Search in a hybrid P2P system using a node map formed by a self-organising map

Kei Ohnishi*

Kyushu Institute of Technology,
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
Fax: +81-948-29-7660
Email: ohnishi@cse.kyutech.ac.jp
*Corresponding author

Hiroshi Yamamoto

Ritsumeikan University,
1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577 Japan
Email: hiroyama@fc.ritsumei.ac.jp

Abstract: We propose a hybrid P2P information search system that combines: 1) the idea that a searcher searches not information itself but people who have KANSEI (human sensitivity) well fitted to the searcher; 2) a P2P network on which we can build a scalable information search system; 3) a folksonomy, which is a system that allows users to classify information by themselves, to quickly and reliably obtain the desired information from on the network. The main characteristic of the proposed system is that it produces a two-dimensional map visualising the KANSEI information of all nodes by a self-organising map, and then utilises the map for searches. The simulation results show that the proposed system provides balanced opportunities from which we obtain our desired information and other unknown information, in comparison with a search method that conducts searches directly in higher-dimensional space and a random search method.

Keywords: hybrid P2P system; search; visualisation; self-organising map; SOM; KANSEI information.

Biographical notes: Kei Ohnishi received his BE, ME and DE degrees from Kyushu Institute of Design, Japan in 1998, 2000, and 2003, respectively. He worked as a Postdoctoral Researcher for University of Illinois at Urbana-Champaign, Kyushu Institute of Technology, and Human Media Creation Center/Kyushu. Since October 2007, he has been an Associate Professor at Kyushu Institute of Technology. His research interests include bio-inspired computing and networking techniques. He is a member of IEEE, IEICE, and SOFT.

Hiroshi Yamamoto received his ME and DE degrees from Kyushu Institute of Technology, Iizuka, Japan in 2003 and 2006, respectively. From April 2006 to March 2010, he worked at Fujitsu Laboratories Ltd., Kawasaki, Japan. From April 2010 to March 2015, he was an Assistant Professor in the Department of

Electrical Engineering, Nagaoka University of Technology. Since April 2015, he has been an Associate Professor in the College of Information Science and Engineering, Ritsumeikan University, Japan. His research interests include computer networks, distributed applications, and sensor networks. He is a member of the IEEE.

1 Introduction

The recent rapid growth of information technology has caused a continuous increase in the amount of available information. This situation is called the information explosion. In this situation, it is difficult for people to find their desired information efficiently and reliably. One approach is that people find their desired information with high quality from the large and diversified amount of information by first finding others who have the human sensitivity (called *KANSEI*) in which they are interested. Then, these people can utilise information that others hold and recommend.

Unlike the client-server network, no node in a peer-to-peer (P2P) network has a fixed role, but rather each node can be both client and server (Lua et al., 2005). Due to this feature, a P2P network is often used as an infrastructure for realising distributed services in which nodes mutually provide services. For example, an information search system built on this infrastructure cannot guarantee a reliable search or a quick search but can be expected to have high scalability to deal with the increasing amount of information and the increasing number of users.

As for a way to classify information that has been increasing in amount, not only centralised classification by system operators, as in conventional web information search systems, but also distributed classification by the users of an information search system has appeared. A web search system that allows users themselves to classify information is called a folksonomy (Mathes, 2004). In a folksonomy, users assign different tags to identical information according to their own KANSEI (human sensitivity). The tag assigned to some information is equivalent to an identifier of the information and can be used for an information search. Therefore, when a searcher looks for particular information, if another person with a similar sensitivity to the searcher has already assigned a tag (for example, a keyword) that the searcher uses as a search key to the particular information, the searcher can quickly find his/her desired information by finding that other person. Thus, if people with similar KANSEI to a searcher use the information search system (folksonomy), the searcher can expect an efficient information search. Furthermore, this approach is theoretically available not only in web search systems as a client-server system but in P2P systems as well.

In this paper, we propose a new information search system that combines the following three components:

1 the concept that a searcher looks not for information itself but for people who have KANSEI matching the searcher's

2   a P2P network that is expected to build highly scalable information search systems to deal with the explosive increase in the amount of information

3   a folksonomy, in which users themselves classify information to be searched.

Specifically, we propose a hybrid P2P information search system that enables us to find persons who have KANSEI matching ours. The salient feature of the proposed system is that we can utilise a map that two-dimensionally visualises KANSEI information of all nodes, that is, the entire network. A hybrid P2P information search system is not a completely distributed system. The system consists of P2P nodes that actually hold information and mutually exchange information between them and a server that manages only the locations of information. When a P2P node searches for particular information, the node asks the server the location of the desired information.

This paper is organised as follows. Section 2 describes related work. In Section 3, we explain the proposed hybrid P2P information search system. Section 4 shows simulation results for evaluating the proposed system. In Section 5, we draw our conclusions and state future work.

2   Related work

The features of the system proposed in the paper are as follows: to utilise visualised network information for searching files, nodes, and so on; to utilise KANSEI information for information searches; and to allow users to freely assign search tags to information, as in a folksonomy. We describe work related to these features in this section.

Network visualisation usually means visualisation of the graph structure of a network. Many tools for such visualisation have been developed, but direct application of such visualisation to information searches has not yet been reported. Visualisation of hyperlinks on the web is not direct utilisation of network visualisation for information searches, but would be useful for web searches. Some studies focus not on visualisation of the linkage between nodes but on visualisation of the relative distances between nodes with different characteristics. For example, a system proposed in Dabek et al. (2004) first calculates the distances between nodes based on the costs of transmission delay and then arranges the nodes in a two-dimensional plane while maintaining the distance relationships of the nodes in the original space as precisely as possible. We consider various uses for this two-dimensional network map, such as understanding the network situation visually and determining routes in the network. However, they are not direct applications for an information search.

In Hayashida and Takagi (2000), search points for an optimisation problem are mapped into two-dimensional space by, for example, a self-organising map (SOM) and visualised two-dimensionally, and then humans generate new search points by utilising the visualised search points. Two-dimensional visualisation of the distribution of solution candidates in a search process for an optimisation problem is also used for understanding the search process and then for developing new search methods (Yamashiro et al., 2006). These two studies (Hayashida and Takagi, 2000; Yamashiro et al., 2006) are not similar to the present study in terms of purpose but are similar to it in terms of procedures.

With respect to information search techniques, the simplest method is to search for objects by explicitly expressing target objects such as file names. However, in order to express the target objects precisely, we need to know what to search for in advance. Such a situation is rare. Therefore, searching for objects from pieces of information included in target objects has been considered. The present web search techniques adopt such a strategy. Pieces of information included in target objects are not target objects themselves, but can be thought of as alternative forms of search tags to search for target objects. Furthermore, as a variation of the pieces of information included in target objects, we can use the features extracted by processing the target objects. P2P systems that have a search method using such search tags, sometimes called metadata, have been extensively investigated (Joseph and Hoshiai, 2003). Some of these systems consider a method for forwarding a search query with the same search tag form to nodes that are likely to have many search objects, such as files with the same search tags. Especially, some P2P systems determine search tags of documents by focusing on semantics of the documents, which is meanings of words, phrases, and sentences in the documents. Then, query forwarding relying on such search tags is referred to as 'semantic routing'. In addition, P2P systems that are organised for the semantic routing to work efficiently or effectively are generally referred to as 'semantic P2P systems'. There have been done many studies on semantic P2P systems and semantic routing so far (Joseph, 2002; Tang et al., 2003; Crespo and Garcia-Molina, 2005; Rostami et al., 2008; Jung, 2010; Mawlood-Yunis et al., 2010; Yeferny et al,. 2013; Himali et al., 20013). Furthermore, a semantic P2P system has been used as a document recommender system, in which documents are recommended to a user according to the user's interests (Vera-del Campo et al., 2012). Our proposed system does not rely on semantics of shared information, but works as a sort of recommender system as well. In our proposed system, a searcher finds a node with interesting KANSEI to obtain recommendations from the found node.

Besides using pieces of information included in target objects, one of the possible forms of search tags that people can assign to search objects is *KANSEI* information, which in Japanese means 'human sensitivity' information and here means how people feel about search objects. Such information search systems are referred to as *KANSEI* information search systems, and several concrete *KANSEI* information search systems exist (Yoshida et al., 1999; Takagi and Noda, 2004; Sato et al., 2007; Yali and Kui, 2008; Miyakawa et al., 2008; Feng and Uchida, 2010; Huang et al., 2011). A key idea of the KANSEI information search system is that a searcher does not explicitly give pieces of information included in search objects, as in a conventional keyword searches, but expresses what he/she wants to subjectively feel about the search objects as the search query. However, the conventional KANSEI information search system requires the correspondences between KANSEI information as search tags and search objects to be established in advance. A recent study has been researching a P2P system that enables us to search objects by using KANSEI information but does not require correspondences between KANSEI information and search objects to be already established (Ohnishi et al., 2009).

The above-mentioned P2P systems have search methods for obtaining desired information or finding nodes holding desired information. In addition, a P2P system sometimes requires a mechanism for finding a trustable and secured node from among multiple nodes that can provide the same services in order to realise a trustable and

secured P2P system. A P2P reputation system is one of such mechanisms and have been studied so far (Gupta et al., 2003; Hao et al., 2009; Pogkas et al., 2009; Azzedin, 2014). In a P2P reputation system, reputation scores of nodes are managed and used for node selection.

In conventional web information retrieval systems, search tags to shared information are assigned in fixed ways, which is a top-down approach. Meanwhile, users have recently become interested in information retrieval systems that allow the users themselves to assign search tags to shared information. Such a bottom-up information retrieval system is referred to as a *folksonomy* (Mathes, 2004). At this moment, several folksonomies are running and most of them allow users to freely assign keywords to shared information (Delicious, http://delicious.com/; Flickr, http://www.flickr.com/). We recently conducted a study for realising the application of folksonomies in a P2P network (Ohnishi et al., 2009).

3   Proposed system

We explain the outline of our proposed system here. The proposed system is an information search system built on a hybrid P2P network. In a hybrid P2P system, a node holds information to be searched and a server, called an 'index server', manages the locations of this information; that is, the server manages which nodes have what information in a centralised manner. The flow of the proposed system is as follows (see Figure 1):

1   each P2P node makes a representation of its KANSEI and sends the representation to an index server

2   the index server generates a map of the KANSEI of all nodes by a SOM (Kohonen, 1990) and distributes the map of the KANSEI to each P2P node

3   every P2P node conducts a search by using the map of the KANSEI during a fixed period of time

4   the procedure begins again at step (1).

When a searcher conducts a search in a general hybrid P2P system, he/she first asks the index server which node holds the desired information and then directly communicates with a node holding the desired information for delivery of that information. In contrast, the index server in the proposed system manages the KANSEI representation of each node, not the locations of information. So, when a searcher conducts a search in the proposed system, he/she first asks the index server which node has the particular KANSEI. The proposed system enables users to find nodes by using two-dimensional visualisation of nodes with various KANSEI representations, that is, by visualisation of the entire P2P network.

We explain the components of the proposed system in detail in the next section.

Figure 1 Flow of the proposed hybrid P2P information search system (see online version for colours)
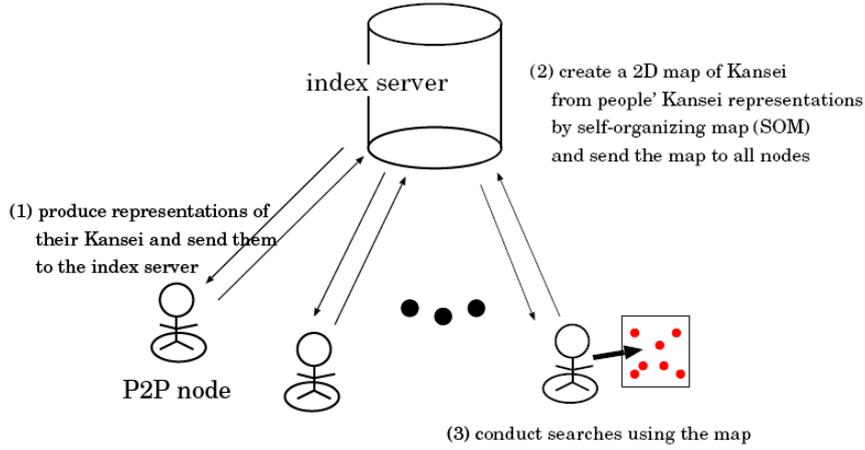

3.1 Method for visualising networks

To two-dimensionally visualise the KANSEI of participating nodes in the system, each node first makes a representation of its KANSEI. Then, the KANSEI information obtained from all nodes is mapped into two-dimensional space and visualised. In the following, we consider some files as the search objects.

Each node expresses its KANSEI as follows. First, each node selects M files that well express its individuality from all of its files. For example, when some node holds many music files, it selects M music files that well represent its taste. Then, each node assigns an S-dimensional numerical vector representing the impression it obtains from each of the M files to each file. For example, it is possible to assign a numerical vector to a file by the semantic differential (SD) method, which evaluates a target as a numerical score (position) on a scale between a pair of bipolar adjectives, for example, 'dark-light'. Then, each node calculates and normalises the sum of all the S-dimensional numerical vectors assigned to the M files, and the normalised S-dimensional numerical vector becomes the representation of the node's KANSEI.

In the simulations described in Section 4, each node selects five files ($M = 5$) and assigns evaluations to them by the SD method, in which four pairs of bipolar adjectives are used and the number of steps in the scale between each pair of bipolar adjectives is five, that is, –2, –1, 0, +1, +2. Suppose that a node assigns the following numerical vectors to its selected five files: (–1, –1, 0, +2), (–2, –1, +1, +2), (–1, –1, +1, +1), (–2, –1, 0, +2), (–1, –1, –1, +2). Then, transforming an element of the vectors such as –2 to 0, –1 to 0.25, 0 to 0.5, +1 to 0.75, and +2 to 1 to normalise the five vectors, the five vectors become as follows: (0.25, 0.25, 0.5,1), (0, 0.25, 0.75, 1), (0.25, 0.25, 0.75, 0.75), (0, 0.25, 0.5, 1), (0.25, 0.25, 0.25, 1). Then, the sum of those transformed five vectors is calculated and results in (0.75, 1.25, 2.75, 4.75). Finally, each element value of the vector is divided by $M(=5)$ and (0.15, 0.25, 0.55, 0.95) is obtained. This normalised vector is the KANSEI representation of the node. The process of representing the KANSEI of a node is shown in Figure 2.

After all nodes generate *S*-dimensional numerical vectors representing their KANSEI, they send their vectors representing their KANSEI to the index server as a component of the proposed system. Then, the vectors are inputted into the SOM and a map visualising all nodes' KANSEI information is obtained. SOM is one type of neural network that models the information processing by the brain (neurons) and generally maps high-dimensional numerical data that is more than three-dimensional data to two-dimensional space while maintaining the distance relationships of the data in the original space as much as possible.

Figure 2   Representing the KANSEI of a node (see online version for colours)
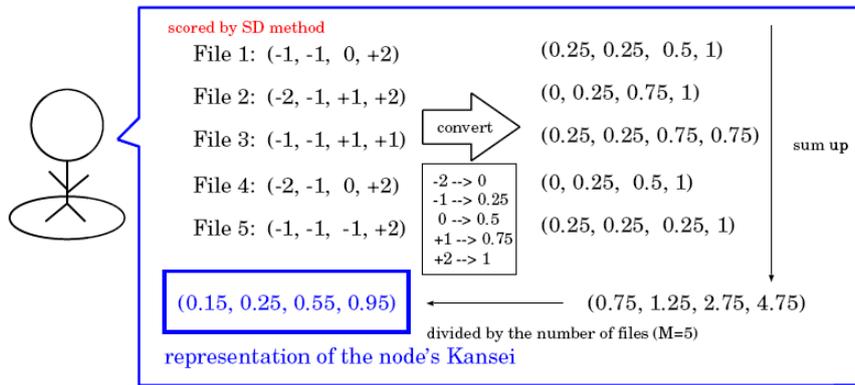
Figure 3   Example of the map representing KANSEI information of all nodes as obtained by the SOM (see online version for colours)
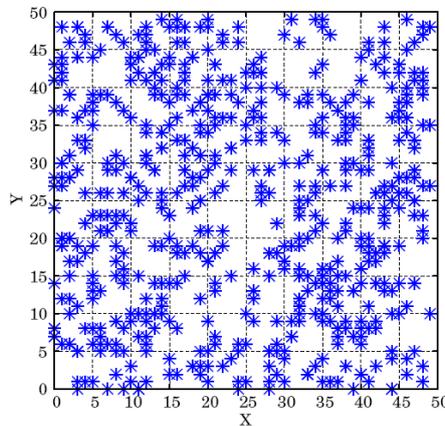
Note: A mark * in the map represents a mapping location for a point in the original four-dimensional space.

A map visualising the KANSEI information of all nodes, which are mapped onto lattice points of the map, is shown in Figure 3 as an $H \times L (H, L \in \mathbb{N})$ square region. In Figure 3, the normalised vector representing the KANSEI of a node is four-dimensional. Some lattice points in the map do not have any mapped KANSEI information of the

represented nodes; these lattice points correspond to neurons in the competitive layer of the SOM and therefore have vectors of weight values whose dimension is four, the same as the vector representing the KANSEI of a node. In addition, it is basically expected that neighbouring points before mapping in the four-dimensional space are close to each other in the map.

3.2 Search using visualised network information

The map visualising the KANSEI information formed by the SOM is distributed to all nodes by the index server. There are two ways to search by using this visualised information.

In the first way, a searcher selects a point in the distributed map and then obtains the contact information (IP address) of the node corresponding to that point or the node closest to that point. Next, file delivery is conducted directly between the searcher and the selected node. It would be useful, in deciding the node to search, if we could refer to the KANSEI information of the node in the map when selecting the node. However, this idea must wait for a real implementation of the proposed system before it can be realised.

Figure 4 Searching for a node by issuing a search query (second way) (see online version for colours)
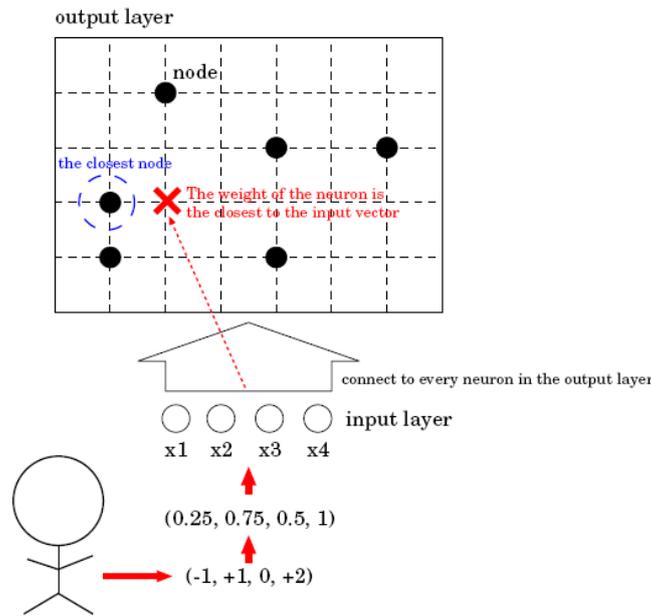
In the second way, a searcher first issues a search query with the same form as the KANSEI information assigned to a file in the proposed system, and the generated search query is mapped onto a point in the map obtained by the SOM. Then, the contact information of the node corresponding to the point or closest to the point is obtained from the index server. We give a concrete example using the SD method described in Section 3.2, where four pairs of bipolar adjectives are used and the number of steps in the

scale between each pair of bipolar adjective is five, that is, –2, –1, 0, +1, +2. Suppose that a searcher makes a search query of (–1, +1, 0, +2). This search query is first normalised as (0.25, 0.75, 0.5, 1) according to transformation rules that transform –2 to 0, –1 to 0.25, 0 to 0.5, +1 to 0.75, and +2 to 1. The normalised search query is then inputted to the map of the KANSEI of all nodes distributed by the index server and equivalent to the particular SOM. Then, a point to which the inputted search query is mapped is obtained by finding a neuron in the competitive layer that has the weight vector closest to the inputted search query in terms of Euclidean distance. Finally, the searcher finds the node corresponding to the point or closest to the point and obtains its contact information (IP address) from the index server. The searcher communicates directly with the found node for file delivery. This procedure is shown in Figure 4. In the simulations described in Section 4, we examine this search.

4   Simulation evaluation

We evaluate the hybrid P2P information search system proposed in Section 3 with simulations presented in this section. One of the salient features of the proposed system is it allows a searcher to determine a search object of a node by selecting a point in the map representing the KANSEI information of all nodes. However, another search way is evaluated through simulations in this section. We need to model the KANSEI of nodes and searchers in the simulations. As mentioned later, we will introduce a simple model for that.

As mentioned in Section 3, the proposed system searches a two-dimensional space whose dimension is smaller than the original space of the KANSEI vector for a node in cases that the searcher selects a point in the map and issues a search query. The SOM is used for mapping points in the original higher-dimensional space to ones in the two-dimensional space, and by this mapping, the distance relationships between the points in the original space are broken at least. Therefore, in the simulations, under the assumption that people with a variety of KANSEI exist, we focus on the difference between search performances when searching a node in the original space and that when searching a node in the two-dimensional space.

The most salient feature of the proposed system is that the proposed system enables its users to search for a node in the first way mentioned in Section 3.2, that is to say, to determine a search object of a node by selecting a point in the map representing the KANSEI information of all nodes. A direct way to evaluate this salient feature is to actually develop the proposed system and conduct subjective tests using the developed system. However, we have not been able to develop the system in this paper. In case of evaluating the proposed system through simulations, we then need a behaviour model on which point a searcher selects in the map. Humans are basically good at finding some tendency included in geometric patterns, so we can expect that in the proposed system, a searcher finds some tendency included in the two-dimensional map and actively determines a node of a search object by considering its tendency. According to this expectation, we can assume particular behaviours of a searcher on which point to select in the map and then introduce the behaviour model into simulations. However, human behaviours represented by this behaviour model are essential for the proposed system and cannot be validated in this paper. We think we should evaluate the salient feature of the proposed system through real experiments in future.

Meanwhile, in simulations described from Section 4.1 onward, we fix a way of selecting a point in the map as issuing a search query. In this way, personality of a searcher on which point to select in the map is excluded. Under the use of this way, we assume a situation that there are differences of KANSEI among a searcher and all nodes and then observe the search performance in the proposed system. We introduce a model to represent the differences of KANSEI among them. We should also validate this model. However, we think the model is appropriate for the simulation study in this paper because the model can express differences of KANSEI among people and also adjust the degree of difference, so it can simulate real situations that people have different KANSEI somehow.

4.1 Simulation model

The simulation model used here consists of one part related to network composition and another part related to the KANSEI of searchers and nodes. We explain each part of the model below.

4.1.1 Network composition

The network consists of one index server and 625 nodes. In the network, 625 types of files are distributed for sharing. In the simulation scenarios, the types of files and their locations on particular nodes are described later. Each node selects five types of files ($M = 5$) for representing its KANSEI information.

4.1.2 KANSEI of searchers and nodes

In a real use of the proposed system, the KANSEI of nodes is represented by vectorised KANSEI information assigned to the selected M files. That is enough to implement the system. However, to execute a simulation model for validating the proposed system, not only the KANSEI information of each node for the selected M files but also the KANSEI information of each node and searcher for all 625 types of files distributed over the network has to be modelled, because the searcher has to determine the KANSEI information of some of the files that he/she obtained by one search in the simulation model. KANSEI information is assigned to a file in the way described in Section 3.1, in which four pairs of bipolar adjectives are used and the number of steps in the scale between each pair of bipolar adjective is five, that is, –2, –1, 0, +1, +2. The difference in KANSEI information between the searcher and the nodes is estimated by determining the KANSEI information of the nodes on the basis of probabilistically changing the KANSEI information of the searcher. Regarding the probability of changing the basic KANSEI information (hereinafter called 'change probability of KANSEI information'), a higher probability corresponds to a larger difference in KANSEI information.

The KANSEI information of the searcher as the basis for others' is obtained in the following way: first, assigning serial numbers from 0 through 624 to the 625 types of files; second, transforming the serial numbers to quinary numbers; third, regarding each digit of the quinary number as an element value of the four-dimensional vector representing the KANSEI information. In other words, it is assumed that the searcher assigns different KANSEI information to each type of file. A rule for transforming the serial number p to a quinary number '$f_1f_2f_3f_4$' ($f_1, f_2, f_3, f_4 \in [-2, +2] \in \mathbb{Z}$) is as follows,

where the corresponding KANSEI information of a four-dimensional vector is (f_1, f_2, f_3, f_4):

$$f_1 = \lfloor p/125 \rfloor - 2,$$ (1)

$$f_2 = \lfloor (p \bmod 125)/25 \rfloor - 2,$$ (2)

$$f_3 = \lfloor (p \bmod 25)/25 \rfloor - 2,$$ (3)

$$f_4 = (p \bmod 5) - 2.$$ (4)

For example, when $p = 323$, the KANSEI information of the searcher for the file with this serial number is $(0, 0, +2, +1)$, because $f_1 = 0, f_2 = 0, f_3 = +2, f_4 = +1$.

Next, the KANSEI information of the nodes is made by probabilistically modifying the KANSEI information of the searcher. As mentioned above, the KANSEI information of the searcher is represented as four-dimensional vectors assigned to 625 types of files. Each element value of each vector that is an integer from –2 to +2 is randomly changed to an integer in the same range with probability p_u. This probability p_u is the 'change probability of KANSEI information' defined above. A higher probability corresponds to a larger difference in KANSEI information.

In a real implementation of the proposed system, both searchers and nodes cannot know their feelings about the files until they have actually seen the contents of the files. However, if we prepare how the searchers feel about the files in advance, as described above, we can simulate a search based on KANSEI information.

4.2   System configurations

System configurations correspond mainly to configurations of the SOM for two-dimensional visualisation of the KANSEI information of nodes. The SOM used in the simulations here arranges the neurons of the competitive layer on integer lattices in the 50×50 square area. In one learning of the neuron weights, each of the vectors of 625 nodes representing their KANSEI information are inputted to the SOM once. This learning is iterated 100 times. The neighbourhood of the best matching unit (the winning neuron) in the learning is eight neurons adjacent to the best matching unit. The learning coefficient α is fixed as 0.01. In the simulations here, the map of the KANSEI information of the nodes is created once before searching in the system begins, and then the created map is used.

4.3   Simulation scenario

We assume that one searcher with the KANSEI described in Section 4.1 issues 625 different search queries, which are identical to the vectors representing the KANSEI information for the 625 types of files with serial numbers from 0 to 624. In this case, a search query for the file with serial number p, which represents what the searcher wants to feel from that file, is given by equations (1) through (4).

We consider three ways to decide the types of files of each node below. Also, we illustrate the three ways in Figure 5.

1 Randomly determining types of files of each node

As mentioned in Section 4.1, each node has five types of files in the simulations. The five types of files of each node are randomly determined from among the 625 types of files.

2 Giving a small bias to types of files of each node

As mentioned before, the 625 types of files are assigned serial numbers from 0 to 624. Also, the 625 nodes are assigned serial numbers from 0 to 624. Then, both the files and the nodes are divided into five groups, where the first group consists of files or nodes with serial numbers 0 through 124, the second group consists of files or nodes with serial numbers 125 through 249, and so on. Thus, five groups are formed by taking 125 types of files from the files with serial number 0, and five groups are formed by taking 125 nodes from the nodes with serial number 0. Then, each node belonging to the first node group randomly selects five types of files without overlapping from the 125 types of files belonging to the first group. Thus, each node belonging to the kth ($k = 1, 2, …, 5$) node group randomly selects five types of files without overlapping from the 125 types of files belonging to the kth ($k = 1, 2, …, 5$) group.

3 Giving a large bias to types of files of each node

The 625 types of files are assigned serial numbers from 0 to 624. Also, the 625 nodes are assigned serial numbers from 0 to 624. Then, both files and nodes are divided into 25 groups, where the first group consists of files or nodes with serial numbers 0 through 24, the second group consists of files or nodes with serial numbers 25 through 49, and so on. Thus, 25 groups are formed by taking 25 types of files from the files with serial number 0, and 25 groups are formed by taking 25 nodes from the nodes with serial number 0. Then, each node belonging to the first node group randomly selects five types of files without overlapping from the 25 types of files belonging to the first group. Thus, each node belonging to the kth ($k = 1, 2, …, 25$) node group randomly selects five types of files without overlapping from the 25 types of files belonging to the kth ($k = 1, 2, …, 25$) group.

If the types of files of each node are determined in the ways described above, then the KANSEI information of each node is determined by modifying the KANSEI information of the searcher on the basis of the change probability of the KANSEI information, p_u, as explained in Section 4.1.2. We use five values, 0, 0.01, 0.1, 0.3, 0.5, as the values of p_u. Then, we conduct simulations with a total of 15 scenarios by combining the three ways to determine the types of files of each node and the five values of the change probability of KANSEI information.

4.4 Observation items

The searcher with the KANSEI information mentioned in Section 4.1.2 issues the 625 different search queries in a one-by-one manner. The queries, which are four-dimensional vectors obtained by converting integers from 0 to 624 to quinary numbers from (–2, –2, –2, –2) to (+2, +2, +2, +2), are equivalent to searching for the 625 types of files with serial numbers 0 through 624 in a one-by-one manner in the

simulations. Then, at the mth $(m \geqq 1)$ search, a search query for the file with the serial number $m - 1$ is issued. A search query for a file with the serial number p is represented by equations (1) through (4).

Figure 5 Three ways to decide the types of files of each node, (a) randomly determining file types of each node (b) giving a small bias to file types of each node (c) giving a large bias to file types of each node

(a)

(b)

(c)
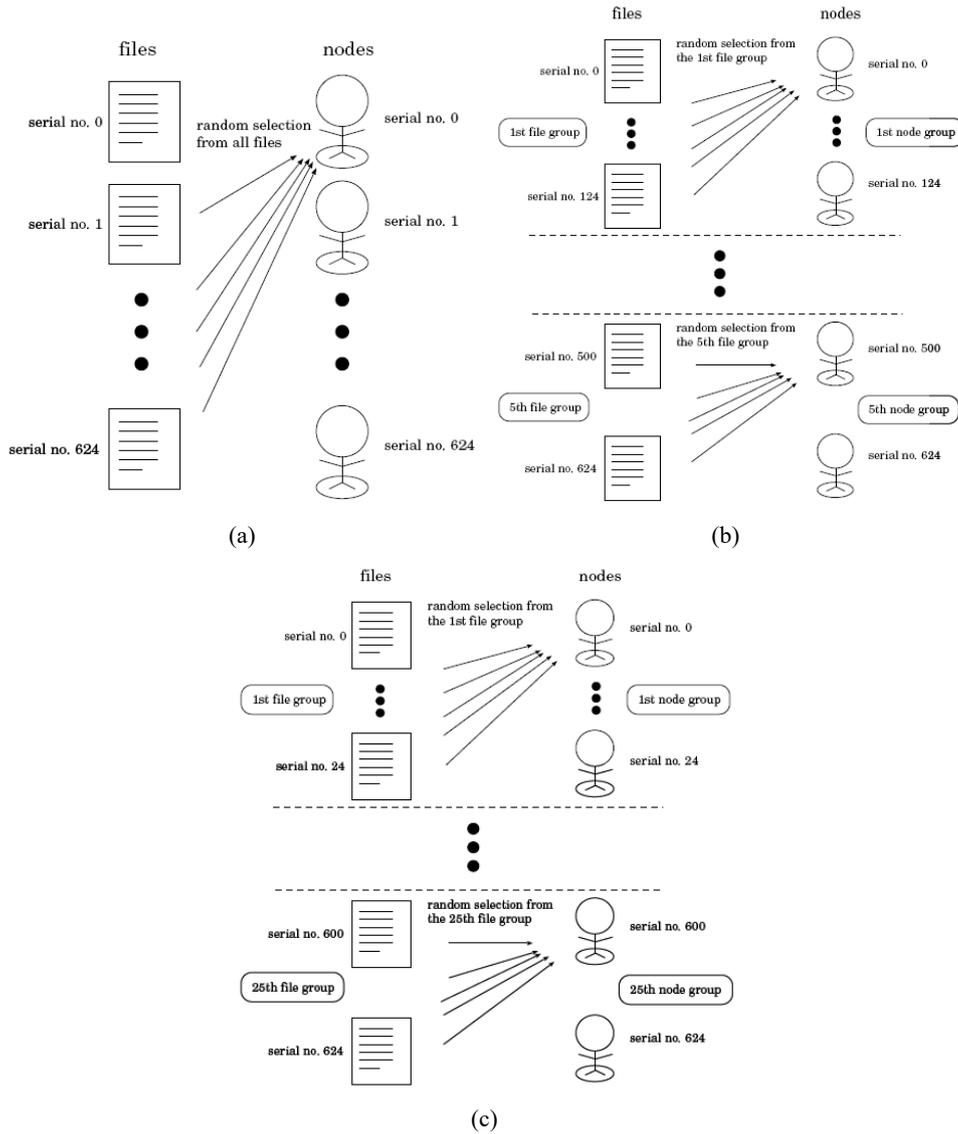
A four-dimensional search query is issued and inputted into the SOM, whose configurations are described in Section 4.2, and then is mapped onto the two-dimensional map visualising the KANSEI information of the nodes. Then, five types of files of the node closest to the mapping point on the map become the search results. In these

simulations, we consider files at only one node as the search results, but files at multiple nodes closer to the mapping point can be search results.

As mentioned above, the searcher obtains five types of files by one search. Then, we consider the following two observation items.

1   The total number of files that the searcher obtained

This means the relationship between the number of searches and the total number of types of files obtained so far. For example, consider the moment just after ten searches were completed, that is, the moment just after the search with a search query representing the KANSEI information of the searcher for a file with the serial number 9 was completed. If 20 types of files out of the total 625 types of files were obtained, the total number of files that the searcher obtained was 20 at that time. It is possible that the searcher obtained the same type of file several times in different searches, but in this case, the count increased by just one.

2   The number of files matching search queries issued by the searcher

This means the relationship between the number of searches and the accumulated number of times that a file matching an issued search query was obtained. A file matching an issued search query is a file with the serial number p when the searcher issues a search query for the file with the serial number p. Here, since the KANSEI information of each node is determined by modifying the KANSEI information of the searcher as the basis, the KANSEI information corresponding to the file with the serial number p can differ between the searcher and each node receiving the issued search query.

4.5 Systems for comparison

We prepare two systems for comparing with the proposed system.

As mentioned at the beginning of Section 4, the proposed system searches a two-dimensional space whose dimension is lower than the original space of the KANSEI information for a node in the case that the searcher points to the map of KANSEI information of the nodes and in the case that the searcher issues a search query. The SOM is used for mapping points in the original higher-dimensional space to points in the two-dimensional space, and by this mapping, the distance relationships between the points in the original space are broken at least. Therefore, as the first system for comparison, we consider a system that searches the original space of KANSEI information, that is, the four-dimensional space for a node with KANSEI information that is closest to an issued search query in terms of Euclidean distance. If the searcher and all nodes give the same KANSEI information for a file, the KANSEI information works very effectively as an identifier of the file in this system for comparison.

The second system for comparison has the opposite objective of the first system. The second system does not use the KANSEI information assigned to files for searching the files. Specifically, the second system for comparison selects a node randomly and obtains its files in one search. This system is expected to work very effectively in a situation that the KANSEI information assigned to a file is totally different between the searcher and the nodes, that is, in the situation that the KANSEI information does not work effectively as identifiers to find the desired files.

4.6   Results and discussion

Figure 6 shows the relationship between the change probability of KANSEI information and the total number of files that the searcher obtained at the end of searches for all 625 types of files. Figure 7 shows the relationship between the change probability of KANSEI information and the number of files matching search queries issued by the searcher at the end of searches for all 625 types of files.

Figure 6   Relationship between the change probability of KANSEI information and the total number of files that the searcher obtained at the end of searches for all 625 types of files, (a) randomly determining the types of files of each node (b) giving a small bias to types of files of each node (c) giving a large bias to types of files of each node (see online version for colours)
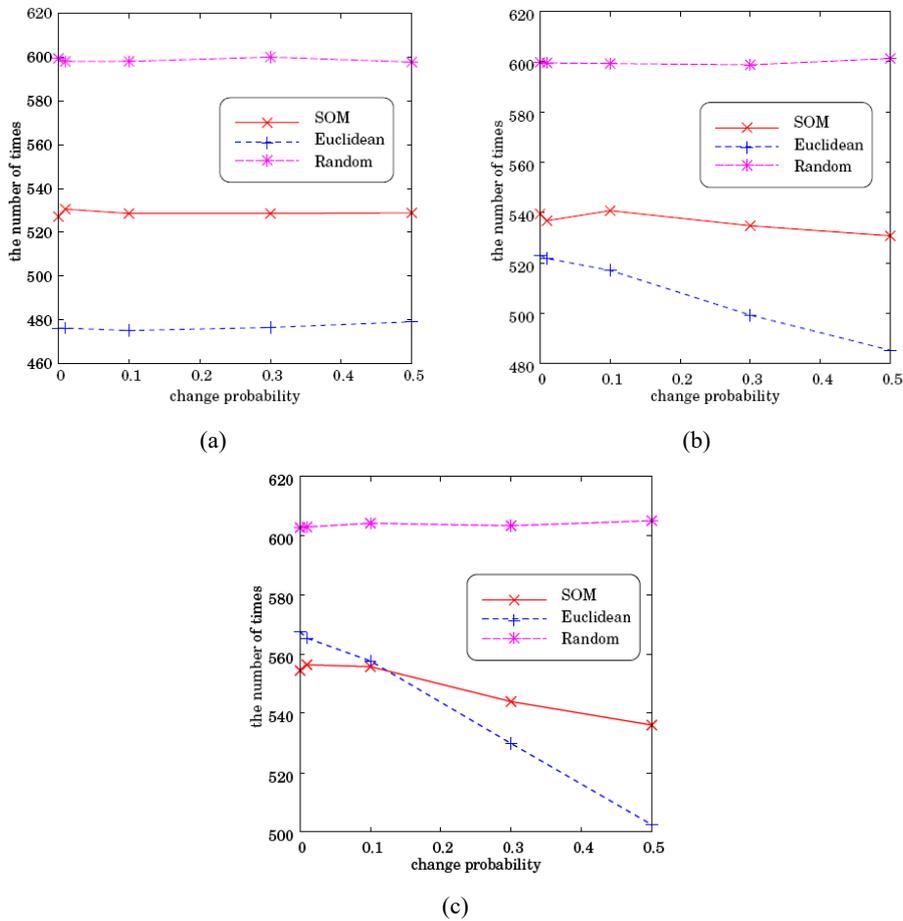

(a)

(b)

(c)

Figure 8 shows the relationship between the number of searches and the total number of files that the searcher obtained for each of the five change probabilities of KANSEI information in the case of randomly determining the types of files of each node. We call the relationship between the number of searches and the number of files that the searcher obtained as the 'time transition' hereinafter. Time here is equivalent to the number of

times of issuing search queries. Similarly, Figures 9 and 10 show the time transitions of the total number of files in the case of giving a small bias to the types of files of each node and in the case of giving a large bias to the types of files of each node, respectively.

Figure 7 Relationship between the change probability of KANSEI information and the number of files matching search queries issued by the searcher at the end of searches for all 625 types of files, (a) randomly determining types of files of each node (b) giving a small bias to types of files of each node (c) giving a large bias to types of files of each node (see online version for colours)
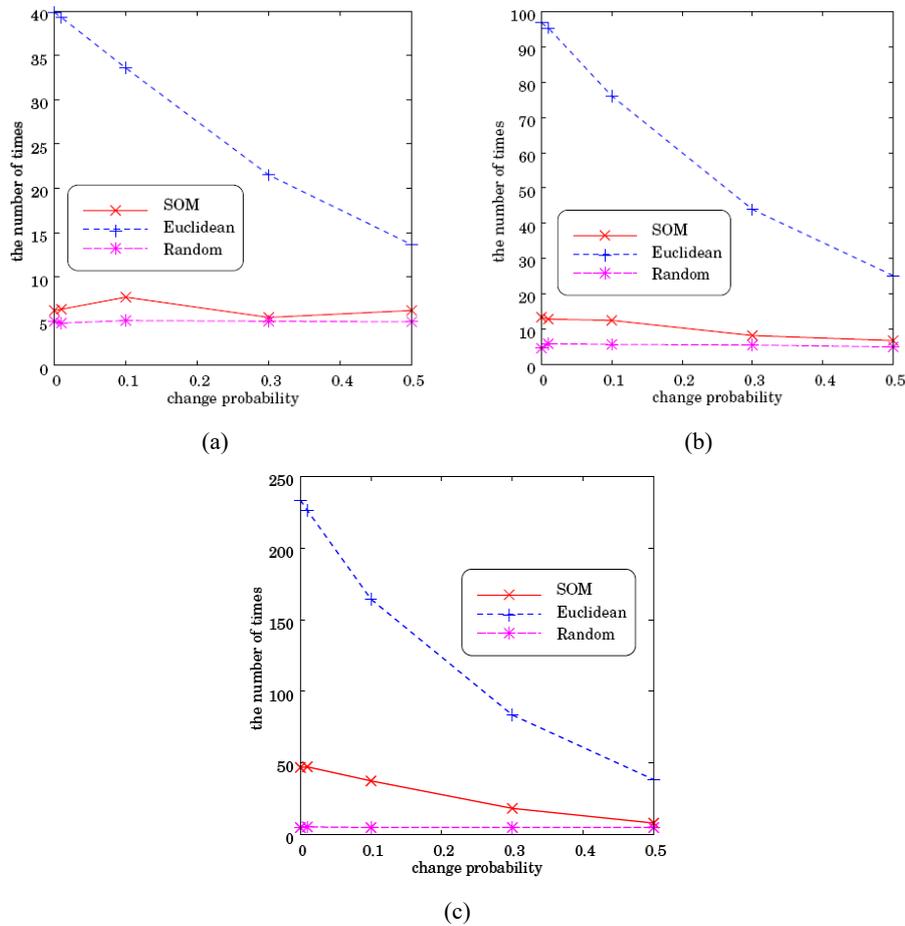

(a)

(b)

(c)

Figure 11 shows the time transition of the number of files matching the search queries issued by the searcher for each of the five change probabilities of KANSEI information in the case of randomly determining the types of files of each node. Similarly, Figures 12 and 13 show the time transitions of the number of files matching search queries issued by the searcher in the case of giving a small bias to the types of files of each node and in the case of giving a large bias to the types of files of each node, respectively.

Figure 8   Time transition of the total number of files in the case of randomly determining types of files of each node, (a) $p_u = 0$ (b) $p_u = 0.01$ (c) $p_u = 0.1$ (d) $p_u = 0.3$ (e) $p_u = 0.5$ (see online version for colours)
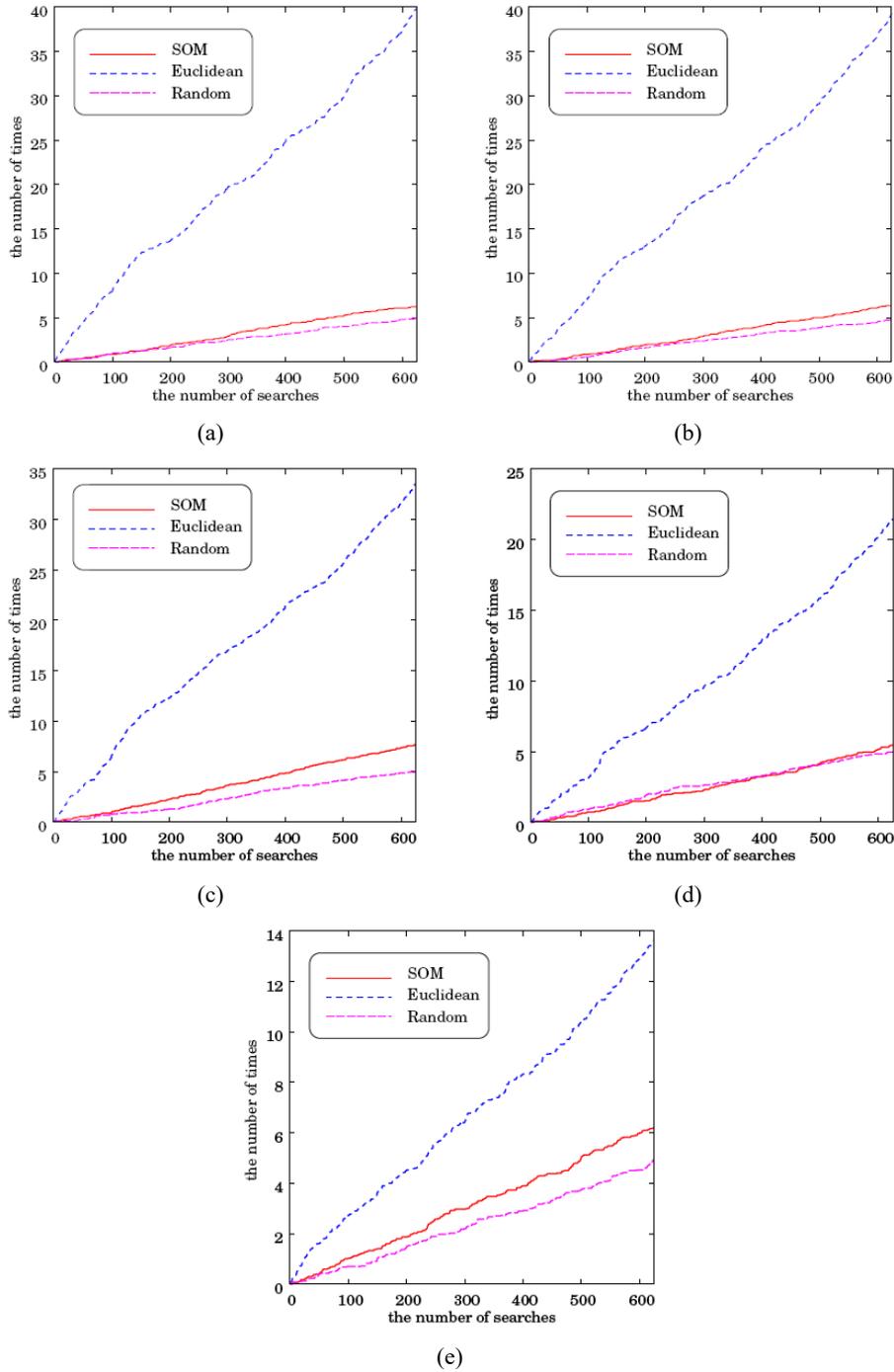

(a)

(b)

(c)

(d)

(e)

Figure 9 Time transition of the total number of files in the case of giving a small bias to types of files of each node, (a) $p_u = 0$ (b) $p_u = 0.01$ (c) $p_u = 0.1$ (d) $p_u = 0.3$ (e) $p_u = 0.5$ (see online version for colours)
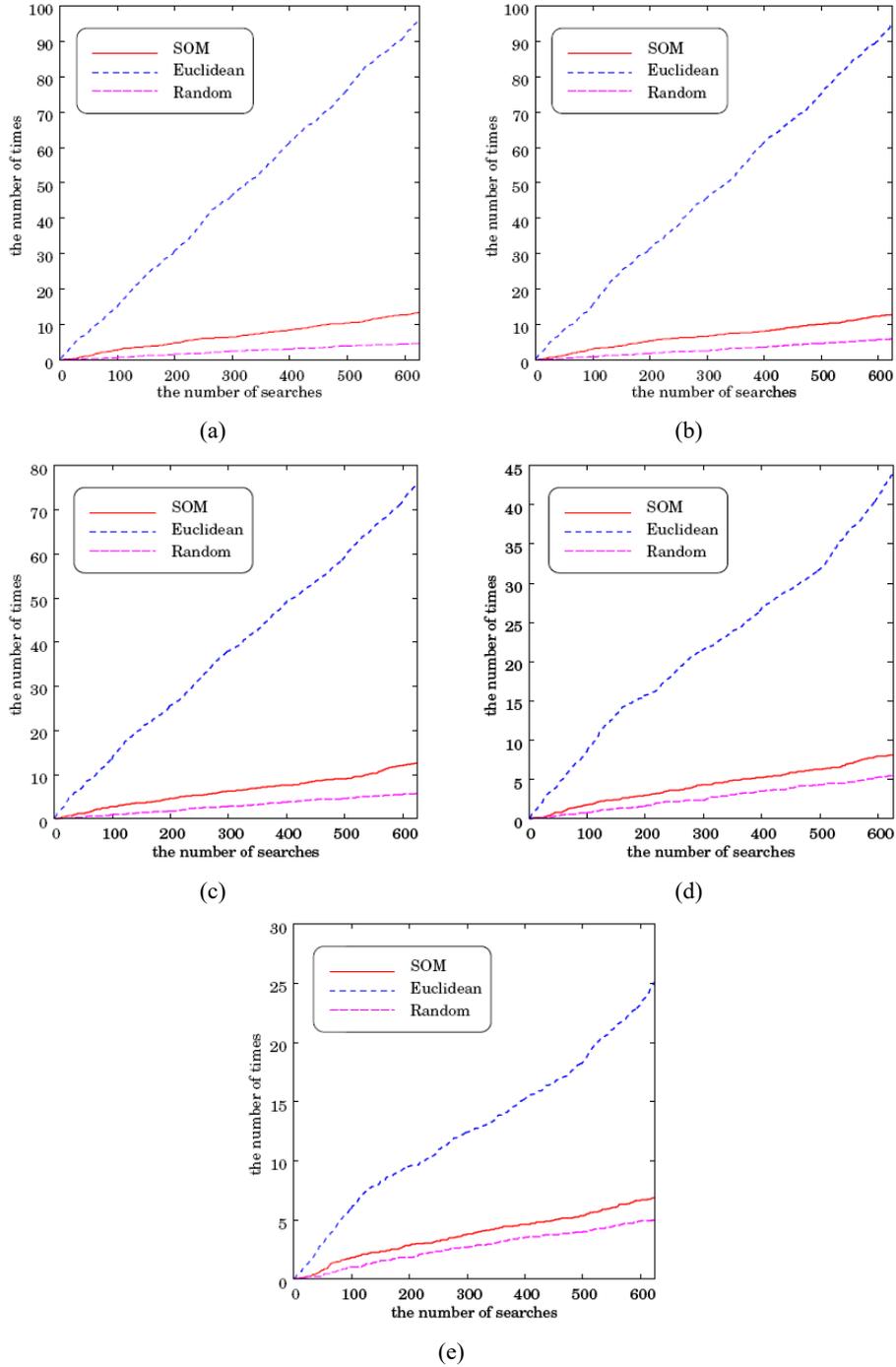

(a)

(b)

(c)

(d)

(e)
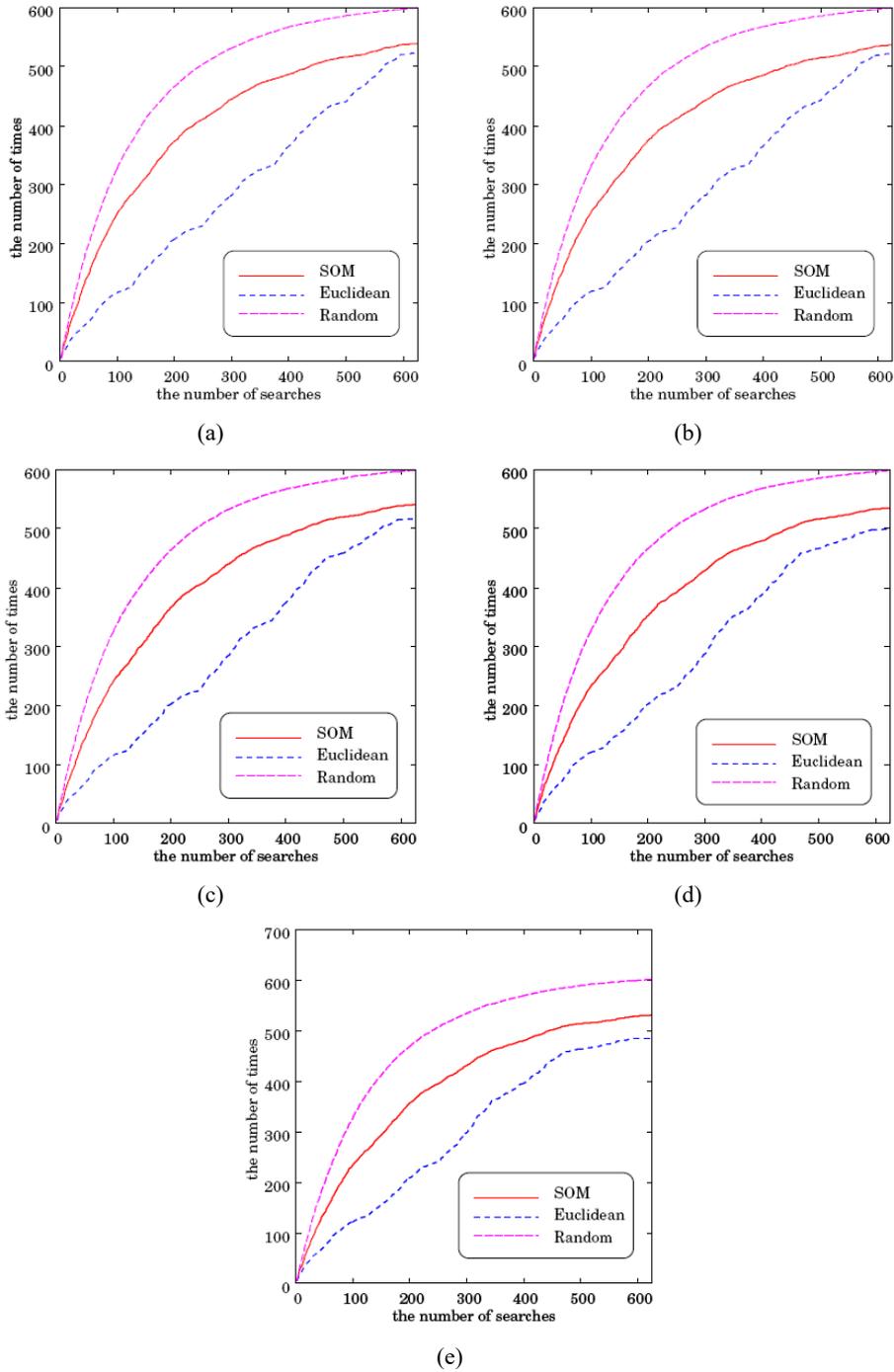
Figure 10   Time transition of the total number of files in the case of giving a large bias to types of files of each node, (a) $p_u = 0$ (b) $p_u = 0.01$ (c) $p_u = 0.1$ (d) $p_u = 0.3$ (e) $p_u = 0.5$ (see online version for colours)


(a)

(b)

(c)

(d)

(e)
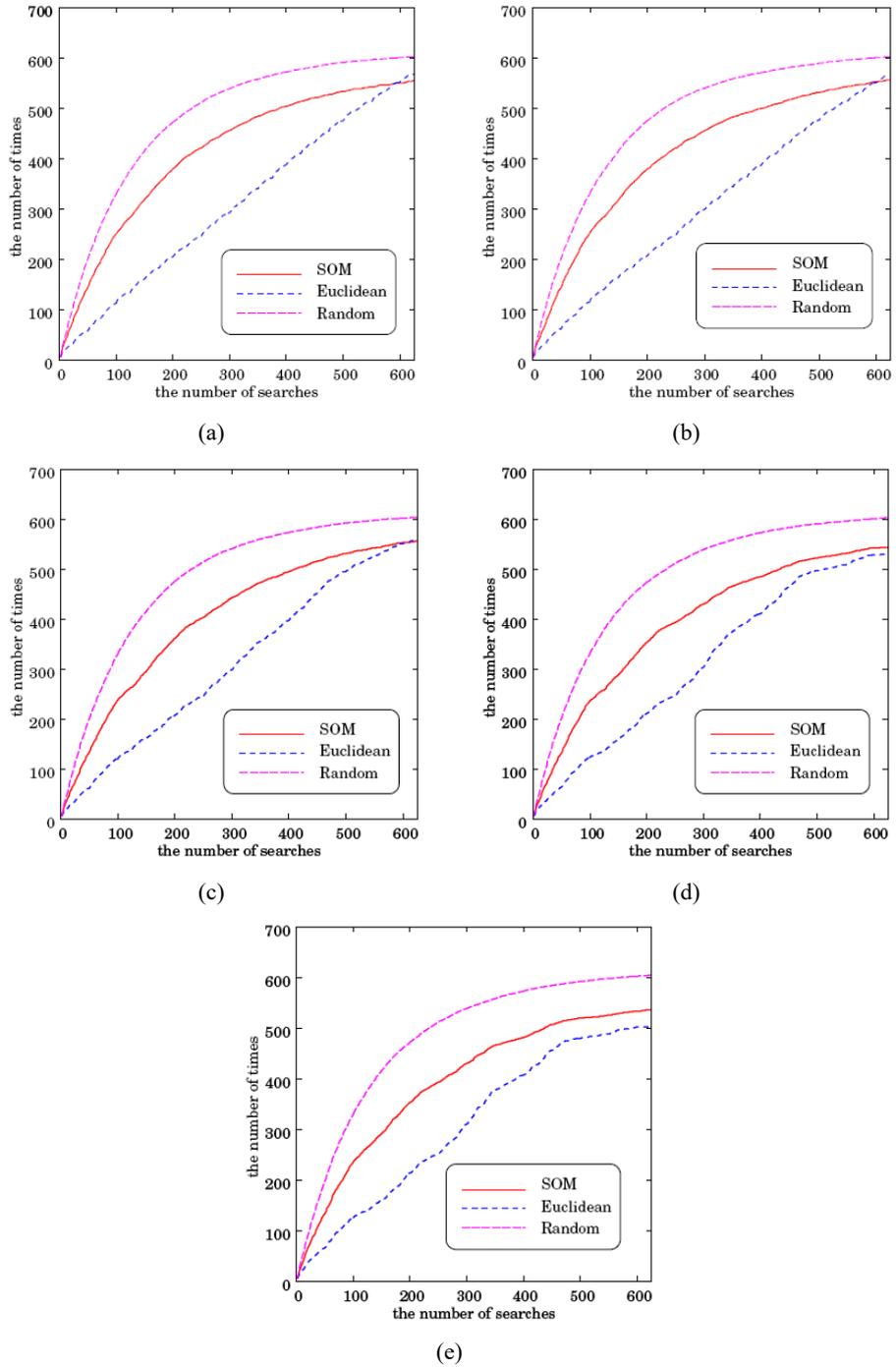
Figure 11   Time transition of the number of files matching search queries in the case of randomly determining types of files of each node, (a) $p_u = 0$ (b) $p_u = 0.01$ (c) $p_u = 0.1$ (d) $p_u = 0.3$ (e) $p_u = 0.5$ (see online version for colours)


(a)

(b)

(c)

(d)

(e)

Figure 12 Time transition of the number of files matching search queries in the case of giving a small bias to types of files of each node, (a) $p_u = 0$ (b) $p_u = 0.01$ (c) $p_u = 0.1$ (d) $p_u = 0.3$ (e) $p_u = 0.5$ (see online version for colours)

Figure 13 Time transition of the number of files matching search queries in the case of giving a large bias to types of files of each node, (a) $p_u = 0$ (b) $p_u = 0.01$ (c) $p_u = 0.1$ (d) $p_u = 0.3$ (e) $p_u = 0.5$ (see online version for colours)
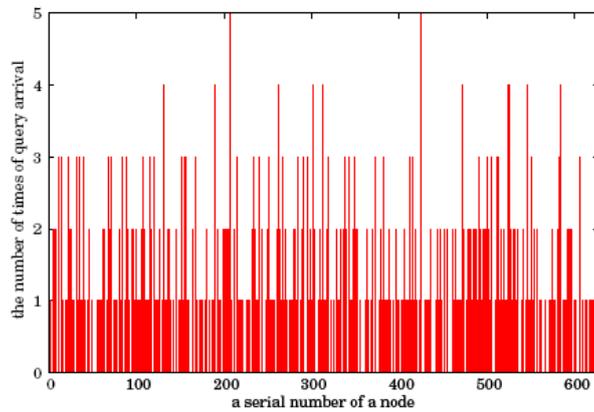

(a)

(b)

(c)

(d)

(e)
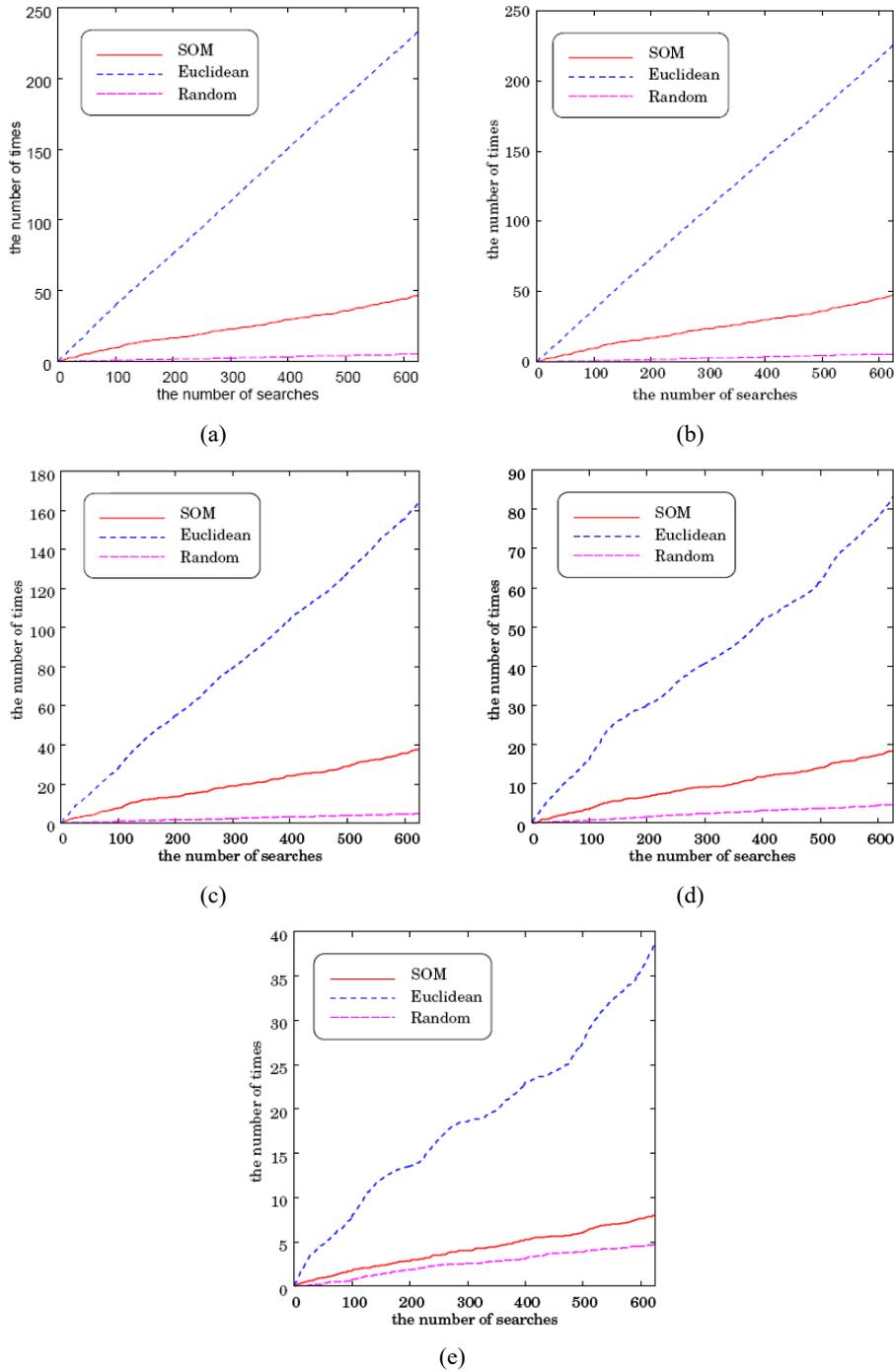
Figure 14   Number of times that a search query reached each node in the case of giving a small bias to types of files of each node and setting the change probability of KANSEI information, p_u, to zero, (a) proposed system (b) system using Euclidean distance (c) system for random searching (see online version for colours)
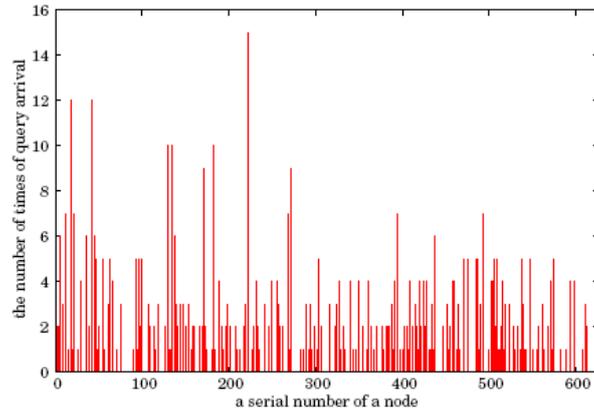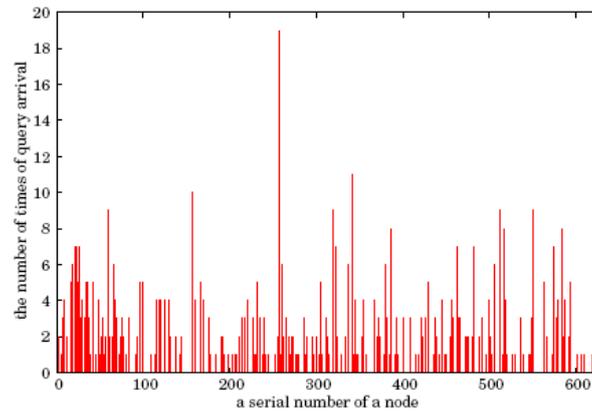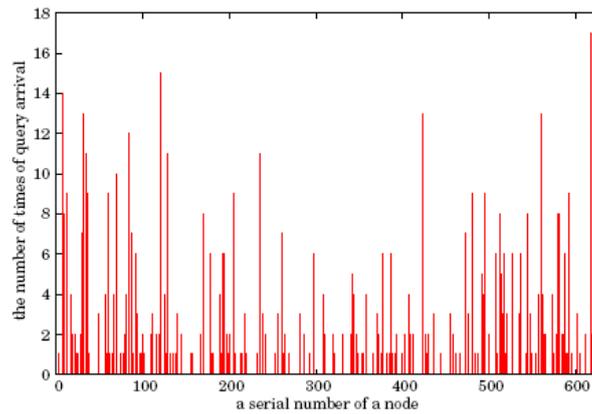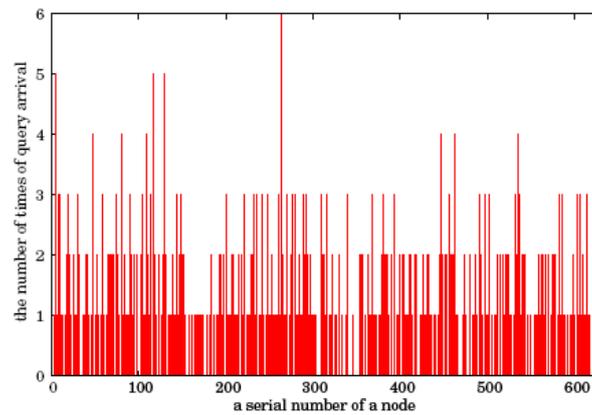

(a)
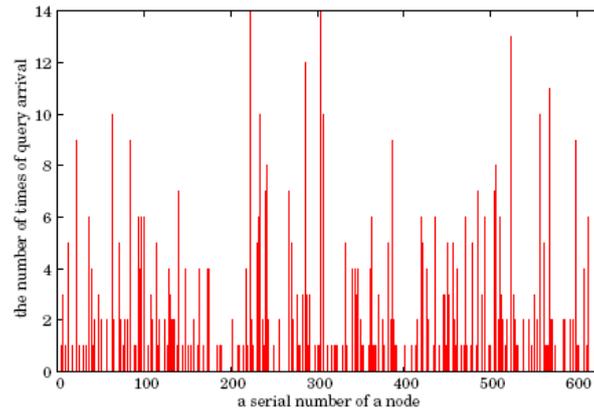
(b)

(c)

Figure 15   Number of times that a search query reached each node in the case of giving a small bias to types of files of each node and setting the change probability of KANSEI information, p_u, to 0.5, (a) proposed system (b) system using Euclidean distance (c) system for random searching (see online version for colours)

(a)

(b)

(c)

In those figures, the label 'SOM' represents the results for the proposed system, the label 'Euclidean' represents the result for the system using the Euclidean distance between an issued search query and the KANSEI information of a node for searching, and the label 'random' represents the system for random searching. All results are the average of 30 independent simulation runs.

We can observe from Figures 6 and 7 that among the three systems, the system for random searching obtains the most total number of files and the least number of files matching the search queries issued by the searcher. In addition, among the three systems, the system using Euclidean distance between an issued search query and the KANSEI information of the node for the search obtains the least total number of files and the most number of files matching the search queries issued by the searcher. The results of the proposed system fall between those two systems in terms of the two types of observed numbers. We examine the following facts below in more detail.

1   The total number of files becomes larger in order of the system for random searching, the proposed system, and the system using Euclidean distance regardless of the ways used to determine the types of files of the nodes.

2   The total number of files in the proposed system becomes slightly smaller when the change probability of KANSEI information becomes higher.

3   The total number of files in the system using Euclidean distance becomes smaller when the change probability of KANSEI information becomes higher.

4   The total number of files in the system for random searching is almost constant regardless of the change probability of KANSEI information and the ways used to determine the types of files of the nodes.

5   The number of files matching search queries becomes larger in order of the system using Euclidean distance, the proposed system, and the system for random searching regardless of the ways used to determine the types of files of the nodes.

6   The number of files matching search queries in the proposed system becomes larger when the bias of the types of files of each node becomes larger.

7   The number of files matching search queries in the system using Euclidean distance becomes larger when the bias of the types of files of each node becomes larger.

8   The number of files matching search queries in the system for random searching is constant regardless of the change probability of KANSEI information and the ways used to determine the types of files of the nodes.

First, we discuss fact (1) mentioned above. The searcher reaches 3,125 (= 625 × 5) files by 625 searches. We observed how many types of files were included in the 3,125 files in the simulations. As shown in the simulation results, the more we rely on precise Euclidean distance between a search query and the KANSEI information for searches, the smaller the total number of files. The reason is that nodes reached by searches among the 625 nodes are more biased when we rely more on the precise Euclidean distance between a search query and the KANSEI information for searches. Figure 14 shows the number of times that a search query reached each node for the three systems when giving a small bias to the types of files of each node and setting the change probability of KANSEI information to zero. Figure 15 shows the same result when giving a small bias to the

types of files of each node and setting the change probability of KANSEI information to 0.5. Each of these results is obtained by just one simulation run. We can observe from Figures 14 and 15 that the bias of nodes that a search query reaches becomes larger in order of the system using Euclid distance, the proposed system, and the system for random searching. These are just examples, but our assumption mentioned above is thought to be correct.

We think that facts (2) and (3) also result from the bias of nodes reached by searches among the 625 nodes. Then, the degree of decrease in the total number of files for the system using Euclidean distance is higher than that of the proposed system when the change probability of KANSEI information becomes higher. The reason is that the bias of nodes reached by searches for the system using Euclidean distance is higher than that of the proposed system. The reason why the bias of the nodes reached by searches becomes larger when the change probability of KANSEI information becomes higher has not been clarified from the results obtained here, but we can speculate that the distribution of KANSEI information with a four-dimensional vector becomes non-uniform for some reason when the change probability of KANSEI information becomes high. In addition, the reason for fact (4) is that the nodes reached by searches among the 625 nodes are not biased even when the change probability of KANSEI information varies.

Next, we discuss fact (5) mentioned above. The number of files matching search queries issued by the searcher, in marked contrast to the total number of files, becomes larger in order of the system using Euclidean distance, the proposed system, and the system for random searching. The reason is also related to the bias of nodes reached by searches among the 625 nodes. From the results, we can assume that the reach of search queries similar to each other is biased not to an irrelevant node but to a relevant node that has at least one file matching one of the search queries, such that five search queries corresponding to serial numbers of files from 0 to 4 (search queries of (0, 0, 0, 0) through (0, 0, 0, 4)) reach a node having a file with serial number 0. Since the system using Euclidean distance relies on KANSEI information with a vector form as the identifiers for searches, the irrelevant bias of nodes that search queries reach would be avoided. The proposed system searches the two-dimensional space reduced from the original four-dimensional space of search queries for a node, and therefore, the distance relationships between the points in the original space are broken at least by the mapping. So, we presume that the bias of nodes that search queries reach is smaller than in the system using the original four-dimensional Euclidean space, but search queries reach irrelevant nodes more frequently.

As for facts (6) and (7) mentioned above, the reasons are thought to be as follows. A larger bias of the types of files of each node corresponds to a higher bias of nodes that the search queries reach and a higher relevancy of nodes reached by the search queries in both the system using Euclidean distance and the proposed system. The reason why search queries reach relevant nodes is that the KANSEI information of a node becomes closer to the KANSEI information assigned to five files of a node when the bias of the types of files of each node becomes larger. In addition, even when the change probability of the KANSEI information is high, if the bias of the types of files of each node is large, the Euclidean distance between the KANSEI information of the nodes and the search query would be useful for finding a file matching the search query.

Next, we discuss Figures 8, 9, 10, 11, 12 and 13. First, we can observe from Figures 8, 9 and 10 that the total number of files in the proposed system and the system

for random searching is large during the early stage of searches and small during the later stage of searches. This means that these two systems obtain many types of files quickly. Furthermore, the system using Euclidean distance obtains almost the same number of files at any moment for all searches. From these results, we presume that even for a small difference in the total number of files at the end of all searches between the proposed system and the system using Euclidean distance, or between the system for random searching and the system using Euclidean distance, if we compare the total number of files obtained by the least number of searches using similar multiple search queries between them, the difference in the total number of files between them would be significant. In addition, we can observe from Figures 11, 12 and 13 that any of the three systems obtains almost the same number of files at any moment, though the increase rates of the files are different between the three systems. From these results, we think that the abilities of the three systems in finding files matching search queries do not change from those shown in Figures 11, 12 and 13, even if we compare the number of files matching search queries obtained by the least number of searches by using similar multiple search queries between them.

In summary, the proposed system is shown to be a more balanced system compared to the system using Euclidean distance and the system for random searching in terms of providing opportunities in which a searcher obtains desired files and in which a searcher reaches other files under the assumption that people have various KANSEI. In addition, only the proposed system can provide a search function with a search query and a two-dimensional map of the KANSEI information of nodes that allows users to decide a node of the search object. We expect the effect of this advantage of the proposed system in its real use. We need to demonstrate the advantage of the proposed system through implementing the system in future work.

5   Conclusions

We proposed a hybrid P2P information search system that combines:

1   an idea that a searcher searches people who have KANSEI well fitted to the searcher, not the information itself

2   a P2P network on which we expect to build a scalable information search system for dealing with the rapid increase of the amount of information

3   a folksonomy, which is a system that allows users to classify information by themselves, to enable us to obtain our desired information from the huge amount of information on the network quickly and reliably.

The proposed system is meant to enable us to find nodes (persons) having our favourite KANSEI. In addition, we evaluated the proposed simulation through simulations.

The salient feature of the proposed system is the building of a two-dimensional map of the nodes' KANSEI information represented by more than two-dimensional numerical vectors, that is, a map of the entire network, by mapping the KANSEI information of nodes onto the two-dimensional space by a SOM and then utilising the map for information searches. In the simulations, we examined the effect of using the two-dimensional map reduced from the original space of KANSEI information under the assumption that people have various KANSEI. We realised from the simulation results

that the proposed system is a more balanced system in terms of providing opportunities in which a searcher obtains the desired files and in which a searcher reaches other files. These results suggest that the proposed system is suitable not only for information searches that seek explicitly desired information but also for heuristic information searches such as searches for persons whose KANSEI are of interest to the searcher.

In future work, we need to implement the system to demonstrate the effect of using the two-dimensional map, which is one of the advantages of the proposed system. Humans are good at finding some tendency included in geometric patterns, and in the proposed system, we can expect that a searcher finds some tendency included in the two-dimensional map and actively determines a node of a search object by considering its tendency.

Acknowledgements

References

Azzedin, F. (2014) 'Taxonomy of reputation assessment in peer-to-peer systems and analysis of their data retrieval', *Knowledge Engineering Review*, September, Vol. 29, No. 4, pp.463–483.

Crespo, A. and Garcia-Molina, H. (2005) 'Semantic overlay networks for P2P systems', in Moro, G., Bergamaschi, S. and Aberer, K. (Eds.): *Agents and Peer-to-Peer Computing, Lecture Notes in Artificial Intelligence*, Vol. 3601, pp.1–13.

Dabek, F., Cox, R., Kaashoek, F. and Morris, R. (2004) 'Vivaldi: a decentralized network coordinate system', in *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp.15–26.

Delicious [online] https://delicious.com/.

Feng, Y. and Uchida, S. (2010) 'How to design KANSEI retrieval systems?', in *Proceedings of Web-Age Information Management, Lecture Notes in Computer Science*, Vol. 6184, pp.405–416.

Flickr [online] https://www.flickr.com/.

Gupta, M., Judge, P. and Ammar, M. (2003) 'A reputation system for peer-to-peer networks', in *Proceedings of the 13th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, New York, NY, USA, pp.144–152.

Hao, L., Lu, S., Tang, J. and Yang, S. (2009) 'An efficient and robust self-storage P2P reputation system', *Distributed Sensor Networks*, Vol. 5, No. 1, p.40.

Hayashida, N. and Takagi, H. (2000) 'Visualized IEC: Interactive evolutionary computation with multidimensional data visualization', in *26th Annual Conference of the IEEE Industrial Electronics Society (IECON 2000)*, pp.2738–2743.

Himali, D.M.R., Navathe, S.B. and Prasad, S.K. (2013) 'SAS: semantics aware search in P2P networks', in *2013 33rd IEEE International Conference on Distributed Computing Systems Workshops (ICDCSW 2013)*, Philadelphia, PA, USA, pp.178–183.

Huang, X., Sobue, S. and Chen, Y-W. (2011) 'Retrieval of clothing fabric images based on KANSEI words', *Information – An International Interdisciplinary Journal*, February, Vol. 14, No. 2, pp.515–525.

Joseph, S. (2002) 'Neurogrid: semantically routing queries in peer-to-peer networks', in *Proceedings of the International Workshop on Peer-to-Peer Computing (Co-located with Networking 2002)*, Pisa, Italy, May.

Joseph, S. and Hoshiai, T. (2003) 'Decentralized meta-data strategies: effective peer-to-peer search', *IEICE Transactions on Communications*, Vol. E86-B, No. 6, pp.1740–1753.

Jung, J.J. (2010) 'Reusing ontology mappings for query routing in semantic peer-to-peer environment', *Information Sciences*, September, Vol. 180, No. 17, pp.3248–3257.

Kohonen, T. (1990) 'The self-organizing map', *Proceedings of the IEEE*, September, Vol. 78, No. 9, pp.1464–1480.

Lua, E.K., Crowcroft, J., Pias, M., Sharma, R. and Lim, S. (2005) 'A survey and comparison of peer-to-peer overlay network schemes', *IEEE Communications Surveys & Tutorials*, Second Quarter, Vol. 7, No. 2, pp.72–93.

Mathes, A. (2004) *Folksonomies – Cooperative Classification and Communication through Shared Metadata* [online] http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

Mawlood-Yunis, A-R., Weiss, M. and Santoro, N. (2010) 'From P2P to reliable semantic P2P systems', *Peer-to-Peer Networking and Applications*, December, Vol. 3, No. 4, pp.363–381.

Miyakawa, A., Ishida, T. and Shibata, Y. (2008) 'Visual and physical feature analysis for KANSEI retrieval method in VR Japanese crafting presentation system', in *22nd International Workshops on Advanced Information Networking and Applications*, Ginowan, Japan, pp.477–482.

Ohnishi, K., Yoshida, K. and Oie, Y. (2009) 'Folksonomical P2P file sharing networks using vectorized KANSEI information as search tags', *IEICE Transactions on Information and Systems*, Vol. E92-D, No. 12, pp.2402–2415.

Pogkas, I., Kriakov, V., Chen, Z. and Delis, A. (2009) 'Adaptive neighborhood selection in peer-to-peer networks based on content similarity and reputation', *Peer-to-Peer Networking and Applications*, March, Vol. 2, No. 1, pp.37–59.

Rostami, H., Habibi, A. and Livani, E. (2008) 'Semantic routing of search queries in P2P networks', *Parallel and Distributed Computing*, December, Vol. 68, No. 12, pp.1590–1602.

Sato, N., Anse, M. and Tabe, T. (2007) 'A method for constructing a movie-selection support system based on KANSEI engineering', in *Proceedings of Human Interface and the Management of Information: Methods, Techniques and Tools in Information Design, Lecture Note in Computer Science*, Vol. 4557, pp.526–534.

Takagi, H. and Noda, T. (2004) 'Media converter with impression preservation using a neuro-genetic approach', *Hybrid Intelligent Systems*, Vol. 1, No. 1, pp.49–56.

Tang, C., Xu, Z. and Dwarkadas, S. (2003) 'Peer-to-peer information retrieval using self-organizing semantic overlay networks', in *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'03)*, Karlsruhe, Germany, pp.175–186.

Vera-del Campo, J., Pegueroles, J., Hernandez-Serrano, J. and Soriano, M. (2012) 'Design of a P2P content recommendation system using affinity networks', *Computer Communications*, December, Vol. 36, No. 1, pp.90–104.

Yali, F. and Kui, C. (2008) 'Indexing wood image for retrieval based on KANSEI factors', in *9th International Conference on Signal Processing (ICSP)*, pp.1099–1102.

Yamashiro, D., Yoshikawa, T. and Furuhashi, T. (2006) 'Visualization of search process and improvement of search performance in multi-objective genetic algorithm', in *IEEE Congress on Evolutionary Computation, 2006, CEC 2006*, pp.1151–1156.

Yeferny, T., Arour, K. and Bouzeghoub, A. (2013) 'An efficient peer-to-peer semantic overlay network for learning query routing', in *IEEE 27th International Conference on Advanced Information Networking and Applications (IEEE AINA)*, Barcelona, Spain, pp.1025–1032.

Yoshida, K., Kato, T. and Yanaru, T. (1999) 'A study of database system with KANSEI information', in *Proceedings of 1999 IEEE international Conference on System, Man, and Cybernetics*, pp.253–256.