Expanding User Interests by Recommending Innovative Blog Entries

Makoto Nakatsuji[†]*, Makoto Yoshida, and Miki Hirano*

Abstract

We describe a method of extracting user interests automatically by analyzing entries in users' blogs and present a new style of recommendation based on innovation detection that provides concepts similar to those in a user's profile as well as concepts that are new to the user. We classify user blog entries into service domain ontologies and extract interest ontologies that express a user's interests semantically as a hierarchy of classes according to interest weight. We also present Semantic-Driven Collaborative Filtering, which is based on similarity measurements between ontologies considering the interest weight assigned to each class and instance. We detect innovative blog entries that include concepts that the user has not thought about in the past based on an analysis of other user's ontologies with high similarities to that of the user.

By introducing class characteristics about those instances, our technique can offer more accurate recommendations than previous collaborative filtering for users with a small number of instances in their user profiles. Results obtained with an online experimental service for recommending innovative blog entries to users on an actual blog portal confirmed the effectiveness of innovation detection.

1. Introduction

Consumer-generated media (CGM) such as blogs and social networking services (SNSs) are becoming more popular for publishing and discussing shared interests among users. In some CGM services like last.fm [1], a music SNS, users disclose their interests not only actively by writing entries in their blogs but also passively through listening to music. Information-sharing systems of these types could enable users to expand their interests by browsing collected blog entries or listening histories published by other users. Online recommendations made by analyzing these published user interests are essential for service providers to increase the sales of their contents. We can automatically create user profiles about their interests from listening histories by checking how many times a user listens to the same songs. By applying collaborative filtering (CF) [2]–[4] to measure the similarity between user profiles, we can recommend several songs to users comparatively easily. However, CF is apt to recommend information resources that are the same as the concepts in the user's profile. Furthermore, there is a major problem called the sparsity problem: this occurs when there is not enough data in the user profile to measure the similarity among users.

In our research [5], we define an *innovation* as a new concept that is likely to be interesting to the user even though it is not included in his or her user profile. We try to expand the scope of user interests significantly by recommending innovative information. In particular, we apply innovation detection to blogs (web logs) because they have become a popular publishing format and targets for searching for information that could expand user interests.

To achieve this purpose, we first construct a user profile automatically as a user-interest ontology, which is a class hierarchy of user interests with interest weights. We present a method of automatically extracting an interest ontology with an interest weight assigned to each class and instance. Bloggers are apt to describe their interests about topics in several service domains freely. Thus, we use blog entries to

† NTT Network Service Systems Laboratories
Musashino-shi, 180-8585 Japan
Email: nakatsuji.makoto@lab.ntt.co.jp

specify user interests by introducing a template ontology, which is a domain ontology of one service. We classify user entries according to a template ontology and remove classification mistakes by using class characteristics and the continuity of descriptions about user interests using a top-down approach. We also provide a user interface to update the interest ontology using a bottom-up approach.

Next, we present a method for measuring the similarity of interest ontologies considering the degree of interest agreement for each class and instance. Then, we present Semantic-Driven CF by applying our measurement of similarities between user-interest ontologies to the previous CF with Pearson correlation coefficients [2]–[4]. Thus, we detect innovative blog entries for each user by analyzing other users' ontologies that have a high degree of similarity to that of the user. We apply our techniques to help users create blog communities by browsing innovative blog entries that include information that is unknown to users and has a high probability of being interesting. We executed offline experiments based on a large number of blog entries (1,600,000 entries of 55,000 users) on an actual blog portal Doblog [6], which is one of the biggest blog portals in Japan. In December 2006, it had about 55,000 users. For these experiments, we used a music template ontology with 114 classes and 4300 instances. We confirmed that our automatic ontology extraction and innovation detection have potential for creating user-oriented blog communities that correspond to user interests.

The specific contributions of this article are as follows.

1. We impose a class taxonomy of instances of user interests in user profiles. Thus, we improve the accuracy of recommendations, especially for users with a small number of instances in their user profiles, by considering the similarities between instances and user-interest ontologies. This is important because many content providers have to handle a lot of users with a small number of entries in their profiles, such as light users who have just begun to use the service. We confirmed the effectiveness of our technique by using extracted results of interest ontologies to compare Semantic-Driven CF with the previous CF.

2. The evaluation was done in two steps. The first step was an offline evaluation for extracting interest ontologies and comparing the previous CF with our Semantic-Driven CF. The second step was an online evaluation for analyzing user reactions to recommendations based on innovation detection. (We ran an experimental service called DoblogMusic for Doblog users from August to December 2006.) Most previous studies evaluated their recommendation techniques by using only offline experimental results. However, it is very important to analyze user reactions to recommendations to check whether the recommendation techniques are actually effective. By analyzing the change in the frequency of user accesses to innovative instances of our recommendations, we confirmed the effectiveness of our ontology extraction and innovation detection. Through an evaluation of the frequency of comments between users who got to know each other through our online recommendations, we also found that recommendations based on innovation detection facilitated the creation of new communities.

2. Interest-ontology extraction

Here, we explain the algorithm for generating an interest ontology by analyzing the distribution of user interests, as shown in **Fig. 1**.

(1) First, we make index files for all blog entries collected through the ping server. Here, we assume that each collected blog entry has a unique user ID.

(2) Second, we classify all collected blog entries into a template ontology. This is knowledge about the content taxonomy such as the hierarchy of genres of music artists; it is constructed by the designers of service providers to manage and promote their contents. For example, consider the template ontology in Fig. 1. Here, the instances are the names of musical bands (e.g., Stone Roses), while classes are the names of musical genres (e.g., Madchester, a term coined to describe the alternative music scene in Manchester, UK, in the late 1980s to early 1990s). We classify a blog entry into the instance "Happy Mondays" of class "Madchester" when the character string "Happy Mondays" is included in the body of a blog entry. However, our algorithm mistakenly classifies blog entries about agricultural farms into the instance "Farm" of class "Madchester" because the only meaning it knows for farm is the name of a British band. To filter out mistakes caused by words with multiple meanings, we make use of characteristics such as class relationships in ontologies and the durability of user interests in a blog.

1. Adjacent classes have similar characteristics. Instances of those classes also have similar characteristics.

2. User interests continue for a certain period and describe an interest lasting two or more days.

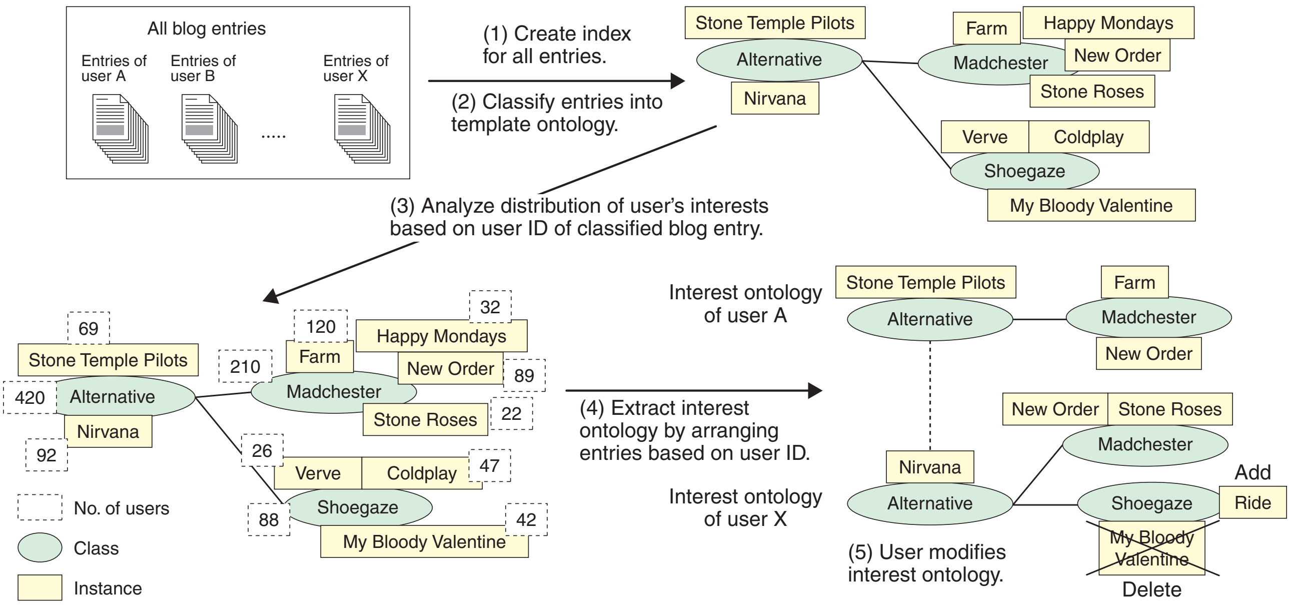


Fig. 1. Procedure for generating interest ontologies.

(3) Then, we measure the number of users interested in each instance of *Ce*, which is one of the end classes in the template ontology. We calculate the number of users interested in class *Ce* by obtaining the number of users interested in all instances in *Ce* and in class *Ce*. Thus, the distribution of interested users in the domain can be measured by recurrently counting the number of users from *Ce* to the root class *Cr*.

(4) Next, by extracting only the classification results for a certain user ID from all classification results, we can extract an interest ontology for this user ID. In Fig. 1, we can extract an interest ontology of user A when the blog entries of this user describe instances of "Stone Temple Pilots", "New Order", and "Farm".

3. Interest-weight-based similarity measurement

We now explain our similarity measurement using **Fig. 2**. First, we define the terminology. We denote the interest ontology of users A and B by O_A and O_B. We use two topologies: T_1 is composed of a class and subclass relationship and T_2 is composed of a class and instance relationship. Furthermore, we define common classes and common instances of both ontologies as *Ci* and *Ii*, respectively. In particular, we define a common class set that formalizes topology T_1 as $C(T_1)$ and a common class set that formalizes topology T_2 as $C(T_2)$. For example, in Fig. 2, $C(T_1)$ has common classes a1 and b2, and $C(T_2)$ has common classes b2, b3, and c4. We also denote the degree of interest agreement of common instance *Ii* as *I*(*Ii*), that of common class *Ci* as *I*(*Ci*), and that of common topology created by common class *Ci* as *It*(*Ci*).

Maedche and Staab [7] calculated the similarity between ontologies considering the degree of similarity between class topologies T_1. We extend this by additionally taking the following ideas from the viewpoint of creating user-interest-based communities.

1. We evaluate the degree of interest agreement between *Ci* s and *Ii* s as a smaller value of interest weight. This idea is for filtering users who only enumerate a lot of instances in an entry and for creating a community among users who have similar or larger interest weight values from the viewpoint of each user.

2. We treat topologies T_1 and T_2 separately because we consider that T_1 reflects the width and depth of a user's interests while T_2 reflects the objects in which users are interested.

3. We achieve low computational complexity by generating the class schema of user-interest ontologies according to that of template ontologies. For ontology mapping, it is important to use a large-scale dataset of the blog community such as that used in our experiments described in Sections 4 and 5.

The procedure of our method is given below.

(1) We analyze classes common to O_A and O_B and extract common classes that belong to $C(T_1)$ and

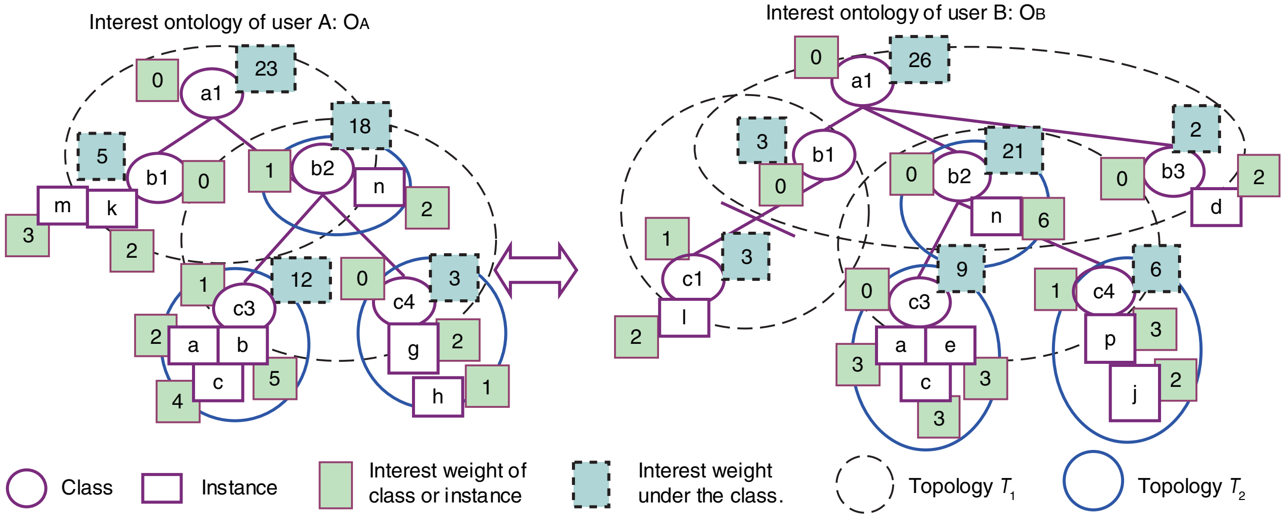

Interest ontology of user A: OA
Interest ontology of user B: OB
a1
b1
b2
b3
c1
c3
c4
m
k
n
a
b
c
e
g
h
l
d
p
j
0
23
26
18
21
5
3
12
9
6
2
1
4
Class
Instance
Interest weight of class or instance
Interest weight under the class.
Topology T_1
Topology T_2

Fig. 2. Measuring similarity based on the degree of interest agreement.

$C(T_2)$.

(2) If common class *Ci* has common instance *Ii* between ontologies, we assign the smaller value of the interest weight of common instances *Ii* to *I*(*Ii*). For example, *I*(*a*) is 2.

(3) Similarly, we assign the smaller value of the interest weight of common class *Ci* to *I*(*Ci*). For example, *I*(*b*1) is 3.

(4) We define product sets of subclasses of *Ci*, which are common to a class set, as *N*(*Ci*), and we define the set union of subclasses of *Ci* among $Ci \in C(T_1)$ as *U*(*Ci*).

For example, $N(a1) = \{b1, b2\}$ and $U(a1) = \{b1, b2, b3\}$. Then, we express *It*(*Ci*) as $\frac{\sum_{Cj\in N(Ci)} I(Cj)}{|U(Ci)|}$. For example, $I_t(a1)$ is given by (3+18+0)/3 = 7. Thus, we obtain degree of interest agreement $S(T_1)$ of $C(T_1)$ as $\sum_{Ci\in C(T1)} I_t(Ci)$. In Fig. 2, S($T_1$) = (3+18+0)/3+(9+3)/2.

(5) We also define an instance set of *Ci* in ontology O_A as $I_A(Ci)$, and we define an instance set of *Ci* in ontology O_B as $I_B(Ci)$ among $Ci \in C(T_2)$. Then, we express *It*(*Ci*) as $\frac{\sum_{Ii\in Ci} I(Ii)}{|I_A(C_i) \cup I_B(C_i)|}$. For example, $I_t(C3)$ is given by ((2+0+3+0)/4) = 5/4. Thus, we assign the degree of interest agreement $S(T_2)$ of $C(T_2)$ as $\sum_{Ci\in C(T2)} I_t(Ci)$. In Fig. 2, S($T_2$) = 2/1+5/4+0.

(6) By using evaluation function f(X) corresponding to the relative degree of importance of a topology, we finally assign the similarity score between ontologies $S_o(AB)$ as $S(T_1)+f(S(T_2))$.

4. Results of offline experiment

Here, we present the results of offline experiments and simulation studies that show the performance of interest ontology extraction and innovative blog-entry detection. We also compared the predictions of Semantic-Driven CF and the previous CF.

We evaluated the performance of our methods using the large-scale blog portal Doblog. We also used the template ontology of the music domain (Fig. 1), which was created by referring to public information on goo music [8], a web portal containing music artist genre information. Our experimental template ontology contained 114 classes as genres and 4300 artists as instances. Each class and instance could have two or more name attributes. For example, the instance “R.E.M.” has the name attributes of “R. E.M.” and “REM”. In total, we gave 7600 name attributes to 4300 instances.

To evaluate the accuracy, we defined correct answers as blog entries that have descriptions of classified classes or instances and evaluated the generated interest ontology by using precision and recall in the classified results. In this paper, precision means the proportion of correct answers in classified results and recall means that of correct answers in all blog entries. When the recall was high, extracted interest ontologies covered user interests better. However, when the precision was lower, created interest ontologies included classification mistakes, and innovation detection for the user was unreliable. Thus, it is essential to achieve high precision. In the evaluation, we applied filtering algorithms to instances consisting of only one word such as “police”, because we considered a single word to have a high possibility of having several meanings. To generate index files of blog entries, we used Namazu [9].

4.1 Measuring the performance of the extracted interest ontology

We evaluated the accuracy of our interest-ontology extraction technique by checking one quarter of the classified blog entries selected at random. As listed in **Table 1**, the achieved precision was higher than 90% with a high recall of 80%. Thus, our filtering algorithm is effective for generating suitable user-interest ontologies.

4.2 Comparison of Semantic-Driven CF and the previous CF

We compared the recommendation results provided by Semantic-Driven CF with the interest-ontology measurement and the previous CF with the Pearson correlation coefficient. In the evaluation, we used the dataset based on the extraction results of user-interest ontologies in Section 2: 54,933 items of interest ontologies with interest weights of 3632 users. To evaluate the performance of the prediction results, the mean absolute error (MAE) has been used in several studies on recommendation systems [2]. MAE is expressed by

$$\frac{MAE = \sum|Pj\text{-}Dj|}{n},$$

where Pj means the prediction value of a user having item j in his/her user profile, Dj means the actual value of the interest weight for a user having item j in his/her user profile, and n means the total number of predicted items. When the value of MAE is smaller, we consider that the prediction performance is better.

In the evaluation, we divided the dataset DATA into two datasets: DATA(test) and DATA(predict). We used DATA(test) in calculating the predicted values of DATA(predict). We prepared five types of DATA(test) with different ratios of DATA(test) to DATA: 16.6%, 33.3%, 50.0%, 66.6%, and 83.3%. A graph of the performance of MAE is shown in **Fig. 3**. Our technique achieved better results than CF when the ratio of DATA(test) to DATA was small, but worse ones when it was larger.

Table 1. Experimental results for our ontology extraction.

Precision	Recall
94.9%	80.3%

We then divided the 3632 users into two user groups: user group U_A composed of users interested in many different artists and user group U_B composed of users who do not belong to user group U_A. We calculated the number of users in these two groups to satisfy

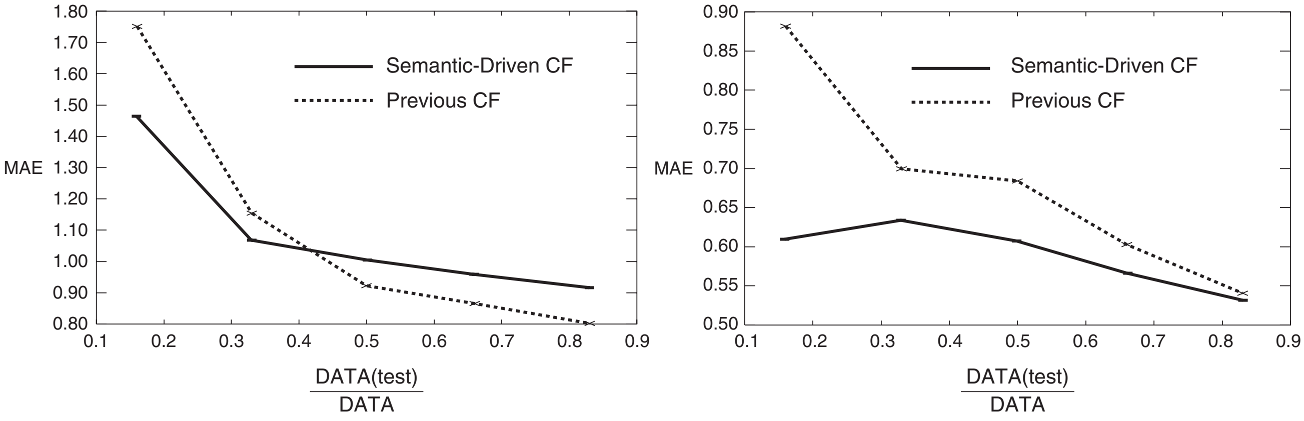
$$\int_{i \in U_A} N(i) = \int_{j \in U_B} N(j), \qquad (1)$$

where the number of artists for user i is N(i). As a result, we obtained $U_A = 680$ and $U_B = 2952$. U_B was much larger than U_A. Graphs of the MAE performance of user group U_B are shown in Fig. 3. For U_B, Semantic-Driven CF outperformed the previous CF regardless of the value of DATA(test)/DATA. This is because Semantic-Driven CF introduces a class taxonomy about instances in user profiles for classifying those instances, especially for users with few entries in their profiles. On the other hand, the previous CF was better than Semantic-Driven CF for U_A. We think that Semantic-Driven CF predicts a wider range of instances than the previous CF by using not only similarities between instances but also the class taxonomy of these instances. Thus, the predictions of Semantic-Driven CF for U_A were worse than those of the previous CF because they offered more ambiguous results. However, we think that these ambiguous characteristics derived from the class taxonomy create the potential for detecting innovation for the user.

According to these results, our technique is effective, especially for users with few entries in their profiles. We think that these results are good considering that U_B is much larger than U_A in typical service systems. However, the purpose of our research is to predict and expand user interests by recommending innovative instances. In the next section, we evaluate innovation detection through an online experimental service.

5. Conclusion

We presented Semantic-Driven CF for expanding user interests significantly by measuring the similarity between user-interest ontologies. By using class characteristics of instances in a user profile, we can detect information that is innovative for users and



(a) Comparison of MAE between Semantic-Driven CF and previous CF.

(b) Comparison of MAE between Semantic-Driven CF and previous CF when focused on users with a small number of entries in thier profiles.

Fig. 3. Comparison of MAE between Semantic-Driven CF and the previous CF.

improve the recommendation results, especially for users with few entries in their profiles. Comparing the effectiveness of Semantic-Driven CF with the previous CF, we found that Semantic-Driven CF achieved good prediction performance for users with few entries in their profiles. Through an online evaluation, we confirmed the effectiveness of innovation detection by checking the access frequency of users that clicked on innovative artists.

Further work is required to study how users control their interest ontologies by providing feedback about their collective knowledge to update class hierarchies of template ontologies constructed by expert designers of ontologies. In addition, we need to evaluate our techniques by comparing them with template ontologies of other domains such as those of fashion or movies.

References

[1] http://www.last.fm/

[2] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of dimensionality reduction in recommender systems—a case study," In ACM WebKDD Workshop, 2000.

[3] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pp. 43–52, 1998.

[4] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proc. of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, North Carolina, ACM, pp. 175–186, 1994.

[5] M. Nakatsuji, Y. Miyoshi, and Y. Otsuka, "Innovation Detection Based on User-Interest Ontology of Blog Community," International Semantic Web Conference (ISWC2006), pp. 515–528, 2006 (http://iswc2006.semanticweb.org/items/Nakatsuji2006dp.pdf).

[6] http://www.doblog.com (in Japanese).

[7] A. Maedche and S. Staab, "Measuring Similarity between Ontologies," Proc. of the European Conference on Knowledge Acquisition and Management, EKAW-2002. Madrid, Spain, LNCS/LNAI 2473, Springer, pp. 251–263, 2002.

[8] http://music.goo.ne.jp/ (in Japanese).

[9] http://www.namazu.org/index.html.en

Makoto Nakatsuji

Emerging Communication Architecture Project, NTT Network Service Systems Laboratories.

He received the B.E. degree in applied mathematics from Kyoto University Faculty of Engineering, Kyoto, and the M.S. degree in systems science from Kyoto University Graduate School of Informatics, Kyoto, in 2001 and 2003, respectively. He joined NTT Network Service Systems Laboratories in 2003. His research interests include Web mining in web 2.0-based systems, semantic-based search systems, and context-aware ubiquitous networks. He received the JSAI SIG Research Award in 2007. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Japanese Society for Artificial Intelligence, and the Database Society of Japan.

Makoto Yoshida

Senior Research Engineer, Emerging Communication Architecture Project, NTT Network Service Systems Laboratories.

He received the B.E. and M.E. degrees in mechanical engineering from Waseda University, Tokyo, in 1992 and 1994, respectively. He joined NTT in 1994. He has more than eight years of experience in marketing and business development in the telecommunications industry. His research interests include service convergence between telecommunications and the Web to create broadband ubiquitous services. Prior to his current position, he was a manager of NTT Advanced Technology Corporation in the USA and played a key role in business development as a liaison between US venture companies and NTT Group.

Miki Hirano

Senior Research Engineer, Supervisor, Group Leader, Emerging Communication Architecture Project, NTT Network Service Systems Laboratories.

She received the B.E., M.E., and Ph.D. degrees in electronics engineering from Kyushu University, Fukuoka, in 1983, 1985, and 2002, respectively. Since joining NTT Musashino Electrical Communication Laboratories in 1985, she has been engaged in the development of switching system architectures, traffic management systems for broadband communication networks, and public ATM switching systems. She is currently involved in R&D of next-generation network systems and new network services. She is a member of IEICE.