A PROPOSAL OF SEMI-AUTOMATIC INDEXING ALGORITHM FOR MULTI-MEDIA DATABASE WITH USERS' SENSIBILITY

Takashi MITSUISHI *Jun SASAKI* *Yutaka FUNYU*

Faculty of Software & Information Science, Iwate Prefectural University
Sugo 152-52, Takizawa, Iwate 020-0193, Japan
{takashi, jsasaki, funyu}@soft.iwate-pu.ac.jp

Abstract: We propose a semi-automatic and dynamic indexing algorithm for multi-media database(e.g. movie files, audio files), which are difficult to create indexes expressing their emotional or abstract contents, according to user's sensitivity by using user's histories of access to database. In this algorithm, we simply categorize data at first, create a vector space of each user's interest(user model) from the history of which categories the data belong to, and create vector space of each data(title model) from the history of which users the data had been accessed from. By continuing the above method, we could create suitable indexes, which show emotional content of each data. In this paper, we define the recurrence formulas based on the proposed algorithm. We also show the effectiveness of the algorithm by simulation result.

Keywords: multi-media database, data mining, indexing of emotional content, sensibility analysis

1 Introduction

Due to the recent remarkable progress of computer and network technologies, video and audio database are constructed with in the computer network, and are accessed via the network. On-line music shops and video on demand systems are provided are few of the many examples. In order to use these multi-media database effectively, it is necessary to retrieve target data efficiently, and is necessary to create appropriate indexes for retrieval.

Though, it is difficult to create indexes, expressing contents themselves of different data, we propose the algorithm to infer the content of each data from partiality of users' accesses, which are caused by their interests. In this paper, we modeled proposed algorithm and define the recurrence formulas based on the model. We also show the effectiveness of the algorithm by simulation result.

This paper is organized as follows. In section 2, we propose a model of our indexing algorithm, and define the recurrence formulas. In section 3, we simulate use of database and evaluate our proposal. In section 3, we compare our proposal with other related works, and conclude our paper in section 5.

2 Semi-automatic Indexing Algorithm with Users' Sensibility

In this section, we describe the necessity of indexes which are expressing contents of data themselves and the problem of existing categorizing methods to create indexes. And then, we propose a semi-automatic and dynamic algorithm to create indexes and define recurrence formulas based on the algorithm.

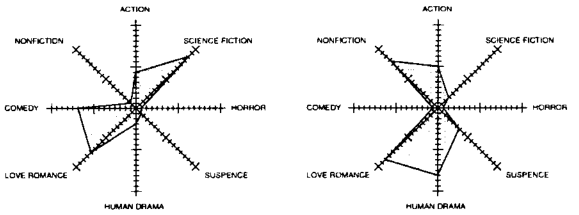
2.1 Retrieve by Directions of Contents

In order to use database effectively, it is necessary that user can retrieve data efficiently by using some intuitive words which are expressing target data, and constructor of database should create appropriate indexes, which shows characteristics of different data, for that purpose.

It is easier to create indexes of data by using accompany information such as movie title, staffs, casts, etc in movie library. However, it is difficult to create indexes that show the contents themselves or directions of contents of the data. The objective of this research is to create the latter such as "science fiction", "action", "love romance", "comedy" and so on, or the abstract and emotional words such as "cool", "exciting", "heart warming", "funny".

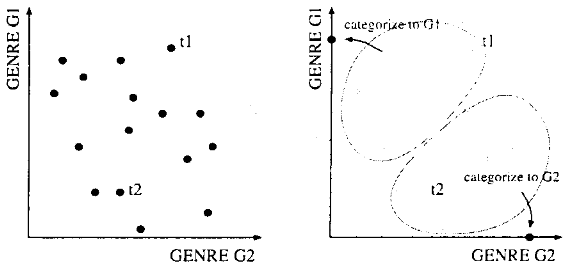
Generally speaking, these directions of contents in movie library or music library are called genres, and most of these data are roughly categorized into these genres roughly. Though we could use these genre names as indexes, most of these data have not only one direction but also many directions complexity.

For example, "Back to The Future", which is mostly categorized into "science fiction". in a way could be further categorized into "love romance" and "commedy" as it also contains both love romance and commedy. Similarly, "TITANIC", which is categorized into "love romance", also contains human dramas blending



(1) Back to The Future (2) TITANIC

Figure 1: complexity of directions



(1)Distribution of Data (2)Categorization of Data

Figure 2: directions and categorization of data

fictitious love story in the historical accident (Fig.1) [1].

As shown in the above, each data has complex directions. If it is possible to create indexes which are appropriately expressing these complex directions of different data, not only we could retrieve target data efficiently; retrieving "Back to The Future" by querying with "Science fiction, also love romance and comedy in some degree", but also a person, who does not know much about "Back to The Future", could find it by querying with "love romance", and he gets new knowledge of it.

In addition to that, when it is possible to create these indexes consisting of complex genre names, we could retrieve data by querying with abstract and emotional words by introducing the corresponding relations between genre names and these words according to the image which user feel about them.

2.2 Problem of Existing Categorization

In case of existing simple categorization, each data would be ordinarly categorized into only one genre, and the genre names do not express directions of the contents of the data. Thus, we could not retrieve data effectively by using these genre names as indexes.

For example, on the assumption that there exist two genres G1 and G2 as directions in contents of data, the distribution of data according to these genres could be shown like Fig.2-(1), and these data would be categorized as Fig.2-(2) by simple categorization. As a result, even though data t1 has a good deal of both directions of G1 and G2, it will be categorized into G1 and it is impossible to express that it has a direction of G2. Whereas, about data t1, which has a little amount of G1 and G2, it will be categorized into G2 even though it has only a little amount of G2 as well as it loses the information that it has a little amount of G1 by this categorization.

Then, the indexes, which are created by ordinary existing simple categorization, might not express the contents of different data, nor could we retirieve data using these genre names as indexes.

On the other hand, there are several researches on presumption of directions of each content of data such as design patterns of images, musical scores, and so on. These are the methods which investigate which patterns will give which impressions in advance, analyze frequency elements, and extract paterns and repetition of them.

However, for the data which includes the amount of data and the content of the data is complex and varied such as video data and sampling audio data, it is difficult to presume the directions by these frequency analysis.

When we want to create indexes showing appropriate directions of contents of data at the time of constructing a database, a constructor of the database should investigate each data closely, comprehend it in detail, and create indexes. However, it might be unfeasible to do for whole data.

2.3 Proposal of Semi-Automatic Indexing Algorithm with Users' Sensibility

In order to presume the directions of contents of data and create indexes according to them, we propose a semi-automatic indexing algorithm with users' sensibility by analyzing users' histories of access to database.

About some data such as movies and musics, some users might know the content of them in some degree beforehand, even though they do not know details of data. These users might tend to use data according to their interest or concern, and we expect there are partiality in their accesses to database. When we analyze users' histories, we could find these partiality and estimate directions of contents of each data.

Data t1 in Fig.2 for an example, in fact has considerable amount of direction of G2 even though it is categorized into G1. A user, who knows this matter and is highly interested in G2, will access t1, and we could prospect users, who tend to access data which are categorized into G2, would access t1. When we analyze histories of users who accessed t1, investigate with other data they has accessed before, guess there interest

[1] It is not by detailed analysis

directing G2, and we could find t1 has some amount of direction of G2.

Hence, we could presume the directions of different users' interest with the histories which data they have accessed, and presume the directions of contents of different data with to the history which users the data have been accessed. We could update the presumption of data when each data is accessed by some user, and create indexes of the data according to the presumption. It means we could create indexes which show the directions of contents of different data semi-automatically and dynamically at the moment of use of database which are simply categorized into several genres,

2.4 Model of The Indexing Algorithm

Based on the theory we mentioned above, we define recurrence formulas in order to renovate an update both vector models showing directions of interets of each user and directrions of contents of each data.

At first, we suppose that there are a set of data $\{m|m \in M\}$ and a set of users $\{n|n \in N\}$, and each direction of contents of data and users' interest can be expressed by vector space with several dimensions. We define these vector spaces expressing contents of data m and interests of a user n at a time of t as title model $\vec{T}_m(t)$ and user model $\vec{U}_n(t)$.

For example, when vector spaces of title models and user models have eight dimension and each element of a vector has magnitude from 0 to 1, $\vec{T}_m(0)$; first value of a title model of data m, which is categorized into G1 at a time of 0 could be expressed as the following.

$$\vec{T}_m(0) = (1,0,0,0,0,0,0,0)$$

And $\vec{U}_n(t_1)$; a user model of a user n at a time of t_1 could be also expressed as the following.

$$\vec{U}_n(t_1) = (0.9,0.4,0.3,0.3,0.2,0.5,0.8,0.1)$$

Thuogh we assume there are some genres which are relatively similar or related to each other in simple categorization, we suppose all genres are independent from others.

Next, we define formulas as methods of renovation of title models and user models from defined vector spaces.

When we express $\{m(n,t)\}$; a set of data which are accessed by a user; n at a time of t as the following,

$$\{m(n,t)\} = \{m|\text{accessed } m \in M \text{ by } n \text{ at } t\}$$

$\{\vec{T}(n,t_1,t_2)\}$; a set of title model which are accessed by a user; n in a moment of $(t_1,t_2]$ could be expressed as the following,

$$\{\vec{T}(n,t_1,t_2)\} = \{\vec{T}_m(\tau)|m \in \{m(n,t)\}, \tau \in (t_1,t_2]\}\}$$

and its average; $aver(\{\vec{T}(n,t_1,t_2)\})$ could be found as the following.

$$aver(\{\vec{T}(n,t_1,t_2)\}) = \frac{\displaystyle\sum_{\tau\in(t_1,t_2]}\ \sum_{m\in\{m(n,\tau)\}} \vec{T}_m(\tau)}{\#\{\vec{T}(n,t_1,t_2)\}}$$

Here, we suppose $\vec{U}_n(t_1)$; a user model of a user n at a time of t and $aver(\{\vec{T}(n,t_1,t_2)\})$; an average of title models which are accessed by m in a moment $[t_1,t_2)$ as Fig.3 and Fig.4 respectively.

When we look the difference between $\vec{U}_n(t_1)$ and $aver(\{\vec{T}(n,t_1,t_2)\})$, we could find a little different in G1, and that the average of accessed data is considerably higher than presumed interest in G2 and lower in G6. Then, we could renovate a presumption of $\vec{U}_n(t_2 + \Delta t)$; a user model after t_2 as shown in Fig.5.

So, we define the following formula in order to renovate a presumption of a user model of user after a time of t_2 from the presumption of the same user at a time of t_1 and the average of title model of data, which are accessed by him/her in a moment of $(t_1,t_2]$.

$$\vec{U}_n(t_2 + \Delta t) = \vec{U}_n(t_1) - \alpha_U \left\{\vec{U}_n(t_1) - aver(\{\vec{T}(n,t_1,t_2)\})\right\} \quad (1)$$

α_U is proportional coefficient in order to deside the degree of diminishing between a user model at t_1 and an average of data models in $(t_1,t_2]$.

Similarly, we define the following formula in order to renovate a presumption of a title model of data m after a time of t_2 from a title model of the same data at t_1 and an average of user model, who accessed it in a moment $(t_1,t_2]$.

$$\vec{T}_m(t_2 + \Delta t) = \vec{T}_m(t_1) - \alpha_T \left\{\vec{T}_m(t_1) - aver(\{\vec{U}(m,t_1,t_2)\})\right\} \quad (2)$$

When we regard $t_1 = t_2 = t$ by making $(t_1,t_2]$ enough short and a user n access only one data n in the moment, there is no change in all user models and title models, and we could re-express formula (1) and (2) as the followings respectively.

$$\vec{U}_n(t + \Delta t) = \vec{U}_n(t) - \alpha_U\{\vec{U}_n(t) - \vec{T}_m(t)\}$$
$$\vec{T}_m(t + \Delta t) = \vec{T}_m(t) - \alpha_T\{\vec{T}_m(t) - \vec{U}_n(t)\}$$

By rewriting $U_n(t)$ and $T_m(t)$ with c_{U_n}; the number of times a user n have accessed through to t and c_{T_m}; the number of times data m have been accessed through to t respectively, we get the following formulas.

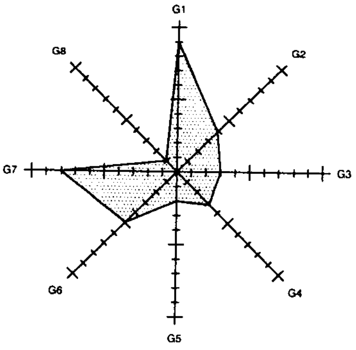


Figure 3: a user model of n at t_1

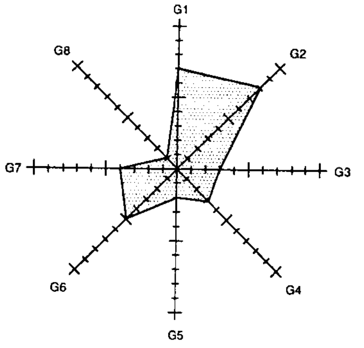


Figure 4: an average of accessed title models by n in $(t_1, t_2]$

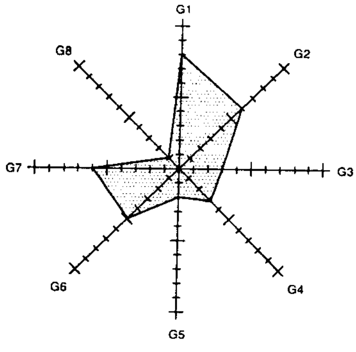


Figure 5: a presumption of a user model after t_2

$$\left.\begin{array}{l}\vec{U}_n(c_{U_n}+1)= \\ \quad \vec{U}_n(c_{U_n}) - \alpha_U\left\{\vec{U}_n(c_{U_n}) - \vec{T}_m(c_{T_m})\right\} \\ \vec{T}_m(c_{T_m}+1)= \\ \quad \vec{T}_m(c_{T_m}) - \alpha_T\left\{\vec{T}_m(c_{T_m}) - \vec{U}_n(c_{U_n})\right\}\end{array}\right\} \quad (3)$$

It means that these are the methods in order to renovate both a user model and a title model at a time of that a user accesses data by the difference between his/her user model and its title model.

On the other hand, when we make $t_1 = 0$ and $\alpha_U = \alpha_T = 1$, we get the following formulas from (1) and (2).

$$\begin{array}{l}\vec{U}_n(t_2+\Delta t) = aver(\{\vec{T}(n,0,t_2)\}) \\ \vec{T}_m(t_2+\Delta t) = aver(\{\vec{U}(m,0,t_2)\})\end{array}$$

By assuming that only one user could access only one data at a time and rewriting $U_n(t)$ and $T_m(t)$ with c_{U_n} and c_{T_m} respectively like the above formulas (3), we get the following formulas.

$$\left.\begin{array}{l}\vec{U}_n(c_{U_n}+1)= \\ \qquad \dfrac{c_{U_n}\times\vec{U}_n(c_{U_n})+\vec{T}_m(c_{T_m})}{c_{U_n}+1} \\ \vec{T}_m(c_{T_m}+1)= \\ \qquad \dfrac{c_{T_m}\times\vec{T}_m(c_{T_m})+\vec{U}_n(c_{U_n})}{c_{T_m}+1}\end{array}\right\} \quad (4)$$

It means that these are the methods in order to renovate both a user model and a title model at a time of that a user accesses data by the average of title models of data which have been accessed by him/her and the average of user models of users who have accessed it in the past respectively.

3 Experiment

In this section, in order to confirm the effectiveness of our proposal we simulate users' retrievals and evaluate renovation of user models and title models with formulas (3) and (4).

3.1 Substance

In this experiment, we suppose there are M data be numbered from 0 to $M-1$, N user be numbered from 0 to $N-1$, and G dimension of directions in contents of data and users' interests. In advance, we set the value of real number; from 0 to 1 at random to each element in vectors as a potential direction of contents of each data $\vec{PT}_m$ or interests of each user $\vec{PU}_n$ apart from a title model $\vec{T}_m$ or a user model $\vec{U}_n$ in order to simulate which data a user will select. We set 1 to an element, which is corresponding to an element having a maximum value in potential direction of contents of each data, and 0 to the other elements in each title model as a first value, and set 0 to all elements in each user model.

We simulate users in order from 0 to $M-1$ select data at random according to the ratio of conformity degree between potential contents $\vec{PT}_m$ and potential interests $\vec{PU}_n$. We regard the above as a set and repeat it for C times. As a result, we evaluate whether the title model $\vec{T}_m$ could reproduce the potential contents $\vec{T}_m$. Here, we assume cosine of an angle be produced by $\vec{T}_m$ and $\vec{PT}_m$ as reproduction degree of title model $\vec{T}_m$ from potential contents $\vec{PT}_m$ as the following.

$$cos\theta = \frac{\vec{T}_m\cdot\vec{PT}_m}{|\vec{T}_m||\vec{PT}_m|}$$

We also assume the cosine of an angle be produced by $\vec{PU}_n$ and $\vec{PT}_m$ as conformity degree between potential interests and potential contents, and we use the p_{th} power of it in order that a user selects data as the following.

$$cos^p\theta = \left\{\frac{\vec{PT}_m\cdot\vec{PU}_n}{|\vec{PT}_m||\vec{PU}_n|}\right\}^p$$

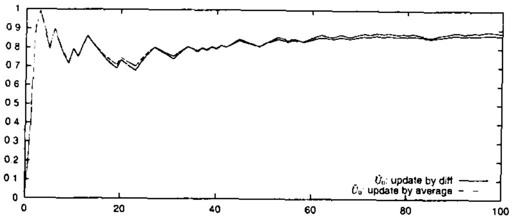


Figure 6: transition of reproduction degree of $\vec{U}_0$

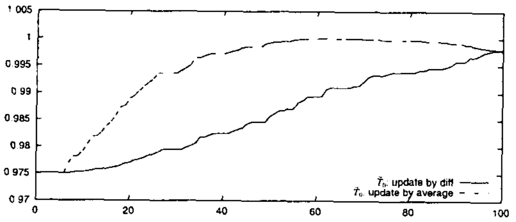


Figure 7: transition of reproduction degree of $\vec{T}_0$

As p is large, difference among $cos^p\theta$ corresponding to θ is large, and partiality will be large also. It means we could define amplitude of partiality with p in simulations.

3.2 Result

At first, we repeat 100 times with condition of the amount of data $M = 100$, the number of user $N = 100$, directions of genre $G = 2$, and degree of partiality $p = 1$, and renovate title models and user models by (3) and (4). We suppose $\alpha_U = \alpha_T = 0.01$ for (3) here. Fig.6 and Fig.7 shows the result of transitions of reproduction degree of $\vec{U}_0$ and $\vec{T}_0$ per times respectively.

Next, we change partiality from 0 to 5 and repeat 1000 times, and examin the difference of reproduction degree from difference of partiality. Fig.8 and Fig.9 shows the result of transition of the average of reproduction degree of title model per times renovated with (3) and (4) respectively.

3.3 Evaluation

From the above results of simulations, we could reproduce the potential values of both user models and title models to a certain extent by the method based on both formulas (3) and (4) till about 200 times repeat of use. About average of title models, though there are diffrences from p, we confirm that reproduction degrees, which is about 0.88 at first, gradually get higher and reach to about 0.98 at 200 times.

Hence, we could get relatively higher reproduction degree in certain repeat times by both methods with formulas (3) and (4) according to the times. That is we could create appropriate indexes expressing contents of different data from these title models based on the algorithm we proposed.

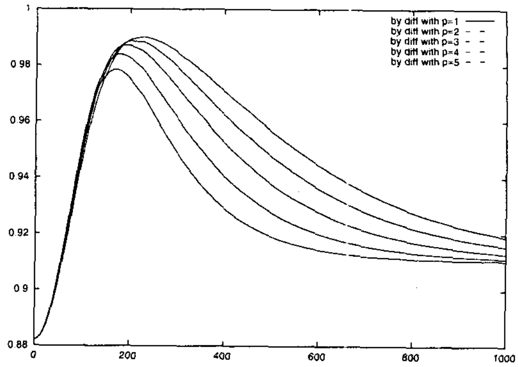


Figure 8: average of reproduction degree of titile model by formulas (3)

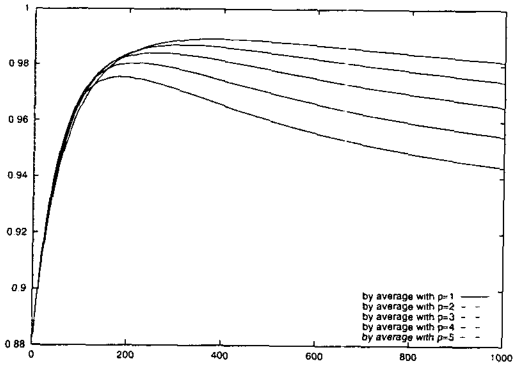


Figure 9: average of reproduction degree of titile model by formulas (4)

However, when we repeat access much more, reproduction degree begins going down by both methods. Especial by a method with formulas (3), it is conspicuous and it converge at about 0.91. By a method with formulas (4), even though it could not be found definitely from Fig.9, we confirmed it also converge at about 0.91 by raising repeat times to 10000 in other experiment. That is to say we could not create appropriate indexes from methods we proposed as they are when access is such many times.

4 Related Works

There exist some reseaches on the methods to find contents or its impression of each data in images and musical scores [1][2][3][7]. These data has comparatively less amount of data in each data, it is easier to get which patterns they contains by frequency analysis, and relations between these patterns and impressions have already been reseached well. However, for video data or audio data, each of which contains large amount of data and they have complex contents, it is harder to analyze whole data by frequency analysis. So, we could neither find their contains nor their directions by these methods.

Another reseach is on the mothod to create a meaning space of each word in some sentences which accompany to pictures and are expressing contents of them [8]. However we could retrieve pictures by these words, we could neither retrieve by contents of data themselves nor directions of them.

In contrast to the aboves, our reseach is not on the mothods to analyze data themselves nor information accompanying with data, it is on the method to analyze users' sensibility from there histories of access to database and presume directions of contents of data they have accessed. With our method, even though we could not find accurate directions, we could find rough directions of different data.

There are several researches on the method to presume users' interest from their histories of access to data on likes of WWW in order to help to retrieve data efficiently or to provide data related to their interest [5][6]. However, in order to presume users' interest, data should have been appropriately categorized in advance. Another reseach on the method to presume correlation among data like HTML documents by analysis of users' histories of access to them [4]. But, is is difficult to create indexes for retrieval only from these correlation among data.

In contrast, with our proposal, we could presume not only users' interest but also potential directions of contents of data from their interests, and also create indexes for retrieval from them.

5 Conclusion

In this paper, we propose an semi-automatic and dynamic indexing algorithm in order to create indexes of data such as multimedia data, which are difficult to be create express indexes for retrieval.

In this algorithm, we categorize data roughly in advance, presume users' interests from their histories of access to database using partiality of use of data caused from their interests, presume directions of contents of data from the histories who accessed them, and renovate these presumptions successively. Base on the algorithm, we define two sort of recurrence formulas in order to renovate an update both vector models showing directions of interets of each user and directrions of contents of each data.

We simulated use of database in roder to confirm effectiveness of our proposals. As a result, we could reproduct potential directions to a certain extent in several times of use. However, as repeating of use much more, values of these vectors would be leveled off. And cosine be producted by two vectors, which we used for calculation of conformity and reproduction degree, does not show the difference in magnitude between vectors. We should define some formulas to calcurate conformity and reproduction degree which concern both directions and magunitude.

In future, we will discuss the aboves, simulate in the case of more than three genres and in the case of new users or new data are added at the time of using database, and revise our models. And also, we will build concrete database and evaluate the model by practical experiment.

References

[1] Manabu Fukuda, Kaoru Sugita, and Yoshitaka Shibata. Perceptional retrieving method for distributed design image database system. *Trans. IPS Japan*, 39(2):158–169, 1998.

[2] Shouji Harada, Yukihiro Itoh, and Hiromasa Nakatani. On constructing shape feature space for interpreting subjective expression. *Trans. IPS Japan*, 40(5):2356–2366, 1999. (in Japanese).

[3] Shigekazu Ishihara, Keiko Ishihara, and Matsuo Nagamachi. Analysis of individual differences in kansei evaluation data based on cluster analysis. *KANSEI Engineering International*, 1(1):49–58, 1999.

[4] Kazuhiro Kazama, Shin ya Sato, Susumu Shimizu, and Takashi Kambayashi. Html document correlation analysis by user's behavior in world wide web navigation. *Trans. IPS Japan*, 40(5):2450–2459, 1999. (in Japanese).

[5] Yoshitaka Kuwata and Masashi Yatsu. An automated follow-up service for technical support help desk. *Trans. IPS Japan*, 40(11):3896–3905, 1999. (in Japanese).

[6] Roberto Okada, Eun-Seok Lee, Tetsuo Kinoshita, and Norio Shiratori. A method for personalized web searching with hierarchical document clustering. *Trans. IPS Japan*, 39(4):867–877, 1998.

[7] Akira Sato, Kouhei Kikuchu, and Hajime Kitakami. A creation method of an affective value of musical piece for impression search. *IPSJ SIG Notes 99-DBS-118*, pages 57–64, 1999. (in Japanese).

[8] Naofumi Yoshida, Yasushi Kiyoki, and Takashi Kitagawa. An implementation method of a media information retrieval system with semantic associative search function. *Trans. IPS Japan*, 39(4):911–922, 1998. (in Japanese).