$ be a point such that dist_p $(SP_i, \boldsymbol{y}^p) = \min_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}_p(SP_i, \boldsymbol{x})$. Then,

$$\begin{aligned} \min_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}_\infty(SP_i, \boldsymbol{x}) &= \mathrm{dist}_\infty(SP_i, \boldsymbol{y}^\infty) \\ &\leq \mathrm{dist}_\infty(SP_i, \boldsymbol{y}^p) \\ &\leq \mathrm{dist}_p(SP_i, \boldsymbol{y}^p) \\ &= \min_{\boldsymbol{x} \in D_{SP_j}} \mathrm{dist}_p(SP_i, \boldsymbol{x}). \end{aligned}$$

The relation above is shown from the minimality of $\boldsymbol{y}^\infty$ and Lemma 1.

From the discussion above, all necessary clusters under arbitrary L_p norm are surely found by the mm-range. For each point in the necessary clusters, the distance of L_p norm from the query is computed, then the ε-similarity search of L_p norm completes.

Theorem 2 The ε-similarity search by arbitrary L_p norm is performed by a mm-GNAT.

Chávez, et al. showed an analysis for compact partitioning algorithm, which contains GNAT and mm-GNAT, using the average and the variance of distance between data points [8] [Section 7.3].

We also analyze search cost by mm-GNAT in another way. Fix a cluster D_{SP_j}. Consider the probability that the cluster is necessary on ε-similarity search. Whether the cluster is necessary or not is determined based on the intersection between the existence range and the query range. Suppose a data set is contained in d dimensional vector space $[0, 1]^d$, then any existence range of L_p norm is contained in $[0, d^{1/p}]$, where 0 and $d^{1/p}$ are the minimum and the maximum distance of L_p norm in the space. Suppose the distance of L_p norm is uniform on the range $[0, d^{1/p}]$ for the simplicity of analysis. The probability of a cluster being necessary is linear with the width of the cluster's existence range. When ε-similarity search is executed, the probability is expressed as the width plus 2ε divided by the maximum width $d^{1/p}$ of the range (see **Fig. 3**). Thus, the probability of the cluster being necessary is

$$\frac{r_{p,i,j}^{\max} - r_{p,i,j}^{\min} + 2\varepsilon}{d^{1/p}},$$

where $r_{p,i,j}^{\min}$ and $r_{p,i,j}^{\max}$ are the minimum and the maximum L_p norm from a separate point SP_i to the cluster D_{SP_j}, respectively.

Then the intersection check is repeated k times. If the cluster passes all intersection checks, the cluster is necessary and we apply Step.2 of framework to the cluster. So, the probability that a cluster is necessary is

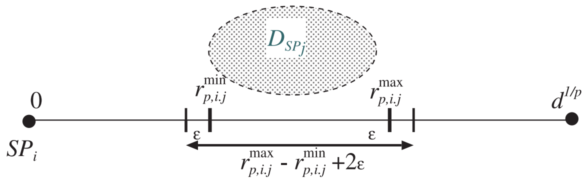

Fig. 3 Probability of a cluster being necessary.

$$\prod_{i=1}^{k}\left(\frac{r_{p,i,j}^{\max}-r_{p,i,j}^{\min}+2\varepsilon}{d^{1/p}}\right).$$

Usually ε is rather small than the width of existence range. The quantity above is approximated as follows:

$$\prod_{i=1}^{k}\left(\frac{r_{p,i,j}^{\max}-r_{p,i,j}^{\min}+2\varepsilon}{d^{1/p}}\right)\sim\left(\frac{\widetilde{\mathrm{diff}}_{p,j}}{d^{1/p}}\right)^{k},$$

where $\widetilde{\mathrm{diff}}_{p,j}$ is the geometric mean of the width of existence range, i.e., $\widetilde{\mathrm{diff}}_{p,j}=\left\{\prod_{i=1}^{k}(r_{p,i,j}^{\max}-r_{p,i,j}^{\min})\right\}^{1/k}$. Therefore, the expectation of the number of distance calculations is expressed by

$$E\left[\sum_{j=1}^{k}\left(\frac{\widetilde{\mathrm{diff}}_{p,j}}{d^{1/p}}\right)^{k}\cdot\left|D_{SP_j}\right|\right].$$

Assume the following two conditions:

- each cluster contains n/k points on the average, where n is the number of data points;
- the probability $\widetilde{\mathrm{diff}}_{p,j}/d^{1/p}$ that a cluster is necessary is independent of the probability of the other clusters being necessary.

Under these assumptions, the expectation is computed:

$$\frac{n}{k}\cdot E\left[\sum_{j=1}^{k}\left(\frac{\widetilde{\mathrm{diff}}_{p,j}}{d^{1/p}}\right)^{k}\right]=\frac{n}{k}\cdot\sum_{j=1}^{k}E\left[\left(\frac{\widetilde{\mathrm{diff}}_{p,j}}{d^{1/p}}\right)^{k}\right]=\frac{n}{k}\cdot\sum_{j=1}^{k}\left(E\left[\frac{\widetilde{\mathrm{diff}}_{p,j}}{d^{1/p}}\right]\right)^{k}$$
$$=\frac{n}{k}\cdot\sum_{j=1}^{k}\left(\frac{E[\widetilde{\mathrm{diff}}_{p,j}]}{d^{1/p}}\right)^{k}=\frac{n}{k}\cdot k\cdot\left(\frac{E[\widetilde{\mathrm{diff}}_{p,j}]}{d^{1/p}}\right)^{k}=n\left(\frac{E[\widetilde{\mathrm{diff}}_{p,j}]}{d^{1/p}}\right)^{k},$$

where $E[\widetilde{\mathrm{diff}}_{p,j}]$ is the expectation of $\widetilde{\mathrm{diff}}_{p,j}$ and denoted by $\overline{\mathrm{diff}}_p$ below.

Finally, adding the expectation above to the number of distance calculation between the query point and k separate points to determine the necessity of clusters, we have

$$k+n\left(\frac{\overline{\mathrm{diff}}_p}{d^{1/p}}\right)^{k}. \tag{2}$$

This value is the expectation of the total number of distance calculations.

Similarly, the number of distance calculations of mm-GNAT can be analyzed. In this case, the maximum width on mm-range is d, and $\widetilde{\mathrm{diff}}_{\mathrm{mm},j}=\left\{\prod_{i=1}^{k}(r_{1,i}^{\max}-r_{\infty,i}^{\min})\right\}^{1/k}$. Let $\overline{\mathrm{diff}}_{\mathrm{mm}}$ be the expectation of $\widetilde{\mathrm{diff}}_{\mathrm{mm},j}$. The expectation of the number of total distance calculations is

$$k+n\left(\frac{\overline{\mathrm{diff}}_{\mathrm{mm}}}{d}\right)^{k}. \tag{3}$$

The first terms of (2) and (3) are fixed when the index structures, GNAT and mm-GNAT, are constructed. We focus on the second terms of (2) and (3) and calculate the ratio between the second terms of (2) and (3):

$$\left[n\left(\frac{\overline{\mathrm{diff}}_{\mathrm{mm}}}{d}\right)^{k}\right]\Big/\left[n\left(\frac{\overline{\mathrm{diff}}_p}{d^{1/p}}\right)^{k}\right].$$

Then, we have the following term without n:

$$\left(\frac{\overline{\mathrm{diff}}_{\mathrm{mm}}}{\overline{\mathrm{diff}}_p}\cdot d^{1/p-1}\right)^{k}. \tag{4}$$

This term corresponds to the ratio of the expectation of the number of distance calculations in Step.2 of the framework of mm-GNAT to that of GNAT for L_p norm.

4. Experiment

In this section, we describe experiments and their results. To investigate the performance of mm-GNAT, we implemented three methods below and compared

Table 1 Data sets for experiment.

	dimension	number of points	distribution	type of data
DB_1	4	100,000	uniform	artificial
DB_2	8	100,000	uniform	artificial
DB_3	16	100,000	uniform	artificial
DB_4	20	100,000	non-uniform	music [15]

them:

(1) *standard method*(GNAT): construct GNAT for each L_p norm, and execute ε-similarity search of L_p norm $(p = 1, \ldots, 10, \infty)$;

(2) mm-GNAT: construct a mm-GNAT based on L_q norm[⋆1] $(q = 1, 2, \infty)$, and execute ε-similarity search of L_p norm$(p = 1, \ldots, 10, \infty)$;

(3) *Yi's method*: construct a GNAT based on L_2 norm, and execute ε-similarity search of L_p norm $(p = 1, 2)$ and $d^{1/2} \cdot \varepsilon$-similarity search of L_p norm $(p = 3, 4, \ldots)$, where d is the dimension of the data.

Experiments were executed for 3 artificial data sets (DB_1, DB_2 and DB_3) and a music data (DB_4) in **Table 1**. We describe search experiments on artificial data below. The experiment on music data is shown in Section 6.

For each method we computed the existence range, executed ε-similarity search and counted the number of points which are within ε from a query point (this number is called the *correct number*).

The search performance of ε-similarity search on GNAT depends on clusters. The clusters are computed from separate points with construction norm. A thousand separate points were randomly selected from data set (1% of the data set). We constructed mm-GNATs based on L_1, L_2 and L_∞ norms, called L_1-based, L_2-based and L_∞-based mm-GNATs, respectively.

[Construction time of index structure] **Table 2** (left) shows the construction times of mm-GNAT and GNAT for 4 data sets. In standard method, an index structure has to be constructed for each search norm, therefore, the whole of construction time and storage are linear to the number of search norms which can be used on the database system. In mm-GNAT, we need only *one* index structure. So, construction time and storage are decreased appreciably.

⋆1 This norm is construction norm which is used for constructing clusters of mm-GNAT, not search norm.

Table 2 Construction time (left), $\overline{\mathrm{diff}}_{\mathrm{mm}}$ and $\overline{\mathrm{diff}}_p$ (right) of mm-GNAT and GNAT.

	construction time				$\overline{\mathrm{diff}}_{\mathrm{mm}}$ for mm-range			
mm-GNAT	DB_1	DB_2	DB_3	DB_4	DB_1	DB_2	DB_3	DB_4
L_1-based	236	403	708	892	1.044	2.974	6.447	3.651
L_2-based	245	415	741	840	1.020	2.941	6.405	3.617
L_∞-based	160	270	514	583	1.058	3.038	6.460	3.774
standard method	construction time				$\overline{\mathrm{diff}}_p$ for existence range			
GNAT(L_1 norm)	206	339	586	824	0.463	1.552	3.200	0.766
GNAT(L_2 norm)	255	387	636	905	0.221	0.533	0.777	0.176
GNAT(L_3 norm)	257	390	636	899	0.194	0.412	0.520	0.123
GNAT(L_4 norm)	288	458	827	944	0.188	0.375	0.441	0.107
GNAT(L_5 norm)	264	399	673	928	0.187	0.361	0.407	0.101
GNAT(L_6 norm)	281	440	753	925	0.188	0.355	0.390	0.097
GNAT(L_7 norm)	271	440	705	920	0.188	0.352	0.381	0.096
GNAT(L_8 norm)	263	423	716	1040	0.189	0.351	0.375	0.095
GNAT(L_9 norm)	271	435	767	1038	0.191	0.350	0.372	0.094
GNAT(L_{10} norm)	282	453	686	977	0.191	0.350	0.370	0.094
GNAT(L_∞ norm)	79	121	191	211	0.200	0.361	0.377	0.096

[Existence range of mm-GNAT] In the experiment, L_1-based, L_2-based, L_∞-based mm-GNATs and GNATs for L_p norm $(p = 1, \ldots, 10, \infty)$ were constructed. We computed $\overline{\mathrm{diff}}_p$, $\overline{\mathrm{diff}}_{\mathrm{mm}}$ from the existence ranges of GNAT, mm-GNAT, respectively. The values of $\overline{\mathrm{diff}}_p$, $\overline{\mathrm{diff}}_{\mathrm{mm}}$ are shown in Table 2 (right).

[Search experiment] To investigate the search performance based on standard method and mm-GNAT, we checked relation between *selectivity* S and the total number N of distance calculations, where selectivity S is the ratio of the correct number to the total number of data points. The selectivity vary with the search radius ε and search norm L_p norm. For example, when selectivity is about 0.03, the ε is equal to 0.513 for L_1 norm and to 0.308 for L_2 norm, inversely, when ε is about 0.3, selectivity is 0.004 for L_1 norm and 0.03 for L_2 norm[⋆2]. The total number N of distance calculations is the sum of the number of distance calculations to obtain all correct answers for a search. The mathematical expression (3) approximates this number.

⋆2 The number of distance calculations is about 10,000 when selectivity is 0.004 for L_1 norm and 0.03 for L_2 norm in Fig. 4 and Fig. 5, respectively.

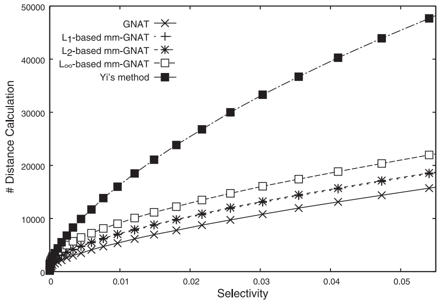

Fig. 4 Selectivity versus number of distance calculations (search by L_1 norm, DB_1).

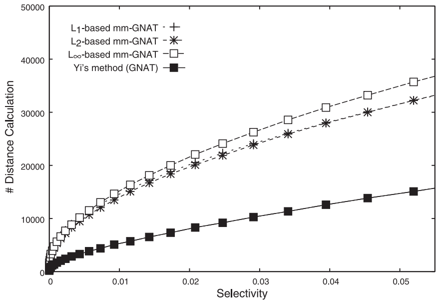

Fig. 5 Selectivity versus number of distance calculations (search by L_2 norm, DB_1).

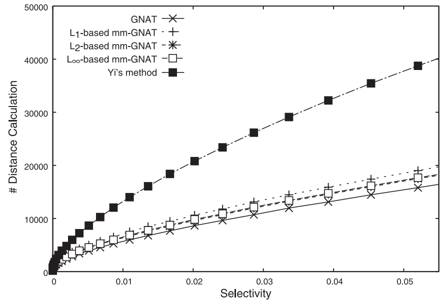

Fig. 6 Selectivity versus number of distance calculations (search by L_∞ norm, DB_1).

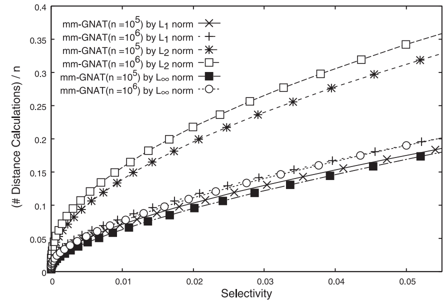

Fig. 7 Selectivity versus the ratio of the number of distance calculations to the number of data n on L_2-based mm-GNATs.

The details of search experiment were as follows. A query point $\boldsymbol{q}$ was randomly selected from the data set. For $\boldsymbol{q}$, we executed an ε-similarity search and counted the correct number $C_{\boldsymbol{q},\varepsilon}$ and the total number $N_{\boldsymbol{q},\varepsilon}$ of distance calculations. We repeated this operation 1000 times, and computed the average correct number C_ε over correct numbers $C_{\boldsymbol{q},\varepsilon}$ for ε-similarity searches. The average number N_ε over $N_{\boldsymbol{q},\varepsilon}$ was also computed.

With changing ε, we computed the average selectivity and the average number of distance calculations with 3 data sets DB_1, DB_2, DB_3 and standard method, L_1-based, L_2-based and L_∞-based mm-GNATs with searching by L_p norm ($p = 1, 2, \ldots, 10, \infty$). Due to space limitations, we show the results of ε-similarity search on 4 dimensional artificial data (DB_1) with searching by L_1 norm, L_2 norm and L_∞ norm in **Fig. 4**, **Fig. 5** and **Fig. 6**, respectively. The graph of L_1-based mm-GNAT (+) is similar to that of L_2-based ($\ast$) in Fig. 4 and Fig. 5, the graph of L_1-based overlaps that of L_2-based.

We also executed search experiments for large artificial data set of a million points of 4 dimensional data. We took 1,000 separate points in each experiments. **Figure 7** shows results for L_2-based mm-GNATs, whose horizontal axis is selectivity and vertical axis is the ratio of the number of distance calculations to the number of data (10^5 or 10^6).

The results show that standard method (GNAT) has the smallest average number of distance calculations. It is easily expected, since the construction norm

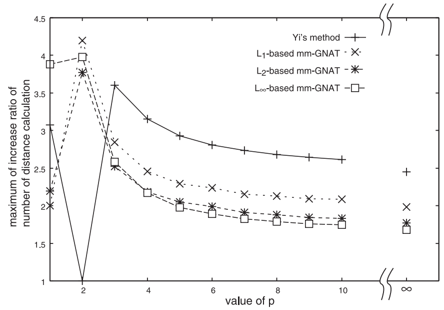

Fig. 8 Relation between search norm and the maximum increase ratio for DB_1.

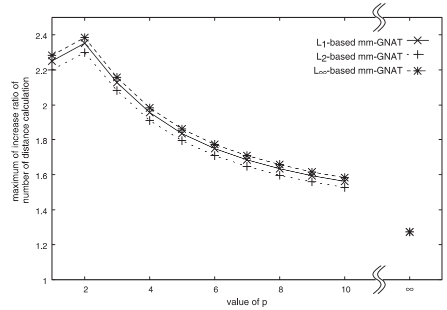

Fig. 9 Expected ratio obtained by substituting the values of DB_1 in Table 2 (right) into parameters (4).

and search norm are the same norm in the standard method. We investigate the ratio of the average number of distance calculations on mm-GNAT to that on standard method. This ratio indicates the performance of search by mm-GNAT and is called the *increase ratio*. To investigate a relation between search norm and the increase ratio, we computed the maximum of increase ratio per search norm (L_p norm) among selectivities for each L_q-based mm-GNAT ($q = 1, 2, \infty$). We summarized the results in **Fig. 8**.

5. Discussion

We discuss the following points:

- experimental confirmation of Theorem 2;
- relation between selectivity and the number of distance calculations;
- comparison with Yi's method;
- effect of construction norm on search performance;
- scalability of mm-GNAT;
- comparison with theoretical analysis.

[**Experimental confirmation of Theorem 2**] The set of the correct points of ε-similarity search on mm-GNAT was exactly the same as that by standard method. Theorem 2 is experimentally confirmed from the results.

[**Relation between selectivity and the number of distance calculations**] We discuss the results for artificial data DB_1, DB_2 and DB_3. Each of data sets has 4, 8, 16 dimension, respectively.

The search experiments for DB_1 are shown in Fig. 4, Fig. 5 and Fig. 6. The numbers of distance calculations of mm-GNATs in Fig. 4 and Fig. 6 are less than about 20,000 and those in Fig. 5 are less than about 40,000. For DB_1, the numbers of distance calculations of mm-GNAT are smaller than that of exhaustive search, which is equal to 100,000.

The volume of ε-ball grows very rapidly as its dimension increases. Therefore, we have to do exhaustive search even if selectivity is small. For DB_2 (8 dimension), pruning unnecessary clusters based on mm-GNAT was effective except for a search by L_2 norm. The search by L_2 norm was much the same thing as exhaustive search even for small selectivity. For DB_3 (16 dimension), the search of any L_p norm was exhaustive search. This phenomenon was also found on standard method. For these data, the number of distance calculations increased with an increase in selectivity and was almost always larger than the number of the data points. Because almost all clusters were regarded as necessary, the distance calculation between the query point and all data points were needed. So, the search became exhaustive search. In addition, the distance calculations for the pruning were also needed.

[**Comparison with Yi's method**] Figures 4 and 6 show that the number of distance computations on Yi's method is larger than those on GNAT and L_q-based mm-GNATs for the search by L_1 norm and L_∞ norm. Thus, L_q-based mm-GNATs ($q = 1, 2, \infty$) has good search performance in the search by L_1 norm and L_∞ norm for DB_1. This is the same for DB_2, DB_3 and DB_4 (Fig. 10 and Fig. 12).

For the search by L_2 norm (Fig. 5), Yi's method has the best performance among all methods. The retrieval by L_2 norm in Yi's method is the same as that on GNAT. The comparison between GNAT and mm-GNAT is already shown in [Search experiment].

Figure 8 shows that the graphs of maximum increase ratio of number of distance calculations for search by L_p norm ($p = 1, 2, \ldots, 10, \infty$) on L_q-based mm-GNAT ($q = 1, 2, \infty$) and Yi's method. In the figure, each mm-GNAT has smaller search cost than Yi's method for L_p search norm ($p = 1, 3, \ldots, 10, \infty$). From the viewpoint of multi-modality support for various L_p norm (except for L_2 norm), mm-GNAT is better than Yi's method in our computational experiment.

[**Effect of construction norm on search performance**] We consider which L_q-based mm-GNAT has good search performance. We look into Fig. 4, Fig. 5 and Fig. 6. Figure 4 shows the result of search by L_1 norm. In the figure, the number of distance calculations on L_1-based mm-GNAT is smaller than those on L_2-based and L_∞-based mm-GNATs. Figure 5 shows the result of search by L_2 norm. The number of distance calculations on L_2-based is smaller than those on L_1-based and L_∞-norm mm-GNATs. Figure 6 shows the results of search by L_∞ norm. The number of distance calculations on L_∞-based mm-GNAT is smaller than those on L_2-based and L_1-based mm-GNATs. These results show that the number of distance calculations is smallest when search norm is the same as construction norm. Otherwise, the number of distance calculations increases on mm-GNAT. In the case of the same norms being used, the pruning of unnecessary clusters works best, but in the other case, some unnecessary clusters are regarded as necessary, therefore, the number of distance calculations increases.

In Fig. 8, L_∞-based mm-GNAT has the best search performance among other mm-GNATs for search by L_p norm ($p = 4, 5, \ldots, 10, \infty$), but L_1-based mm-GNAT has best for search by L_1 norm. This case can be explained as follows. The search performance is best when search norm is construction norm. When the subscript q of the construction norm (L_q norm) is near to that of the search norm (L_p norm), the search performance of L_p norm is better rather than other mm-GNATs. Thus L_1-based, L_2-based and L_∞-based mm-GNATs have best search performance for L_1, $L_p(p = 2, 3)$ and $L_p(p = 4, 5, \ldots, 10, \infty)$ search norms, respectively.

[**Scalability of mm-GNAT**] We executed search experiments for two 4 dimensional artificial data set DB_1 (10^5 data points) or DB_5 (10^6 data points). The results are shown in Fig. 7. Three pairs of curves are shown in the figure ($\{\times, +\}$, $\{\ast, \square\}$ and $\{\blacksquare, \bigcirc\}$). The pair ($\{\ast, \square\}$) are searches by L_2 norm. The maximum of increasing ratio ($\square$) of search by L_2 norm for DB_5 is only 10% larger than that ($\ast$) for DB_1, while the size of DB_5 is 10 times larger than that of DB_1. The similar relation is found in the searches by L_1 norm ($\{\times, +\}$) and by L_∞ norm ($\{\blacksquare, \bigcirc\}$).

[**Comparison with theoretical analysis**] We compare the expected ratio (4) of the number of distance calculations with that in Fig. 8. The expected ratio (4) depends on $\overline{\mathrm{diff}}_{\mathrm{mm}}$, $\overline{\mathrm{diff}}_p$, which are determined by index structure, the number of separate points k, dimension d and the search norm. Substituting p, d and the values of $\overline{\mathrm{diff}}_{\mathrm{mm}}$, $\overline{\mathrm{diff}}_p$ in Table 2 (right) to (4), we have a graph with the same vertical and horizontal axes as those of Fig. 8. We have a graph of expected ratio (4) shown in **Fig. 9** by substituting $d = 4$, $k = 1$ and the values of $\overline{\mathrm{diff}}_{\mathrm{mm}}$, $\overline{\mathrm{diff}}_p$ for DB_1 in Table 2 (right). The graph of Fig. 9 has a peak at $p = 2$ and a shape similar to that of Fig. 8. Since each value of (4) is positive and equal to the corresponding the value of Fig. 9 to the power of k, the graph of expected ratio (4) has the shape similar to that of Fig. 9. Thus it is shown that the behavior of the ratio in Fig. 8 is approximated by the expected ratio (4).

We also focus attention on search norm of expected ratio (4). Suppose construction norm is fixed. The $\overline{\mathrm{diff}}_{\mathrm{mm}}$ and the dimension d are constant, then the value of (4) depends on $d^{1/p}/\overline{\mathrm{diff}}_p$. The enumerator $d^{1/p}$ monotonically decreases, for example, for $d = 4$ the $d^{1/p}$ decreases from 4 to 1 when $p = 1, \ldots \infty$. The denominator $\overline{\mathrm{diff}}_p$ decrease from 0.463 to 0.221 when $p = 1, 2$ and the values for $p = 3, \ldots, 10, \infty$ are contained between 0.188 to 0.200. So, the denominator is regarded as constant for $p = 3, \ldots, 10, \infty$. Thus the value of (4) depends on only $d^{1/p}$ for $p = 3, \ldots, 10, \infty$.

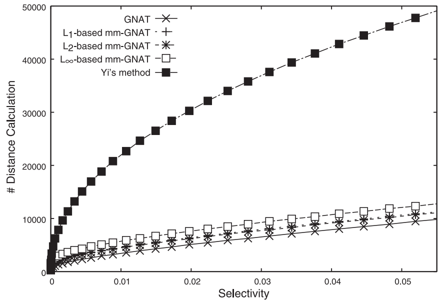

Fig. 10 Selectivity versus number of distance calculations (search by L_1 norm, music data).

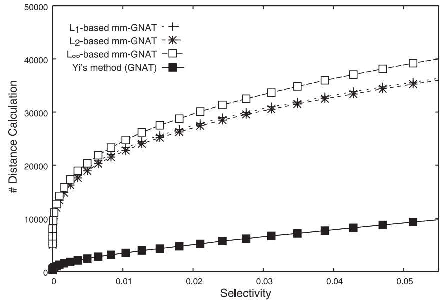

Fig. 11 Selectivity versus number of distance calculations (search by L_2 norm, music data).

6. Application to Music Data

Retrieval of music data is a hot topic [13]. We have proposed features for retrieval of music data [14),15)].

In this section we describe search experiment on music data set. We prepared 1,023 pieces of music from 89 CDs and then applied TwinVQ encoder to each piece of music. In the encoding step of TwinVQ, we extracted an autocorrelation coefficient vector $\boldsymbol{r}_{u,m} = (r_{u,m,1}, \ldots, r_{u,m,20})$ of the m-th frame of the u-th piece of music. Out of the extracted autocorrelation coefficient vectors, 100,000 autocorrelation coefficient vectors were randomly selected. We call the set of the selected vectors "DB_4" in Table 1. We computed the number of distance calculations for ε-similarity search by the same method applied to artificial data. The results are shown for L_1-based, L_2-based and L_∞-based mm-GNATs in **Fig. 10**, **Fig. 11** and **Fig. 12**, respectively. Axes of figures are the same as those of Fig. 4. The graph of L_1-based mm-GNAT overlaps that of L_2-based in Fig. 10, Fig. 11 and Fig. 12. All graphs in Fig. 12 except for Yi's method overlap each other.

The results of search experiments are similar to those for DB_1 (see Fig. 4, Fig. 5 and Fig. 6). Principal component analysis for DB_4 showed that the cumulative contribution ratio is 99.07% (up to 4th axis) and 99.52% (up to 5th axis). This implies that the music data (DB_4) can be regarded as 4 dimensional data. This experiment suggests that mm-GNAT works well for high dimensional data if the

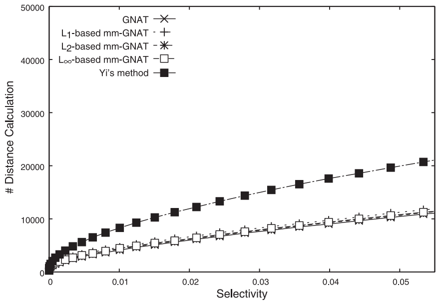

Fig. 12 Selectivity versus number of distance calculations (search by L_∞ norm, music data).

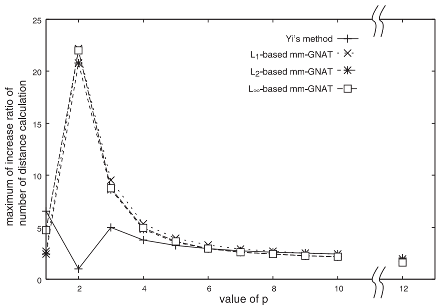

Fig. 13 Relation between search norm and the maximum increase ratio for music data.

data are highly correlated, as is often the case with real data.

To investigate a relation between search norm and the increase ratio, we also computed the maximum of increase ratio for search by L_p norm ($p = 1, 2, \ldots, 10, \infty$) on L_q-based mm-GNAT ($q = 1, 2, \infty$) and on Yi's method and search norm on music data. **Figure 13** shows a graph of the maximum of increase ratio. Note that the shape of the graph is similar to that in Fig. 8. The search costs of mm-GNATs for music data are smaller than Yi's method when $p = 1, 6, \ldots, 10, \infty$ in Fig. 13.

7. Conclusion

In this paper we proposed a new multi-modaility support index structure, *mm-GNAT*, for ε-similarity search by arbitarary L_p norm. mm-GNAT is realized with mm-range on GNAT and without extending search radius. The index structure is designed using the following inequalities

$$\mathrm{dist}_\infty(\boldsymbol{x}, \boldsymbol{y}) \le \mathrm{dist}_p(\boldsymbol{x}, \boldsymbol{y}) \le \mathrm{dist}_1(\boldsymbol{x}, \boldsymbol{y}) \quad \forall \boldsymbol{x}, \boldsymbol{y},\ p = 1, 2, \ldots, \infty.$$

From this relation the existence range for any search norm is always included in the mm-range. Therefore, search of arbitrary L_p norm is realized by using mm-range.

We implemented index structures GNATs, mm-GNATs for several search norms L_p norm ($p = 1, 2, \ldots, 10, \infty$) and executed experiments of ε-similarity search on the index structures. We confirmed that ε-similarity search is correctly executed on the mm-GNATs. We compared mm-GNAT with Yi's method. mm-GNAT has better performance on retrieval by L_p norm ($p = 1, 3, \ldots, \infty$), while Yi's method has better performance on retrieval by L_2 norm in our experiments.

The number of distance calculations was reduced by using mm-GNAT as index structure rather than exhaustive search for uniform 4 dimensional data set DB_1. The numbers of distance calculation on GNAT and mm-GNAT increase rapidly for the data sets DB_2, DB_3 (uniform 8, 16 dimensional data, respectively). GNAT and mm-GNAT need to be improved for high dimensional uniform data set. When we executed search experiments on L_q-based mm-GNATs by L_p norm ($p = 1, 2, \ldots, 10, \infty$), the search by L_q norm (*i.e.*, $L_p = L_q$) has the best performance.

Moreover, we executed experiments on large scale data set, which has 1,000,000 data points. The search performance is about 1.1 times on the data set while the size of the data set is 10 times.

We also performed search experiment on music data. The search performance was similar to that for 4 dimensional artificial data. mm-GNAT was useful for efficient retrieval of music data in MPEG-4/TwinVQ domain.

Acknowledgments This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research, 22500097, 2010.

References

1) Guttman, A.: R-trees: A dynamic index structure for spatial searching, *SIGMOD '84: Proc. 1984 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, pp.47–57, ACM (1984).

2) White, D. A. and Jain, R.: Similarity Indexing with the SS-tree, *ICDE '96: Proc. 12th International Conference on Data Engineering*, Washington, DC, USA, pp.516–523, IEEE Computer Society (1996).

3) Katayama, N. and Satoh, S.: The SR-tree: An index structure for high-dimensional nearest neighbor queries, *SIGMOD '97: Proc. 1997 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, pp.369–380, ACM (1997).

4) Yianilos, P. N.: Data structures and algorithms for nearest neighbor search in general metric spaces, *SODA '93: Proc. 4th Annual ACM-SIAM Symposium on Discrete Algorithms*, Philadelphia, PA, USA, pp.311–321, Society for Industrial

and Applied Mathematics (1993).
5) Ciaccia, P., Patella, M. and Zezula, P.: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces, *VLDB*, Jarke, M., Carey, M.J., Dittrich, K.R., Lochovsky, F.H., Loucopoulos, P. and Jeusfeld, M.A. (Eds.), pp.426–435, Morgan Kaufmann (1997).
6) Brin, S.: Near Neighbor Search in Large Metric Spaces, *VLDB*, Dayal, U., Gray, P.M.D. and Nishio, S. (Eds.), pp.574–584, Morgan Kaufmann (1995).
7) Böhm, C., Berchtold, S. and Keim, D.A.: Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases, *ACM Comput. Surv.*, Vol.33, No.3, pp.322–373 (2001).
8) Chávez, E., Navarro, G., Baeza-Yates, R. and Marroquín, J.L.: Searching in metric spaces, *ACM Comput. Surv.*, Vol.33, No.3, pp.273–321 (2001).
9) Yi, B.-K. and Faloutsos, C.: Fast Time Sequence Indexing for Arbitrary L_p Norms, *VLDB*, Abbadi, A.E., Brodie, M.L., Chakravarthy, S., Dayal, U., Kamel, N., Schlageter, G. and Whang, K.-Y. (Eds.), pp.385–394, Morgan Kaufmann (2000).
10) Lee, S., Kwon, D. and Lee, S.: Minimum Distance Queries for Time Series Data, *Journal of Systems and Software*, Vol.69, No.1-2, pp.105–113 (2004).
11) Ciaccia, P. and Patella, M.: Searching in metric spaces with user-defined and approximate distances, *ACM Trans. Database Syst.*, Vol.27, No.4, pp.398–437 (2002).
12) Kimura, A., Onishi, K., Kobayakawa, M., Hoshi, M. and Ohmori, T.: Distance Conversion Rule for Arbitrary L_p distance, *IPSJ Trans. on Databases*, Vol.46, No.SIG 8 (TOD 26), pp.93–105 (2005).
13) Pardo, B.: Introduction, *Comm. ACM*, Vol.49, No.8, pp.28–31 (2006).
14) Kobayakawa, M., Hoshi, M. and Onishi, K.: A method for retrieving music data with different bit rates using MPEG-4 TwinVQ audio compression, *MULTIMEDIA '05: Proc. 13th Annual ACM International Conference on Multimedia*, New York, NY, USA, pp.459–462, ACM (2005).
15) Onishi, K., Kobayakawa, M., Hoshi, M. and Ohmori, T.: A Feature Independent of Bit Rate for TwinVQ Audio Retrieval, *ICME 2001: Proc. IEEE International Conference on Multimedia and Expo*, Los Alamitos, CA, USA, pp.409–412, IEEE Computer Society (2001).

(Received March 19, 2010)
(Accepted July 6, 2010)

(Editor in Charge: *Kunihiko Sadakane*)

Kensuke Onishi received his Doctor degree in science from the University of Tokyo, Tokyo, Japan in 1999. From 1999 to 2004, he was with Graduate School of Information Systems, University of Electro-Communications. He is a lecturer of Department of Mathematical Sciences, Tokai University. His current research interests include data structure and algorithm, especially computational geometry and index structure of database, graph algorithm and discrete mathematics. He is a member of ACM, IEEE CS and IPSJ.

Michihiro Kobayakawa received his Dr.E. degree from University of Electro-Communications in 2001. From 2001 to 2008, he was on Graduate School of Information Systems, University of Electro-Communications. From 2008 to 2009, he was an Associate Professor of Okinawa National College of Technology. He is an Associate Professor of Tokyo Metropolitan College of Industrial Technology. His interests include a content-based multimedia retrieval, and algorithms and data structures for multimedia data retrieval. He is a member of ACM, IEEE CS and IEICE.

Mamoru Hoshi received his Dr.E. degree in mathematical engineering from the University of Tokyo in 1985. He is an Emeritus Professor of University of Electro-Communications. His research interests include algorithms and data structures for searching, multimedia retrieval, and random number generation. He is a member of ACM, IEEE CS, IEEE IT, IEICE, IPSJ, and SITA.