

Ontological Data Mining from Blog Entries Focusing on User Interests

**NTT Network Service Systems Laboratories,
NTT Corporation**

Makoto Nakatsuji, Makoto Yoshida and Hiroshi Sunaga

Outline

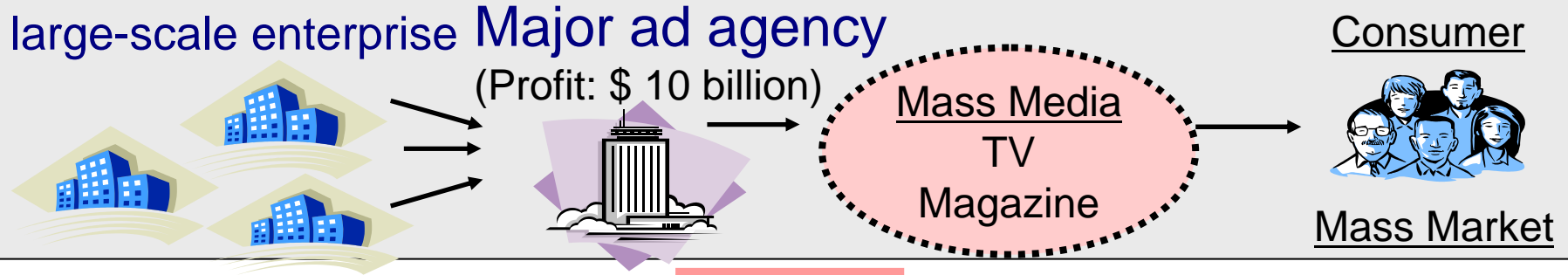
1. Motivation
2. User-Interest Ontology Generation Algorithm
3. Detecting innovative topics by measuring similarity between interest ontologies
4. Experiment and Results
5. Summary and future plan

Motivation

The vision of community business

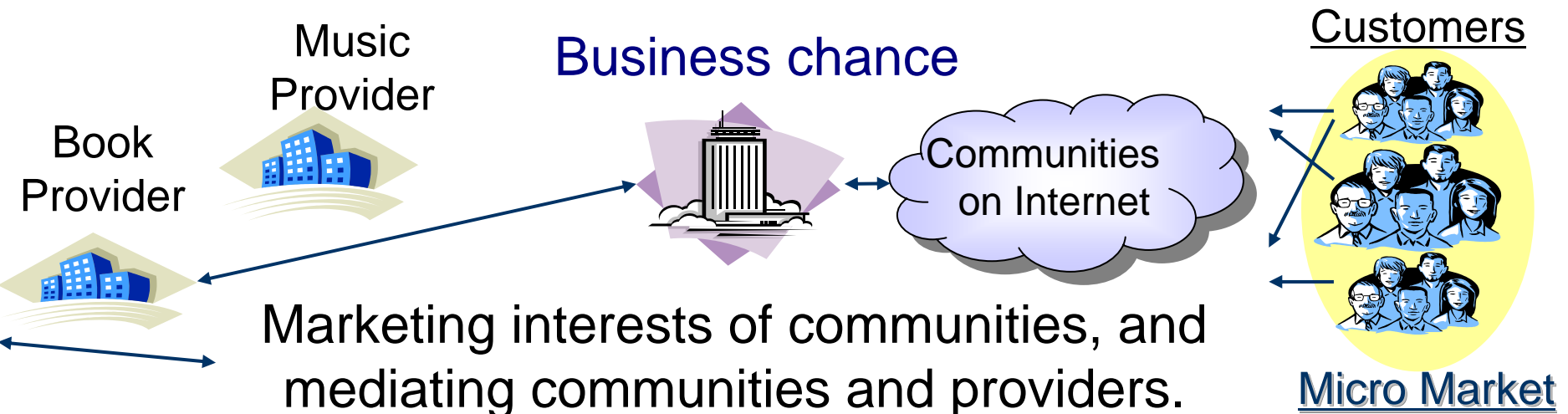
20th century

Mass production and consumption based on mass marketing.



21st century

Micro markets and niche communities are formed on Internet.



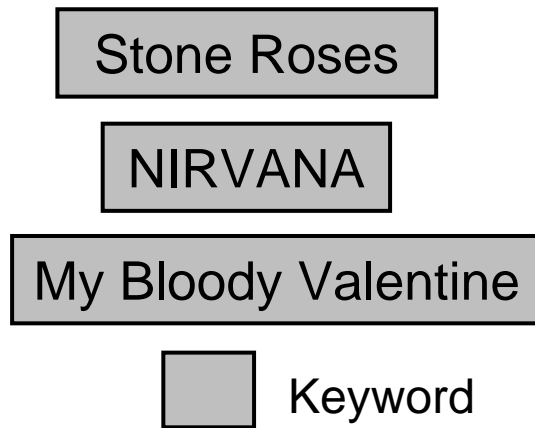
Our research purpose and approach

- Matching users based on their interests and increasing the activity in communities.
 - Extracting *user-interest ontology* that expresses user profile about interests as a hierarchy of classes according to user's interest weight assigned to each class and instance.
 - We consider *instances as topics* and *classes are defined by taxonomy of those topics*.
 - Detecting *innovative topics for a user* that include topics in new classes that seem to be interesting to the user even though they are not included in the user profile.
- ↓
- We try to expand the width of user interests by letting users browse *innovative topics* in other users' blog entries.

Comparison of previous profile and interest ontology

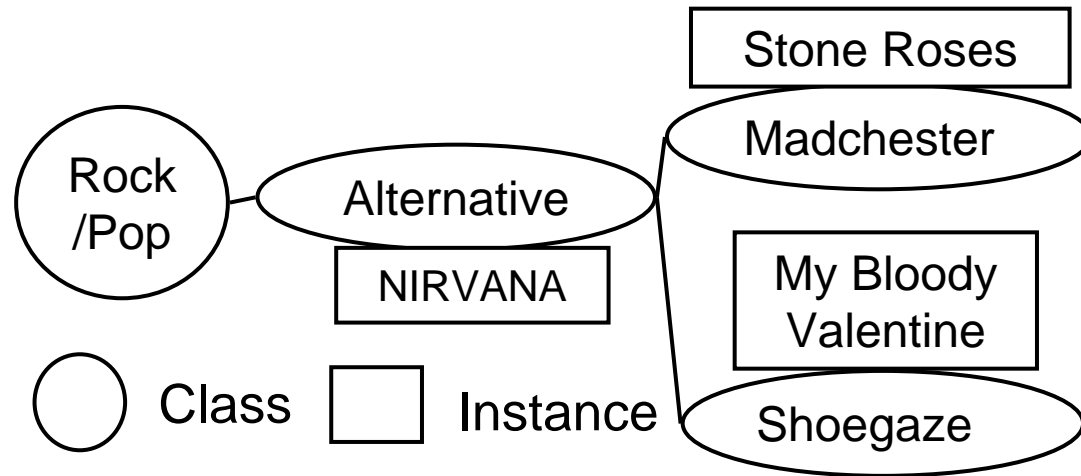
previous user profile

Storing keywords that may be interesting to a user



user interest ontology

Storing topics as instances with information about taxonomy of topics in a service domain about those topics.



Amazon, Lastfm

Simple recommendation

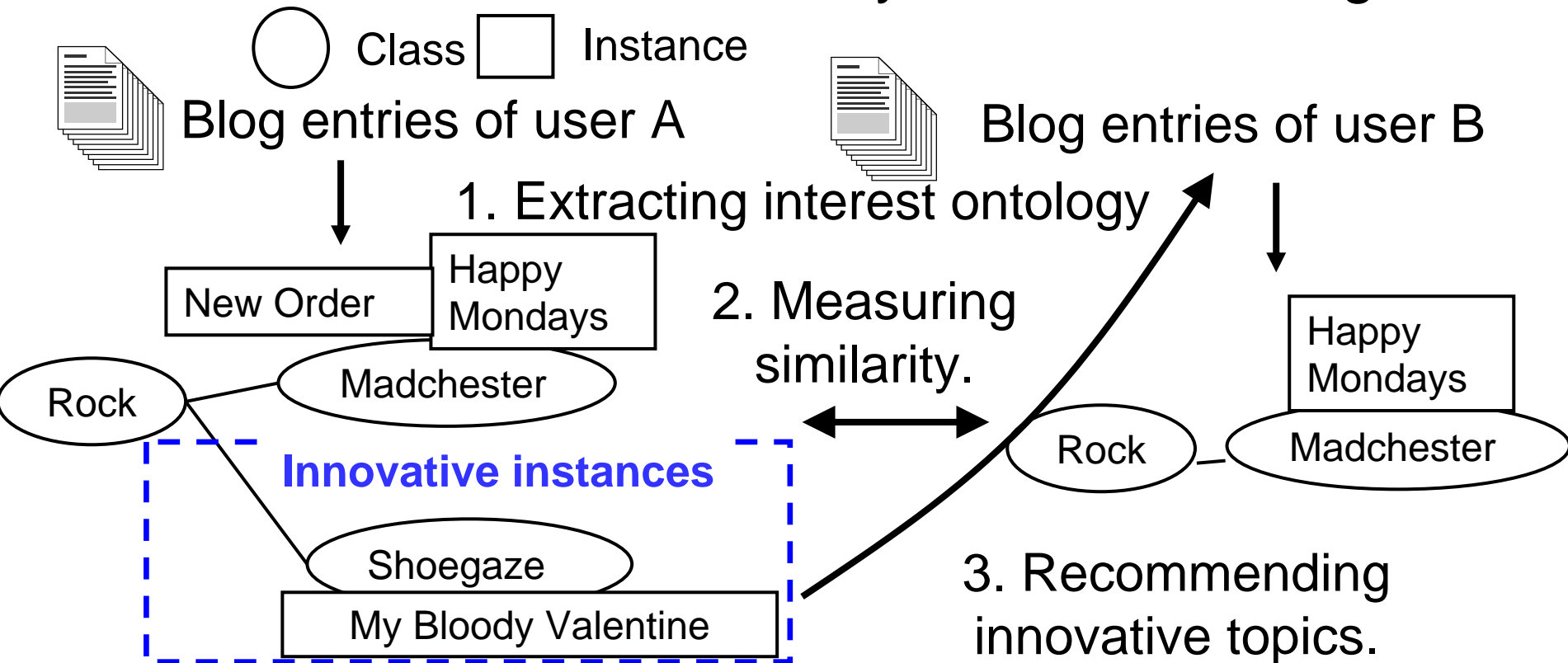
- Keyword base, not using taxonomy of topics.

Semantically rich recommendation

- Detecting innovative instances for a user by analyzing differences of hierarchy of classes.

Our approach for detecting innovative topics

1. Extracting an interest ontology by classifying users' blog entries into *a service domain ontology*.
2. Generating *interest-sharing group* G_u by measuring similarity between ontologies.
3. Detecting innovative instances by analyzing differences of class hierarchy between ontologies.



2. User Interest Ontology Generation Algorithm

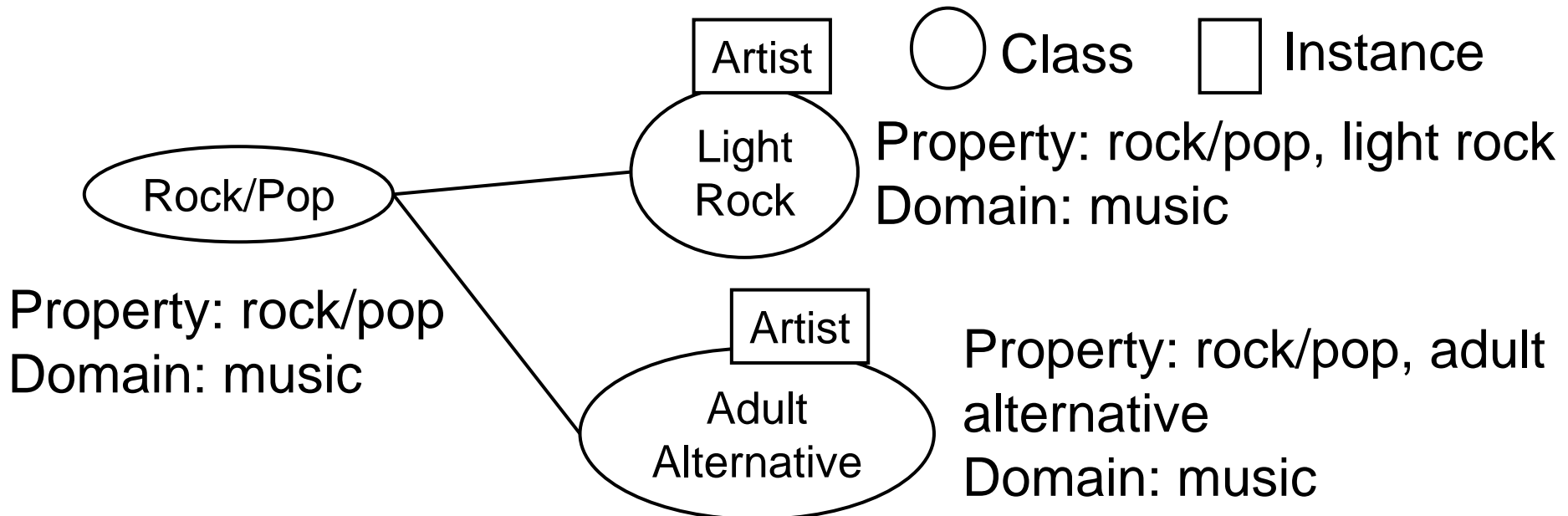
Service domain ontology for creating community

- It defines specification of service for creating communities.
- We use this ontology to generate interest ontologies.
 - Thus, taxonomy of instances in user interest ontology is according to that of a domain ontology.

An example of service domain ontology in music domain.

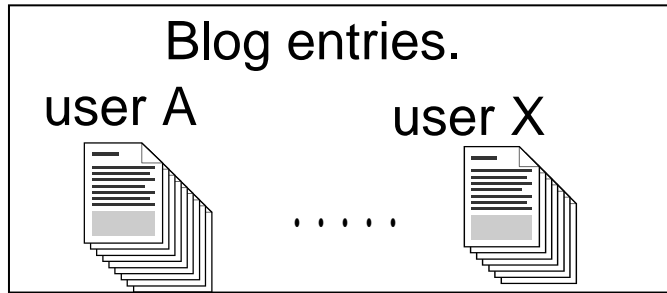
We can make use of topic directory in several services.

- For example, genre hierarchy of AMG or ListenJapan.

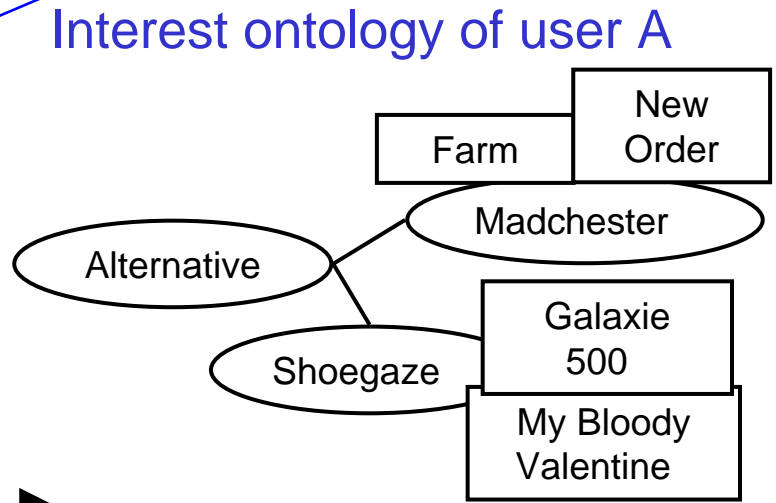
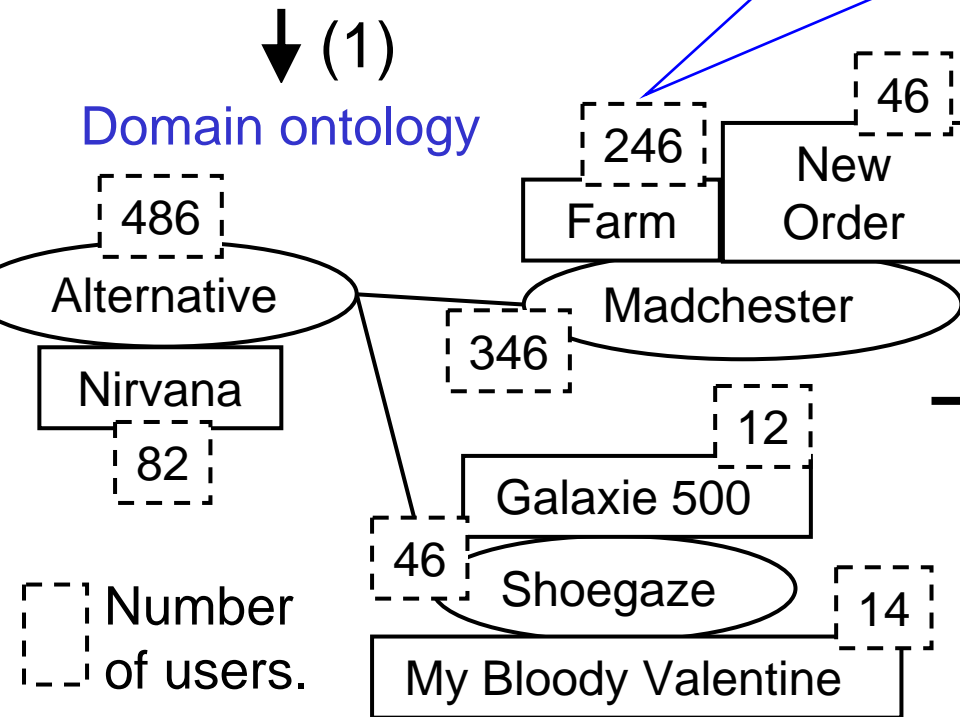


Procedure of generating interest ontologies.

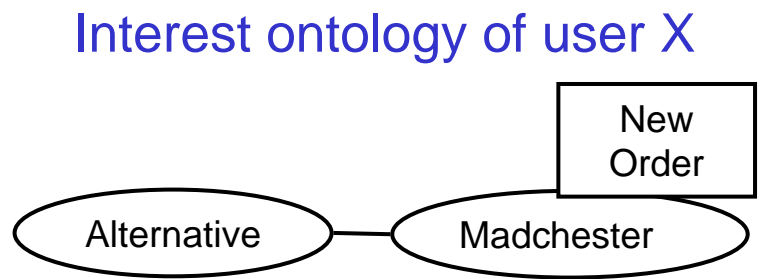
1. Classifying blog entries of users into a *domain ontology*.
2. Filtering classification mistakes caused by multisense words.
3. Extracting a user interest ontology by checking the user ID.



(2) Filtering multisense words such as "Farm" (see Next page)



(3) →

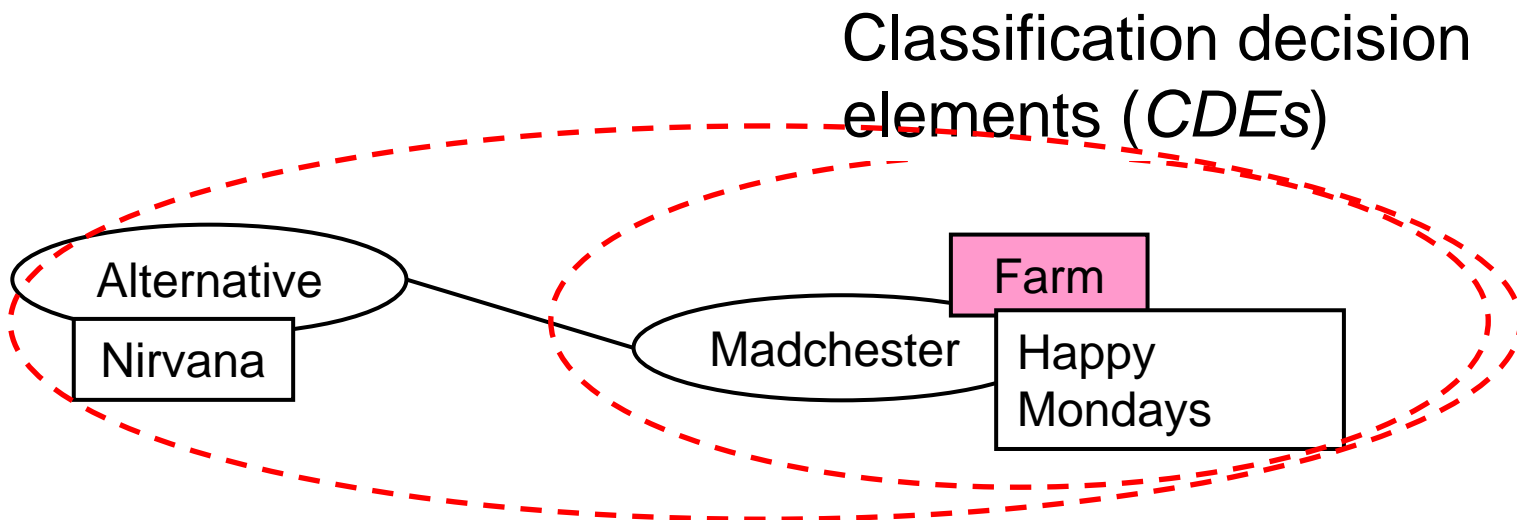


Filtering Algorithm

Filtering mistakes by words with several meanings.

- (1) Using knowledge of taxonomy of domain ontology.
- (2) Using user's continuity of descriptions about same class of interests in several period of time.

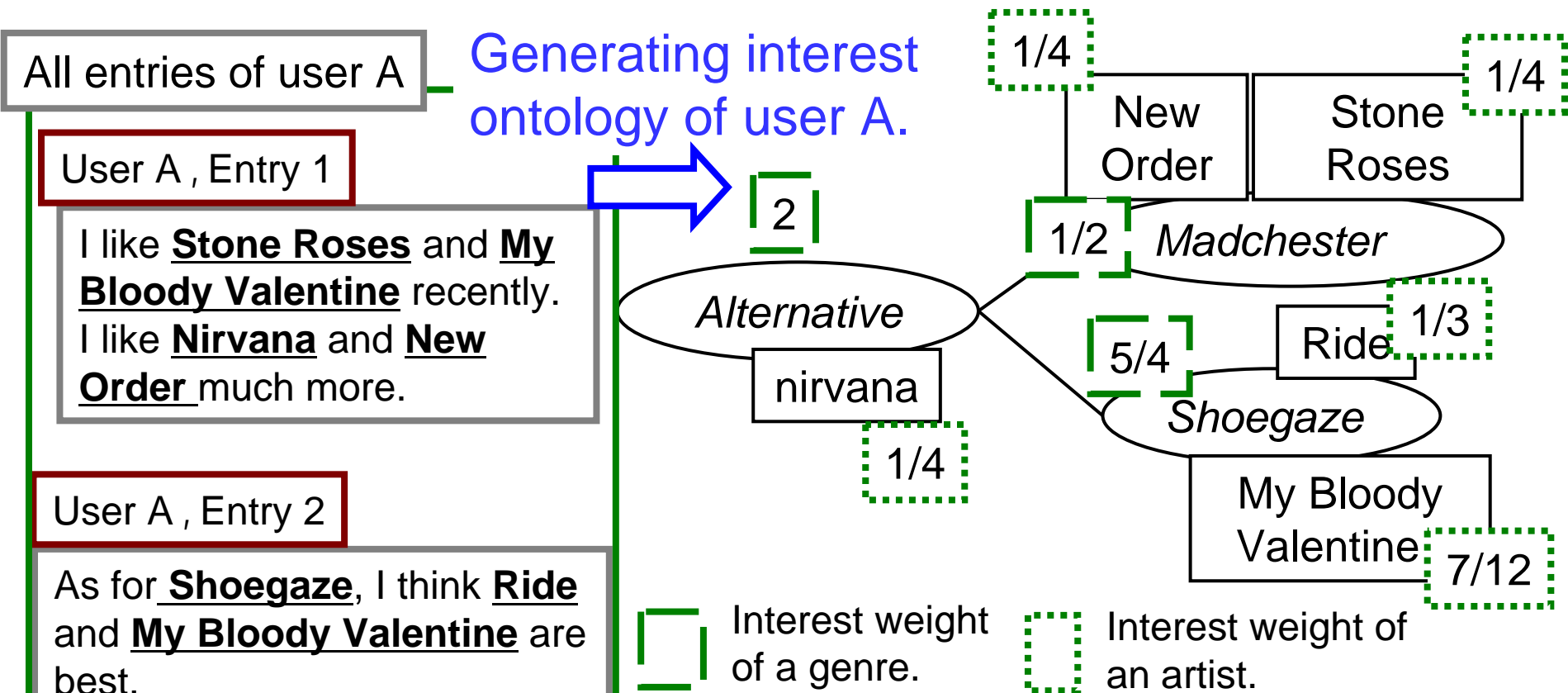
e.g.) Considering word "Farm" as artist "Farm" of "Madchester" genre, if "Farm" and "Happy Mondays" co-occurs in blog entries of a user for a certain period of time.



Introducing interest weight to ontology

A parameter that indicates the degree of interest for each class and instance of interest ontology.

- Interest weight of an entry is 1
- If entry has N types of artists, interest weight of each artist is $1/N$
- Interest weight of instances is reflected in that of the class



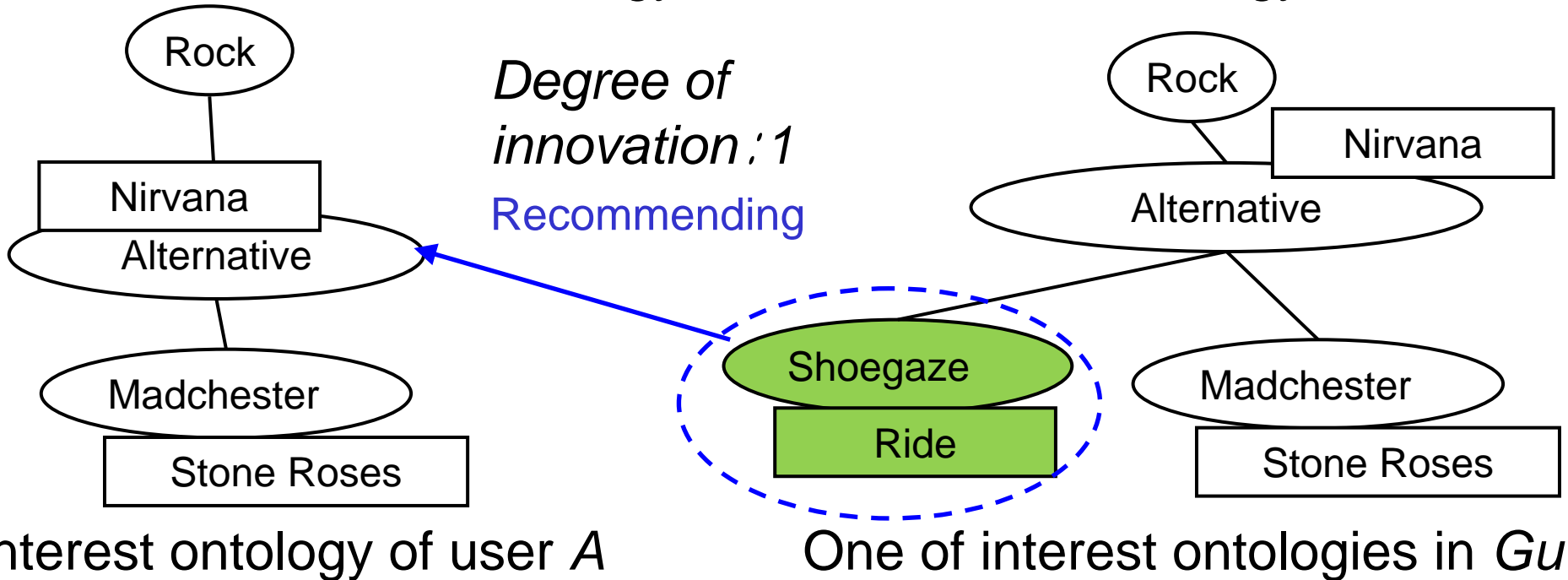
3. Detecting innovative topics by measuring similarity between interest ontologies

- 3.1 Procedure of detecting innovative topics
- 3.2 Similarity measurement algorithm

3.1 Procedure of detecting innovative topics

1. We measure the similarity between user interests based on the degree of interest agreement between interest ontologies.
2. Extracting innovative instances, which user A does not have, even though users in interest-sharing group G_u have those with a high possibility. It is important to set appropriate size of G_u for recommendation. (We evaluate this in experiment.)
3. Recommending instances to user A with degree of innovation.

Degree of innovation indicates how many hops we need to get from different instances of ontology of G_u to class of ontology of user A .



3.2 Similarity Measurement Algorithm

What is the contribution of our similarity measurement algorithm

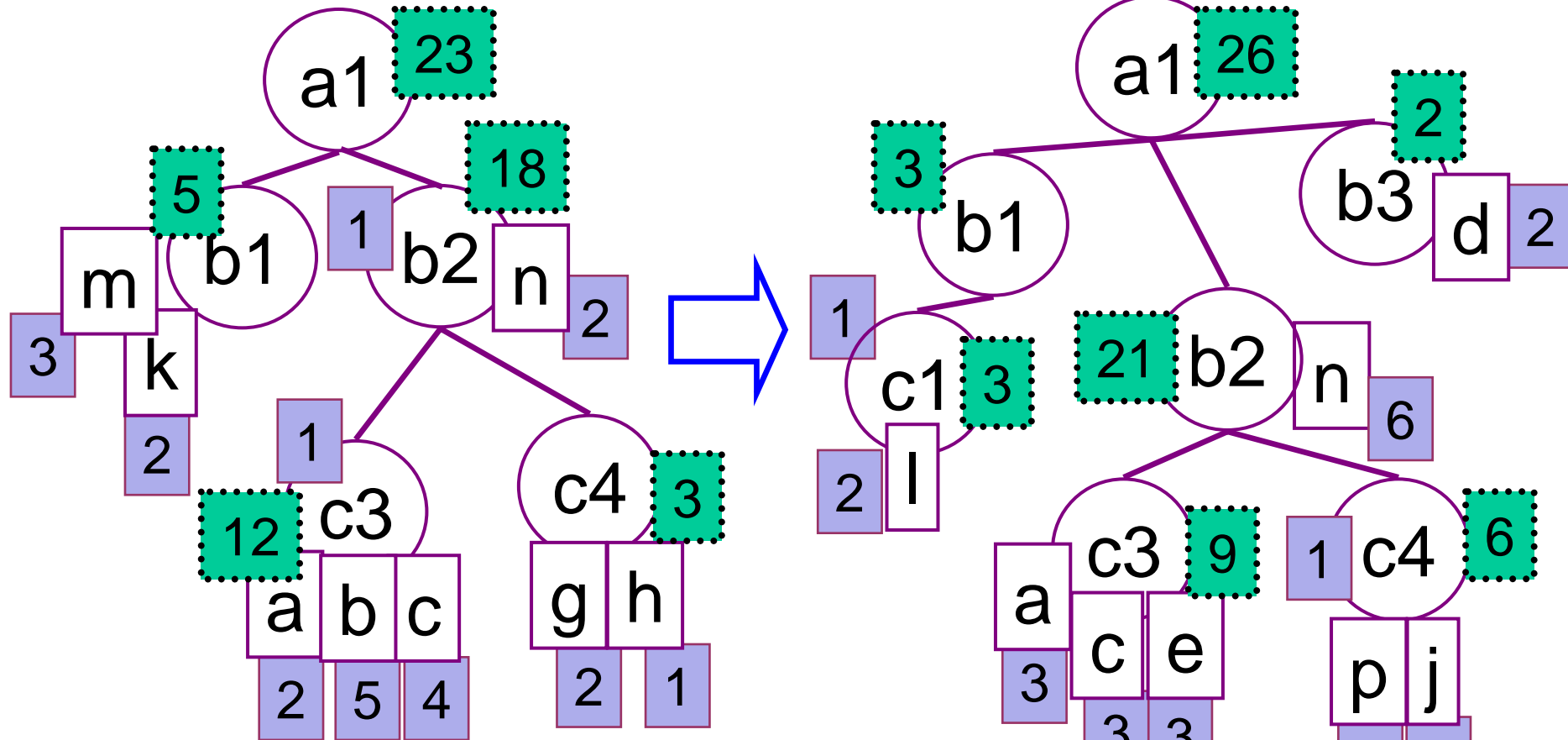
- We measure similarity between user interests by considering the interest agreement between classes and between instances based on personal weights on user-interest ontologies.
- Thus, we can detect suitable innovative topics for each user by setting appropriate size of interest-sharing group G_u .

3.2 Similarity Measurement Algorithm

We explain algorithm by using example below.

Interest ontology of user A

Interest ontology of user B



○ class

□ instance

■ Interest weight of the class or instance



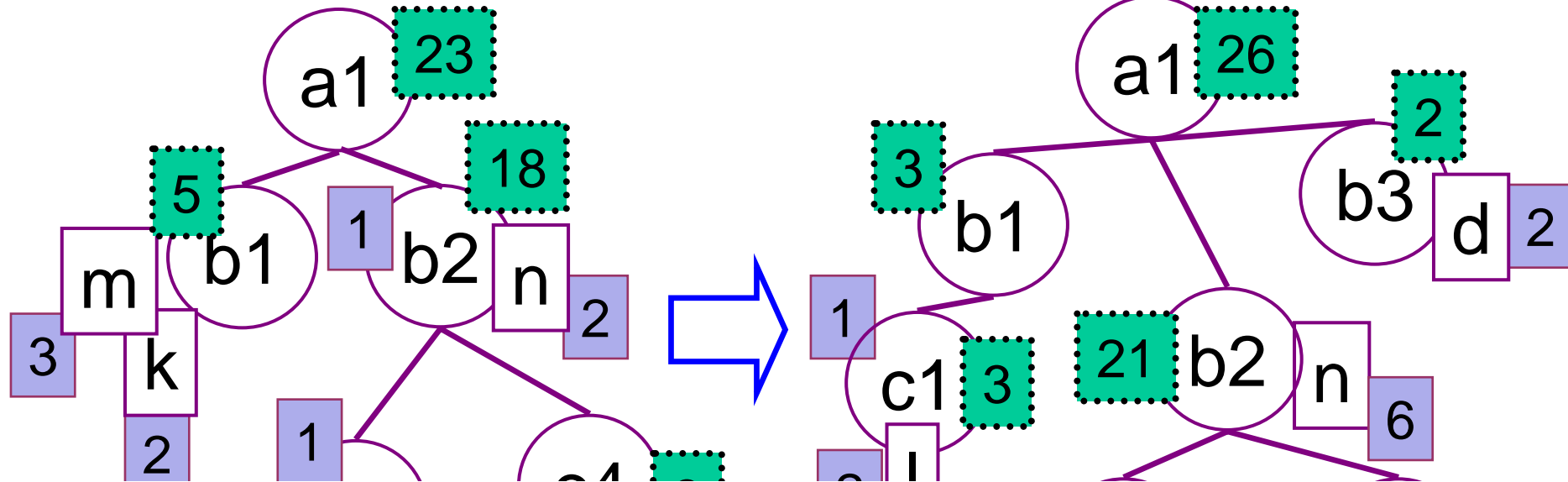
Interest weight under the class

3.2 Similarity Measurement Algorithm

We explain algorithm by using example below.

Interest ontology of user A

Interest ontology of user B



IDEA 1: We achieve a low computational complexity by generating user-interest ontologies according to the taxonomy of topics in a domain ontology.

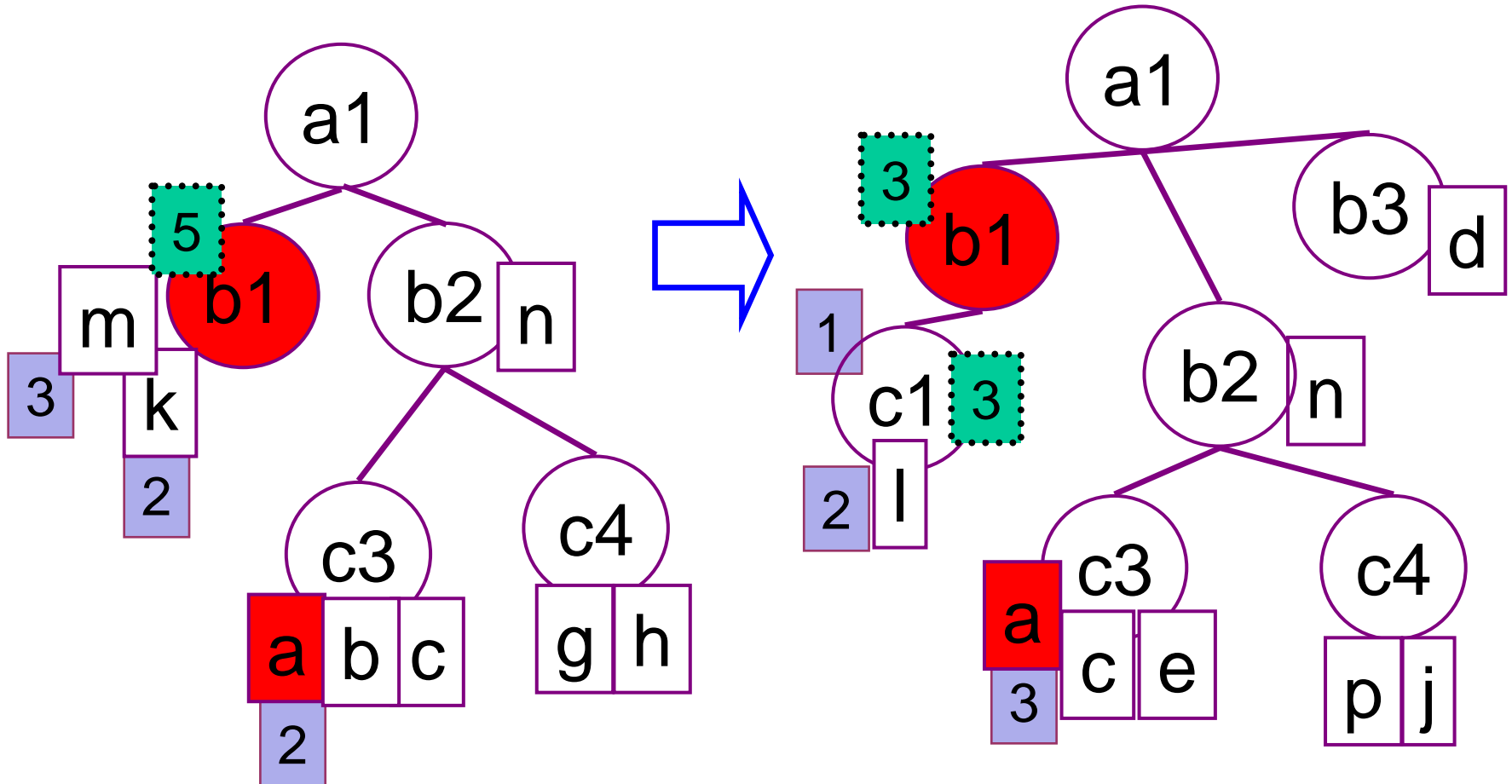
□ instance □ class or instance □ under the class

IDEA 2: We evaluate the degree of interest agreement between classes and instances as a smaller value of interest weight.

- Because we want to create a community among users who have similar or larger interest weight values from the viewpoint of each user.

IDEA2: We evaluate the degree of interest agreement between classes and instances as a smaller value of interest weight.

- the degree of interest agreement of instance *a* is 2
- the degree of interest agreement of class *b1* is 3

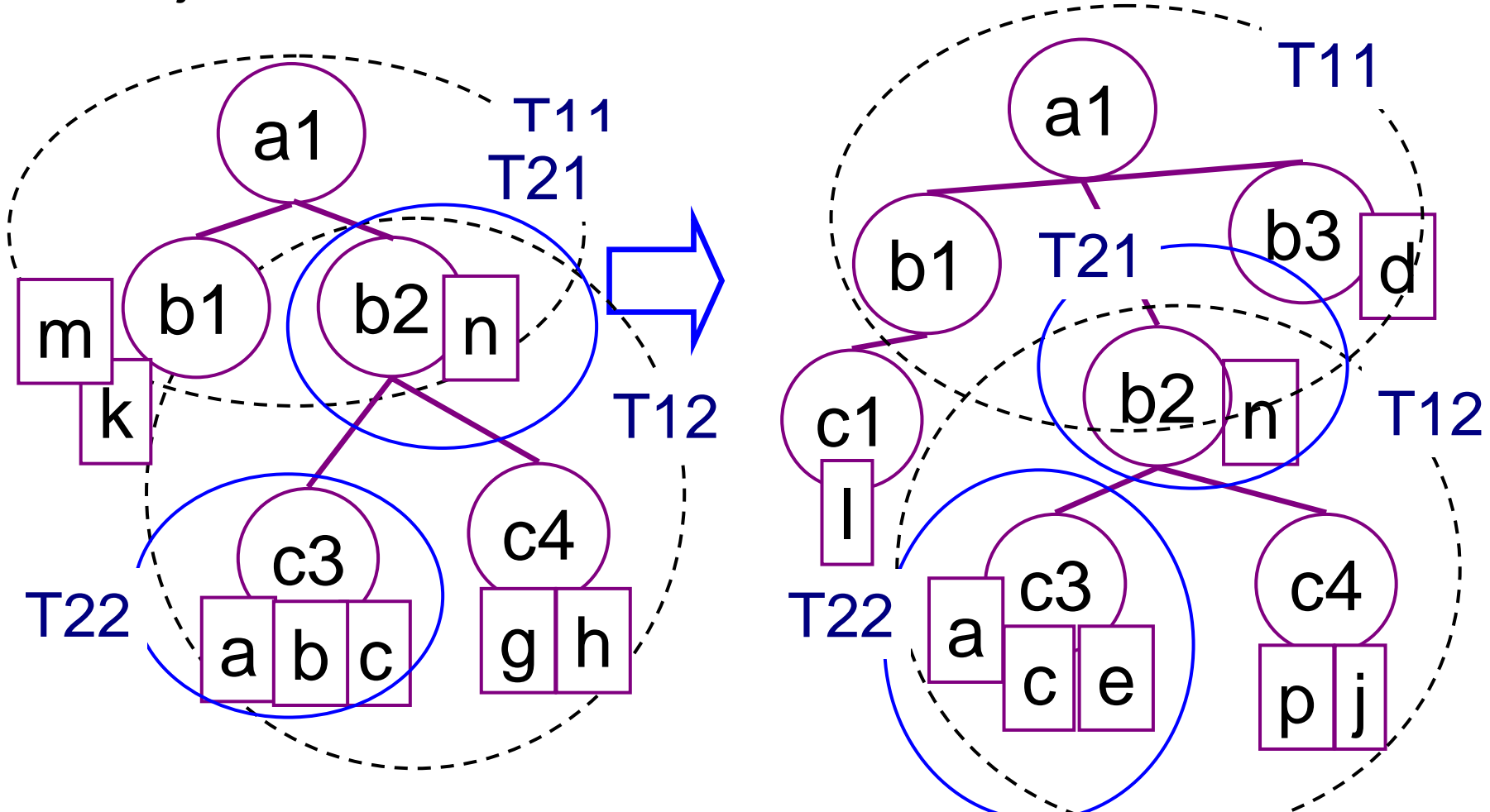


Interest ontology of user A

Interest ontology of user B

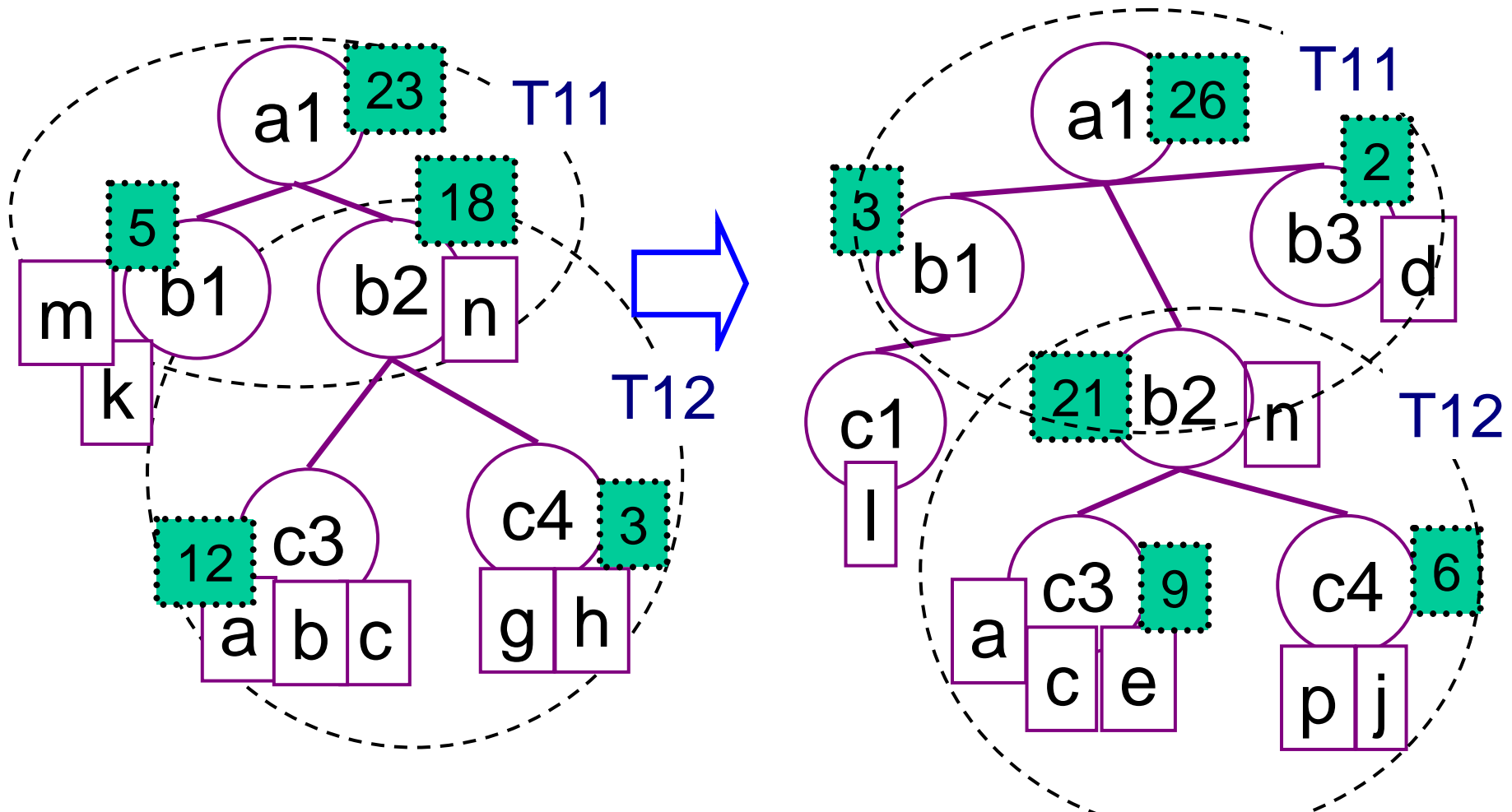
IDEA3: We treat class-class topologies $T1$ and class-instance topology $T2$ separately.

- because we consider that $T1$ reflects the width and depth of a user's interests and $T2$ reflects the objects in which users are interested.



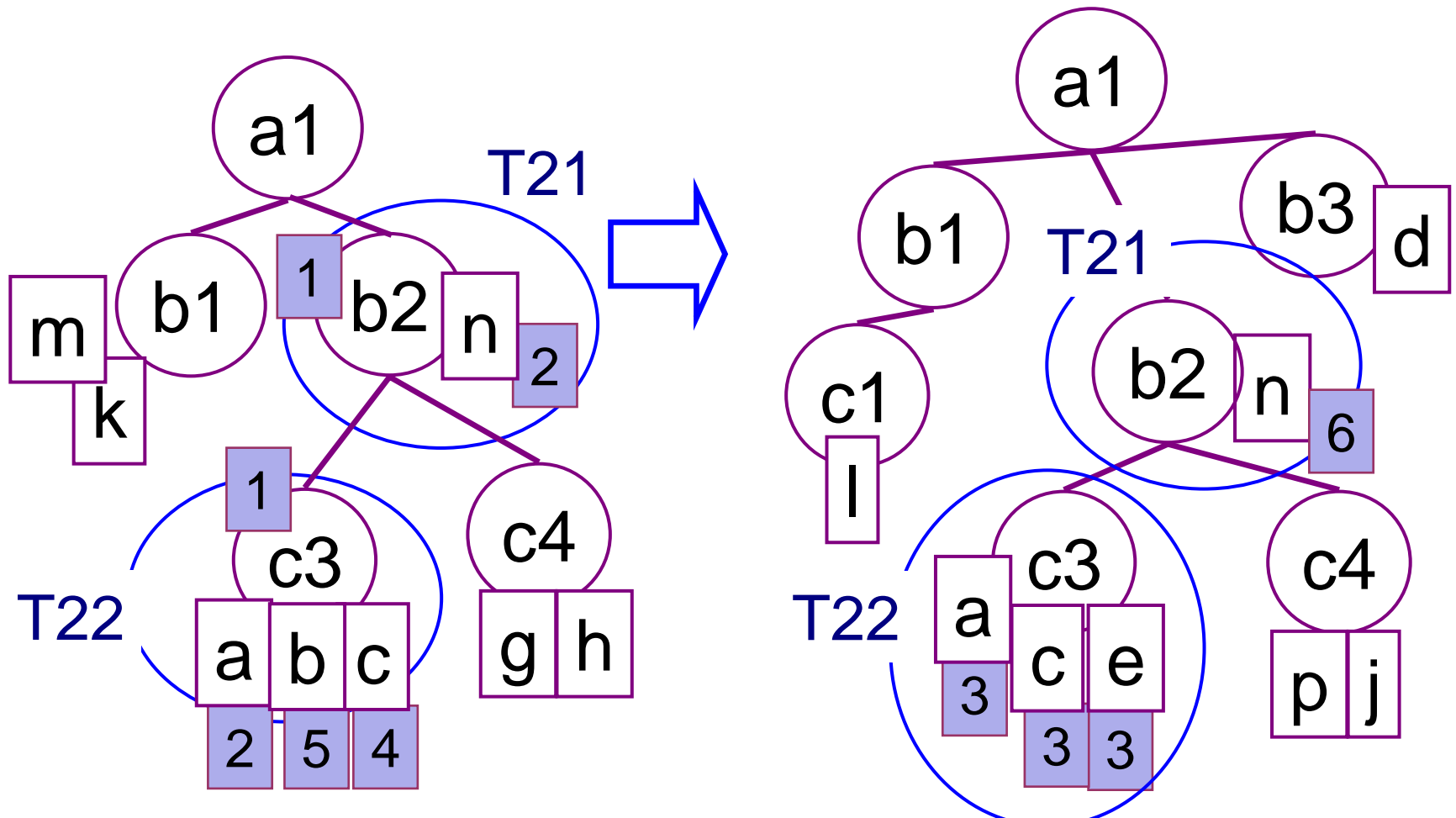
We calculate the interest agreement of $T1$ based on product sets and set union of each topology.

- $$S(T1) = S(T11) + S(T12) = (3 + 18 + 0) / 3 + (9 + 3) / 2$$



We calculate the interest agreement of $T2$ based on product sets and set union of each topology.

- $$S(T2) = S(T21) + S(T22) = (2/1) + ((2+0+3+0)/4)$$



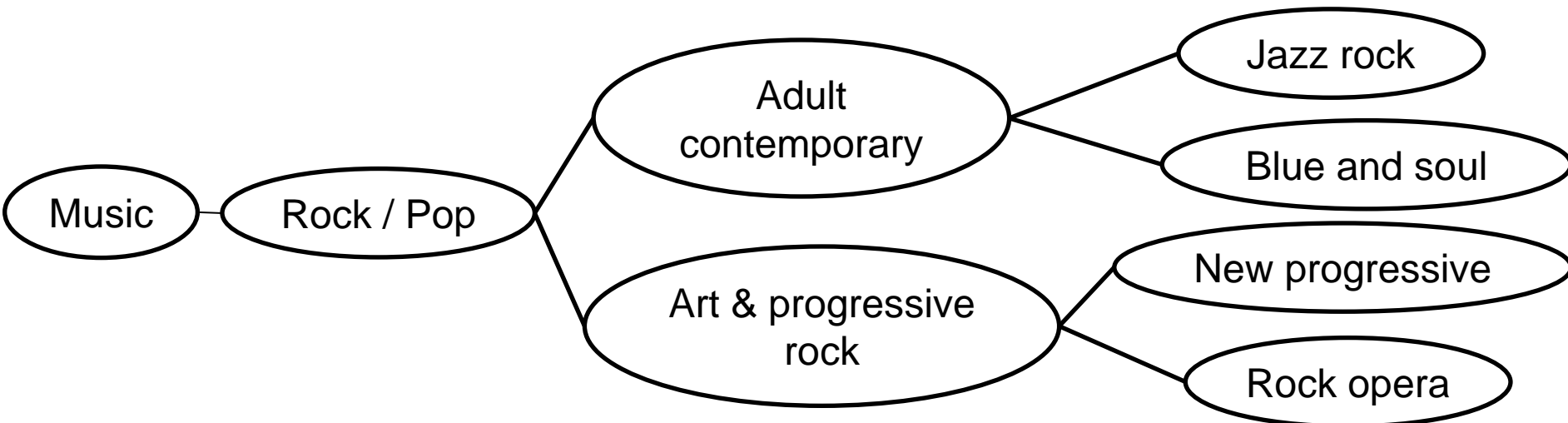
Finally, we give similarity between ontologies as $S(T1) + F(S(T2))$

- $F(X)$: the relative degree of importance of a topology

4. Experiment and Results

Datasets of offline experiments

- Whole entries of blog portal Doblog, one of large-scale blog portals in Japan. (<http://www.doblog.com>)
 - 1.6 million entries of 55,000 users
- Music service domain ontology (114 classes, 4300 instances of ListenJapan <http://listen.jp/>)
 - We used detailed domain ontology, which has 4 class hierarchies, suitable for creating niche communities.
 - Instances have several name attributes (eq R.E.M., REM).



Evaluations

We evaluated our results focused on the following 5 points.

- The accuracy generated interest ontology
- The accuracy of our detecting results.
- From the viewpoint of the degree of innovation.
- Suitable granularity of the interest-sharing group *Gu*.
- Comparing our technique with previous collaborative filtering.

Users' reaction against innovative topics and the increase of activity of communities were evaluated in online evaluation.

(Not included in this slide.)

Evaluating the accuracy of generated interest ontology

We randomly check *1/10* classified results.

Precision	Recall
94.9%	80.3%

Generating interest ontology with high precision.

We can apply interest ontology to recommend innovative topics.

Evaluating the accuracy of our detecting results

1. We set instances in manually created recommendation list in a music portal ListenJapan as correct answers.
2. We evaluated recall by changing the number of users, X in interest-sharing group Gu .
3. We evaluated our innovative instances by checking the recall in the change of X .

The change of recall of our technique

	$X=30$	$X=60$	$X=90$
Recall	64.8%	76.7%	80.1%

Improved significantly

Improved slightly

Evaluating from the viewpoint of the degree of innovation.

We compared manually created recommendations and our results of detecting innovation from the viewpoint of the degree of innovation.

1. manually created recommendations of ListenJapan

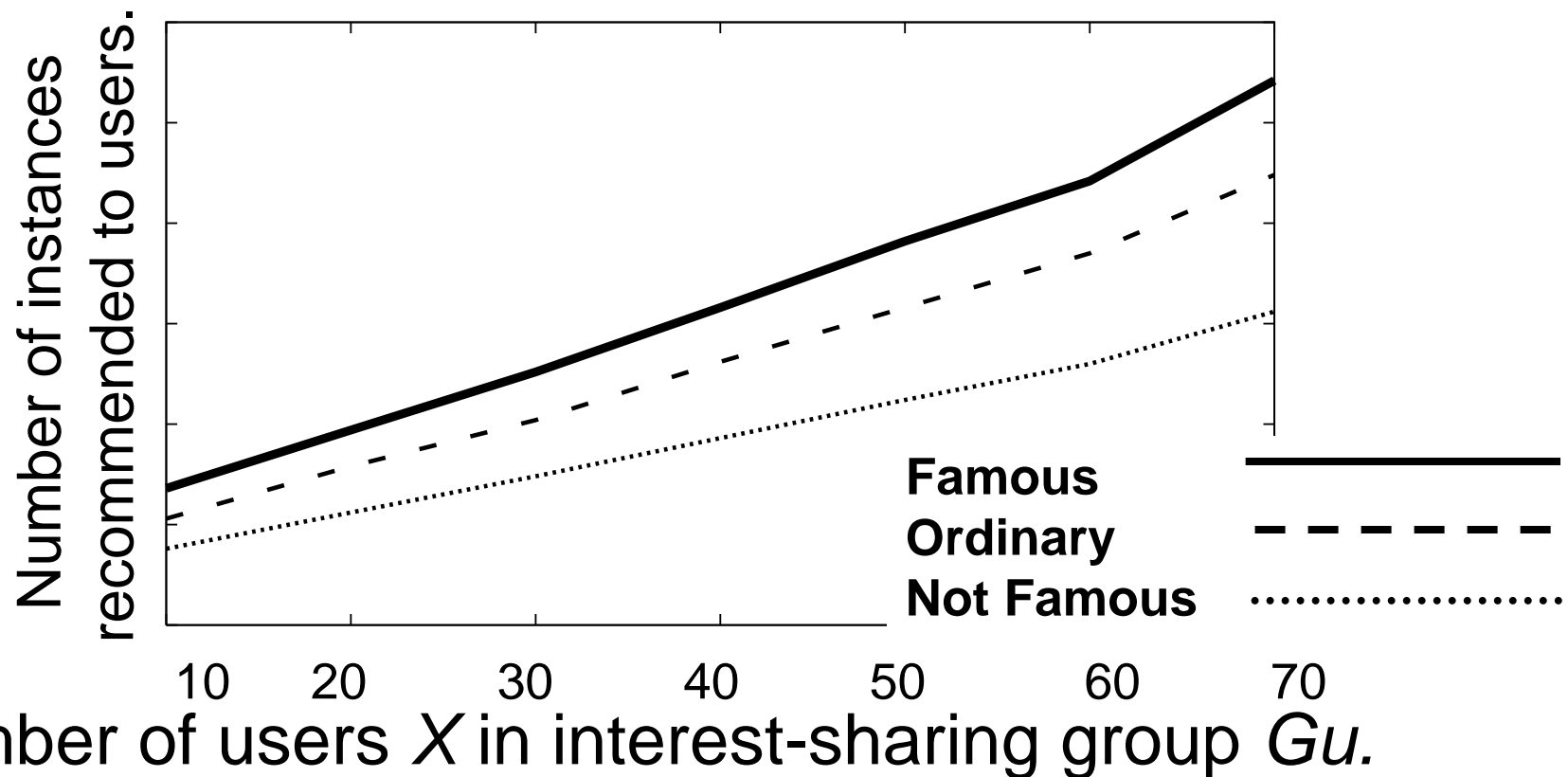
Degree of innovation	0	1	2	3
Proportion	57.6%	15.2%	23.2%	4.0%

2. Our results of detecting innovation

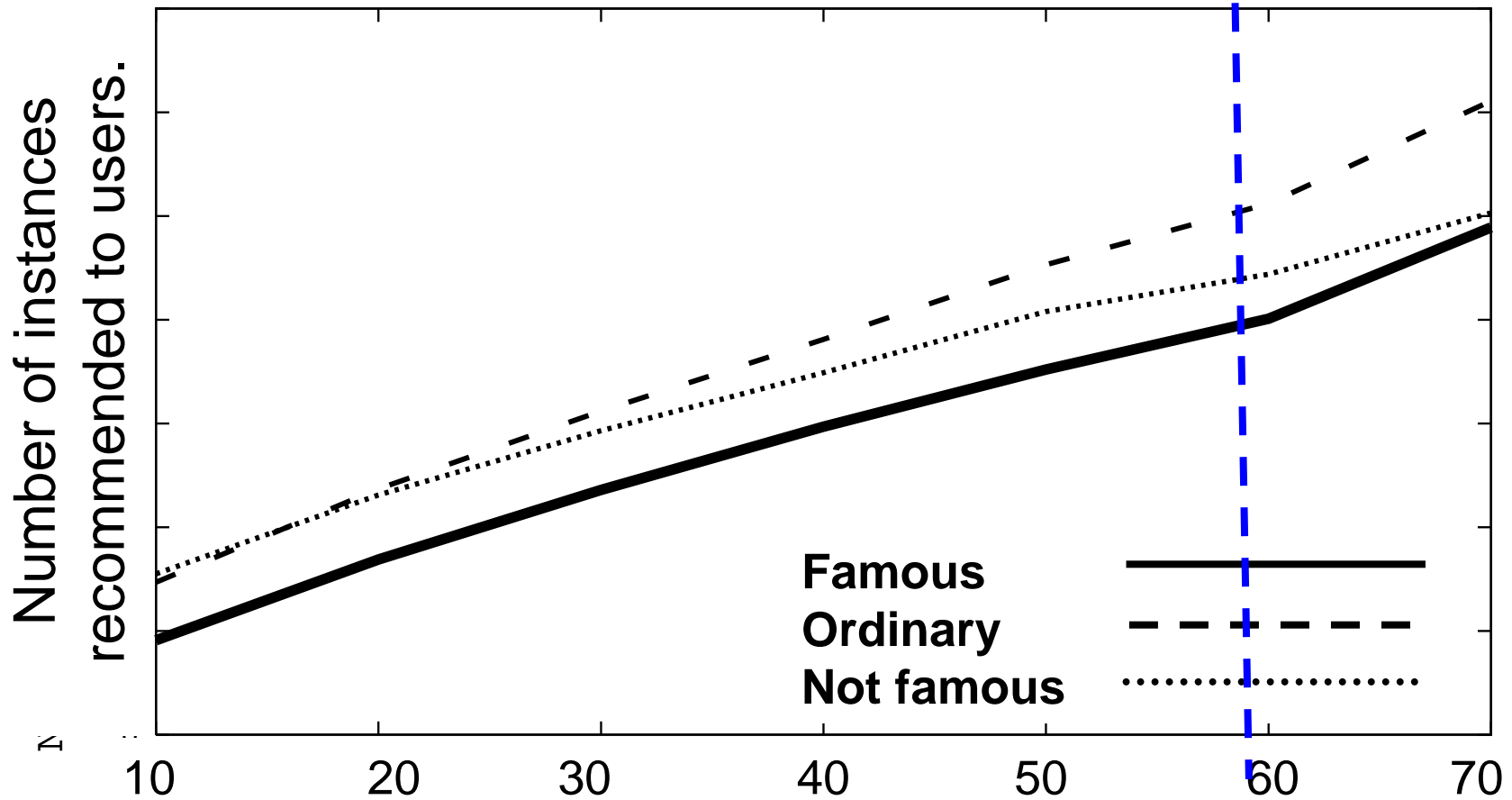
Degree of innovation	0	1	2	3
Proportion	23.4%	23.1%	44.3%	9.2%

Analyzing suitable granularity of interest-sharing group $Gu - 1$

1. We divided instances into 3 groups : a famous, an ordinary, and a not famous group.
2. We calculated the number of instances recommended to users by changing the number of users X in interest-sharing group Gu from 10 to 70.

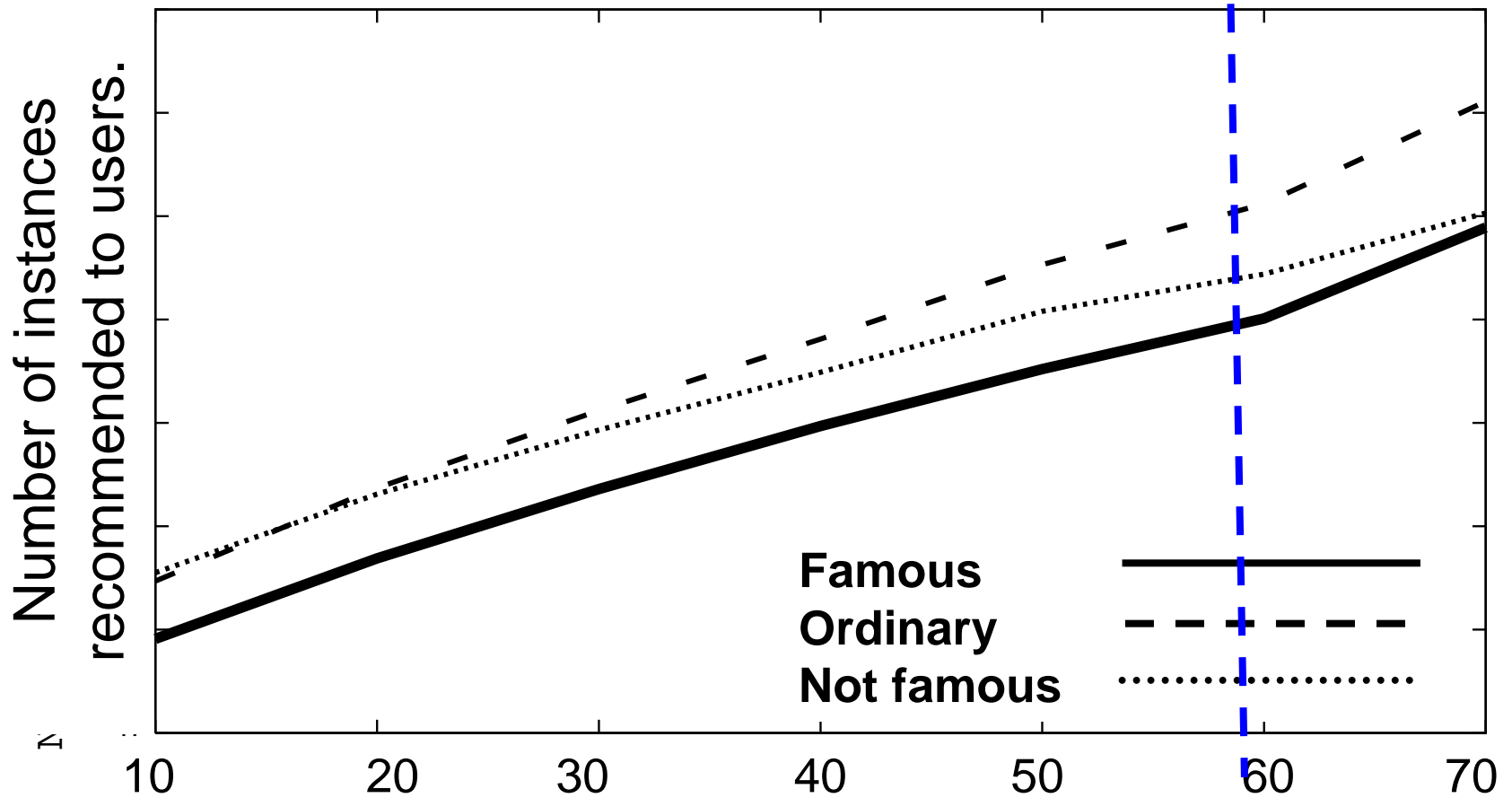


Analyzing suitable granularity of interest-sharing group *Gu*-2



(b) number of users with high interest weight in their ontology.

Analyzing suitable granularity of interest-sharing group *Gu*-2



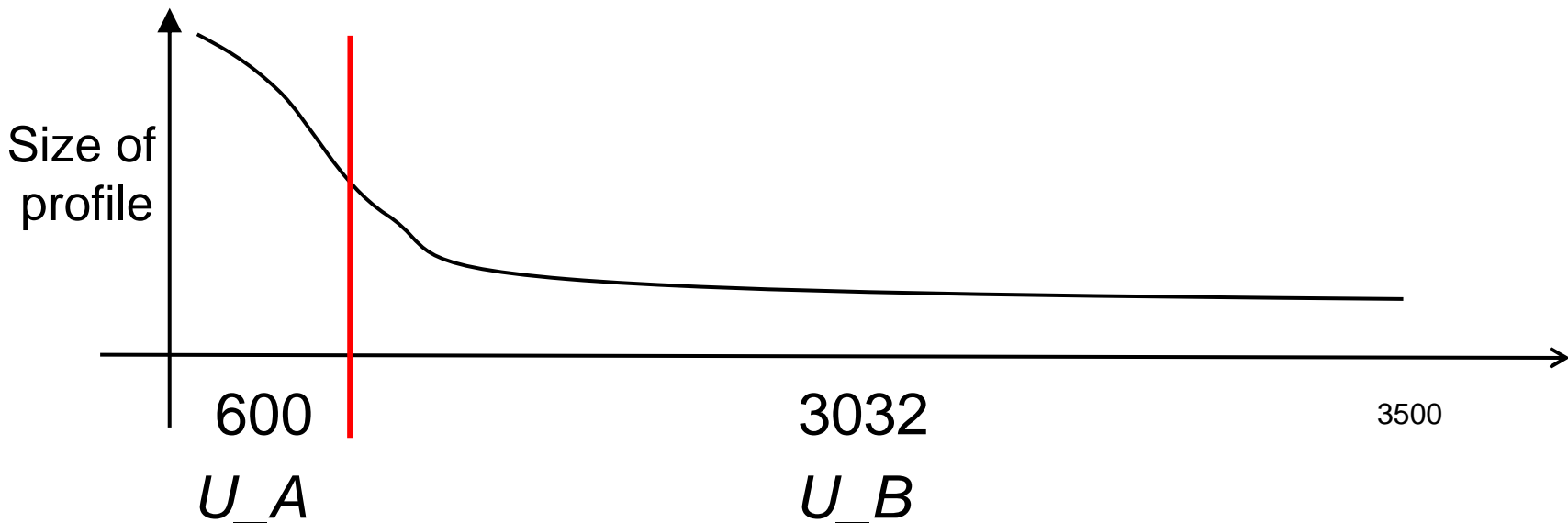
Instances with a low possibility of being interesting come to be recommended more often because the gap between a user's ontology and ontologies of group *Gu* is larger.

Comparing our technique with previous CF

- We compared our technique with previous collaborative filtering (CF) from the viewpoint of accuracy of predicting users' interests
- we divided the dataset *DATA* into the following two datasets: *test* and *predict*. *Test* is used for calculating the predicted values of *predict*.
- We prepared five types of *Test* by changing the proportion *Test/ Predict* to 16.6, 33.3, 50.0, 66.6 and 83.3%.

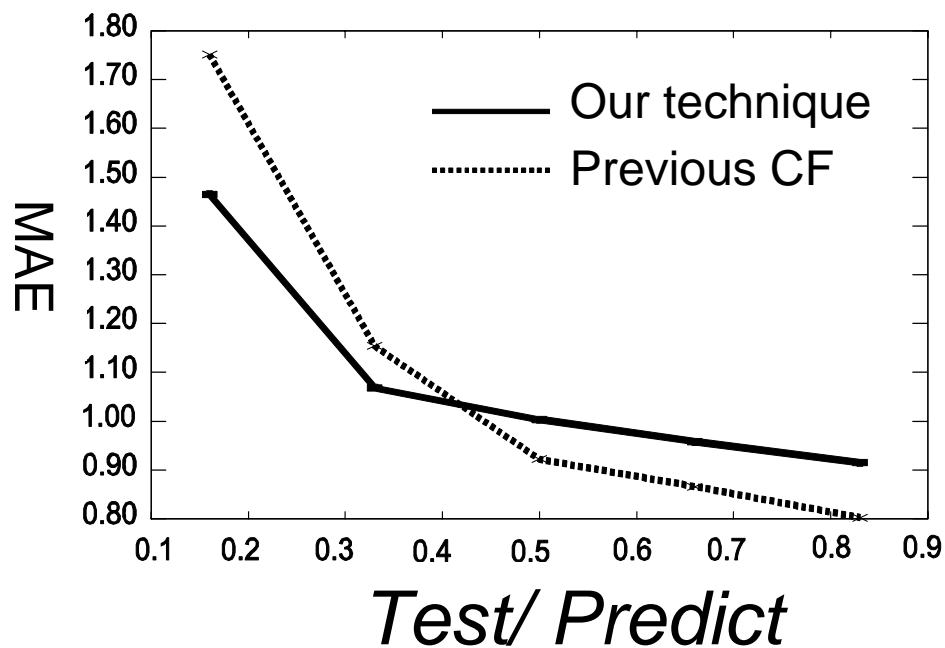
Comparing our technique with previous CF

- We divided 3,632 users into two user group: user group U_A which is composed by users who are interested in a lot of artists, and user group U_B which is composed by users who do not belong in user group U_A .

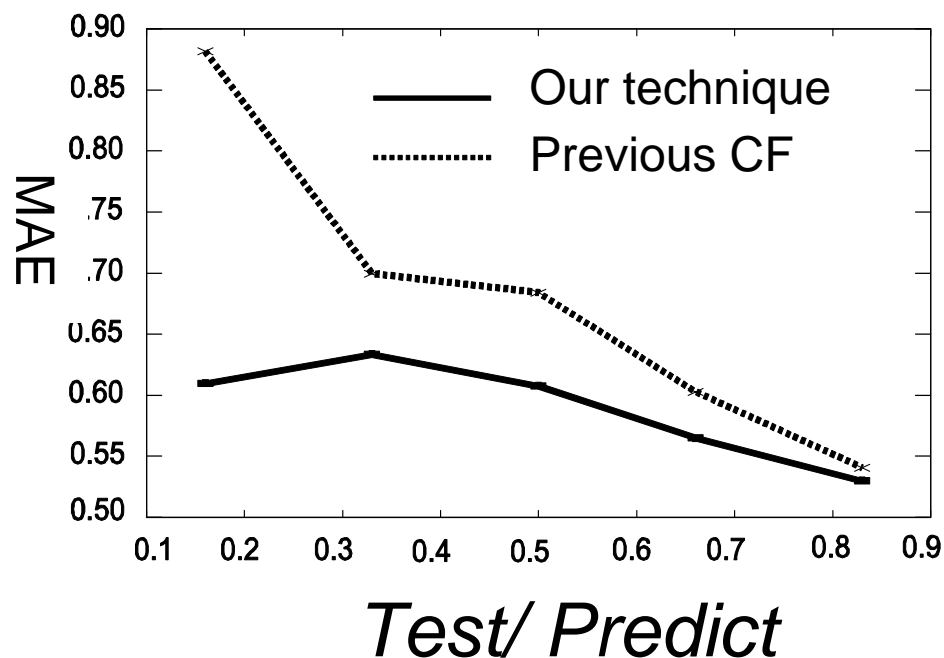


The results about MAE (mean absolute error)

- We analyze MAE about values of prediction by algorithms to about values of correct answers.



(a) We focused on all users.



(b) We focused on U_B (Light and middle users)

Discussion about comparison

- Our technique predicted user interests more accuracy than previous CF for the majority of users who hadn't stored many instances in their profiles by imposing taxonomy of topics about instances in a profile.
- We think our technique is effective for typical services most of whose users don't store a lot of instances in their profile.
- Our purpose is expanding the width of user interests by letting users browse *innovative topics, not predicting user interests*. We evaluate the effectiveness of detecting innovative topics based on actual user reaction in online evaluation.

5. Summary and future plan

Summary

- Our contributions are mainly 3 points below.
 - We extracted user-interests with high accuracy from blog-entries based on ontological mining technique.
 - We measured the similarity between user interests based on personal weight on user-interest ontology. Thus, We could detect innovative topics for a user by setting the suitable number of users G_u whose interest ontologies are similar to that of the user and by analyzing interest ontologies of G_u .
 - We evaluated our technique based on an actual large-scale blog portal. As a result, we confirmed our innovative topics were effective for each user and could increase the activity of blog communities.

References

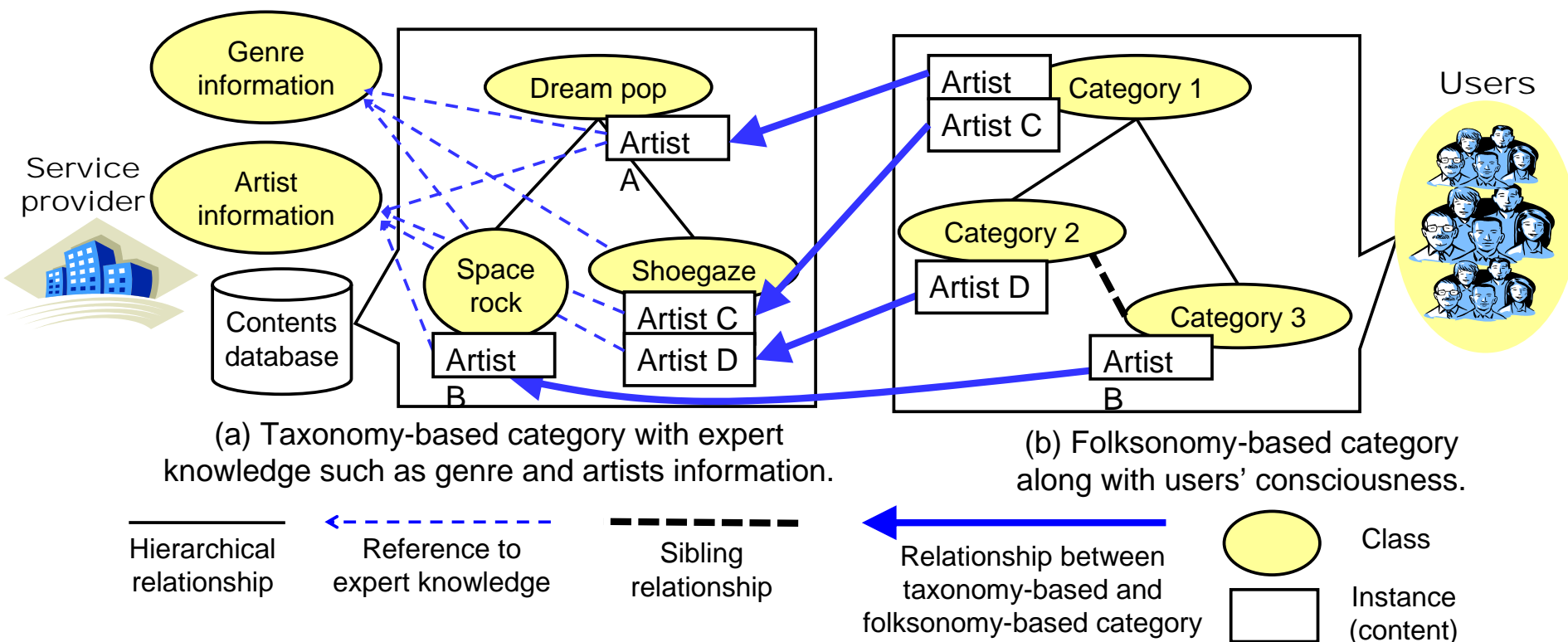
- [1] Maedche, A. and Staab, S.: Measuring Similarity between Ontologies, *Proc. Of the European Conference on Knowledge Acquisition and Management -EKAW-2002. Madrid, Spain, LNCS/LNAI 2473, Springer , pp. 251–263 (2002).*
- [2] Nakatsuji, M., Miyoshi, Y. and Otsuka, Y.: Innovation Detection Based on User-Interest Ontology of Blog Community., *International Semantic Web Conference (ISWC2006), pp. 515–528 (2006).*
- [3] Sarwar, B. M., Karypis, G., Konstan, J. A. and Reidl, J.: Item-based collaborative filtering recommendation algorithms, *World Wide Web, pp. 285–295 (2001).*
- [4] Zhang, Y., Callan, J. and Minka, T.: Novelty and redundancy detection in adaptive filtering, *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval , pp. 81–88 (2002).*

Future plan

- Future work is required to study how users control their user-interest ontologies by providing feedback of their collective knowledge to update class hierarchies of service domain ontologies constructed by expert designers of ontologies.
- We try to coordinate services based on interests, moods, locations, presences and vital information for establishing real ubiquitous network society.

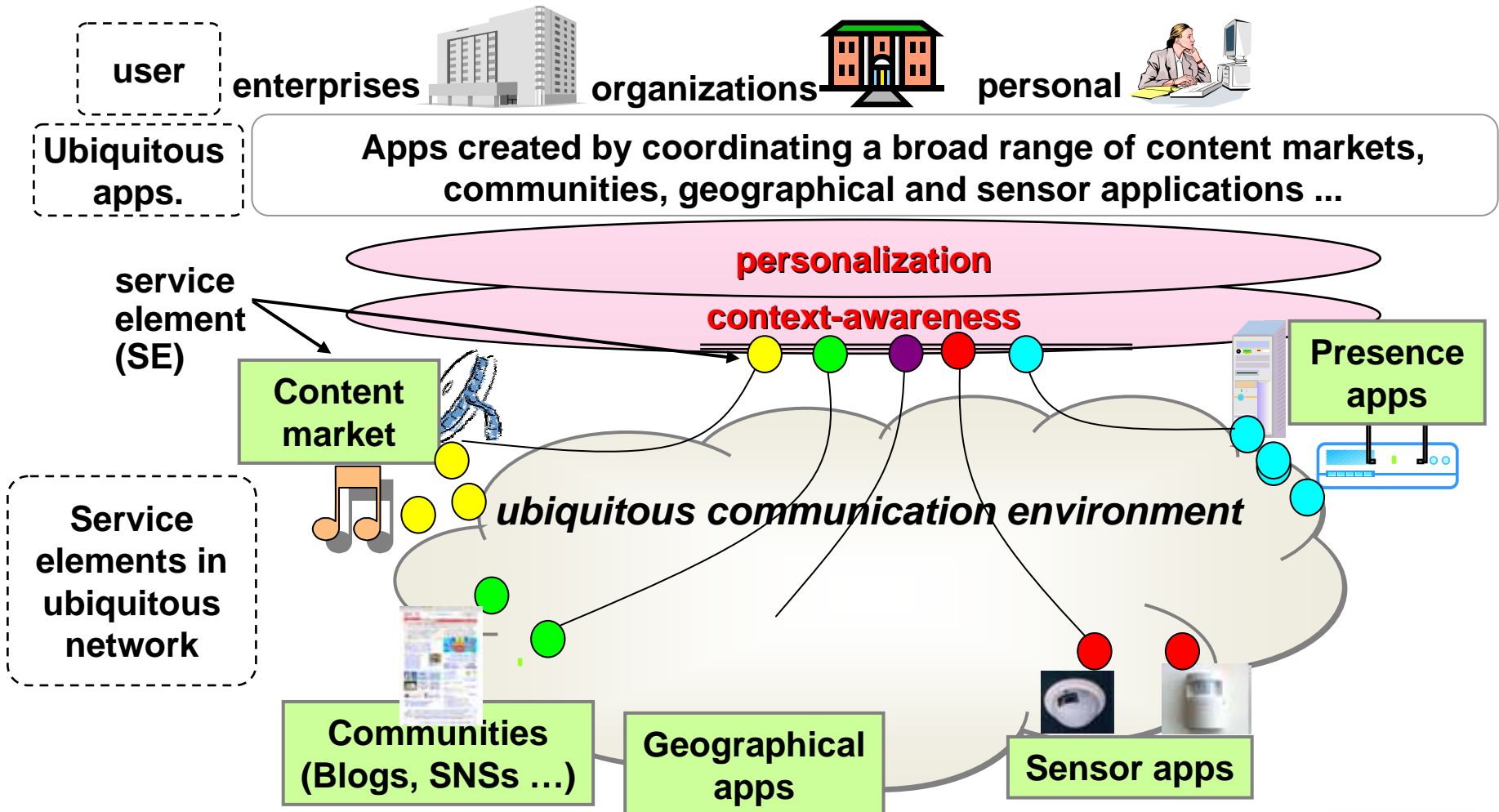
Image of relationships between taxonomy-based category and hierarchical category of community of interests.

- Users can browse content easily using users' interests and also refer to the expert knowledge about the content.
- Expert designers can get feedback about users' knowledge about the taxonomy of content by analyzing those relationships.



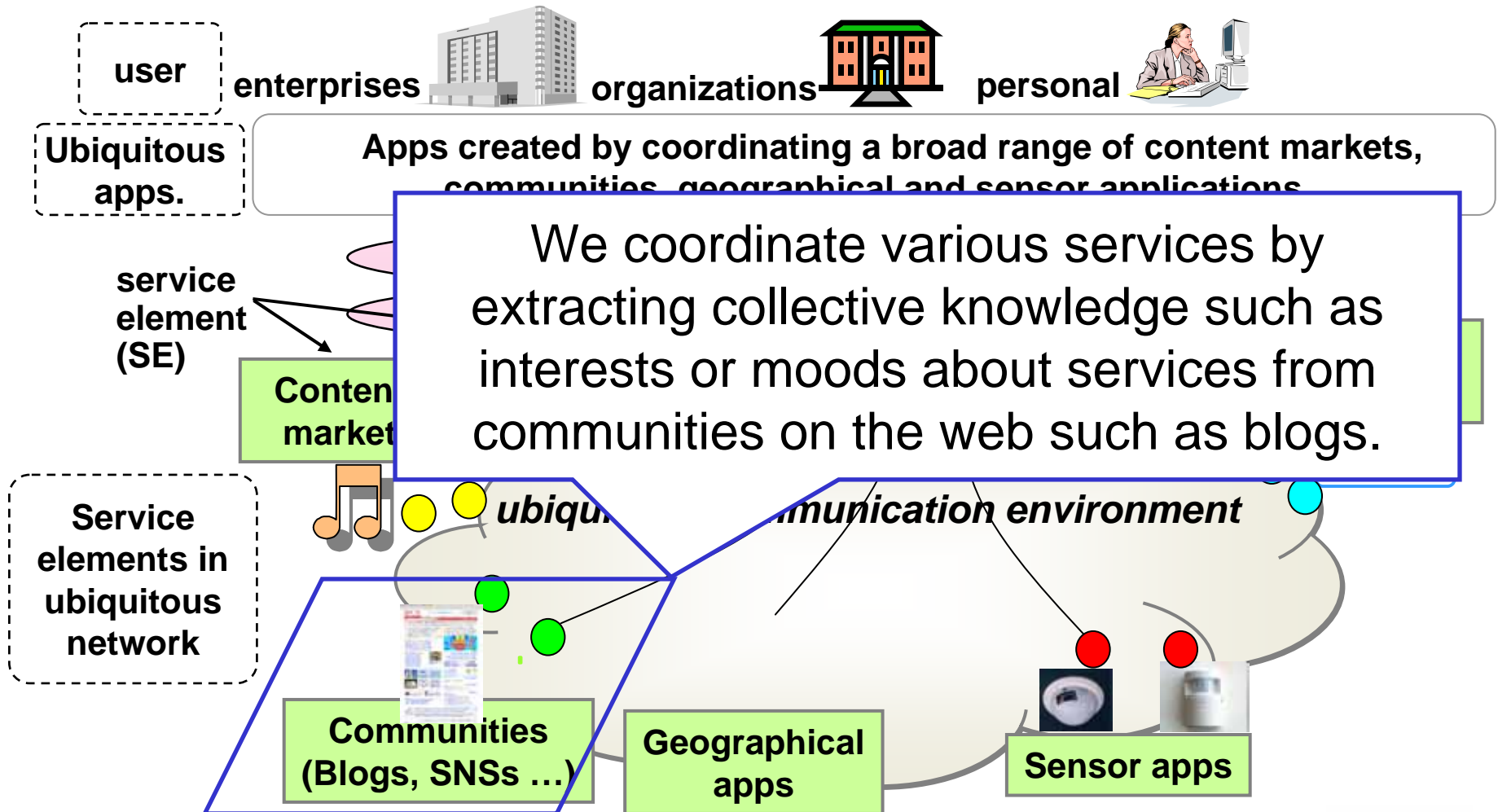
Context-awareness for coordinating services

Personalization based on context-awareness is a key technology to coordinate services in ubiquitous communication environment.



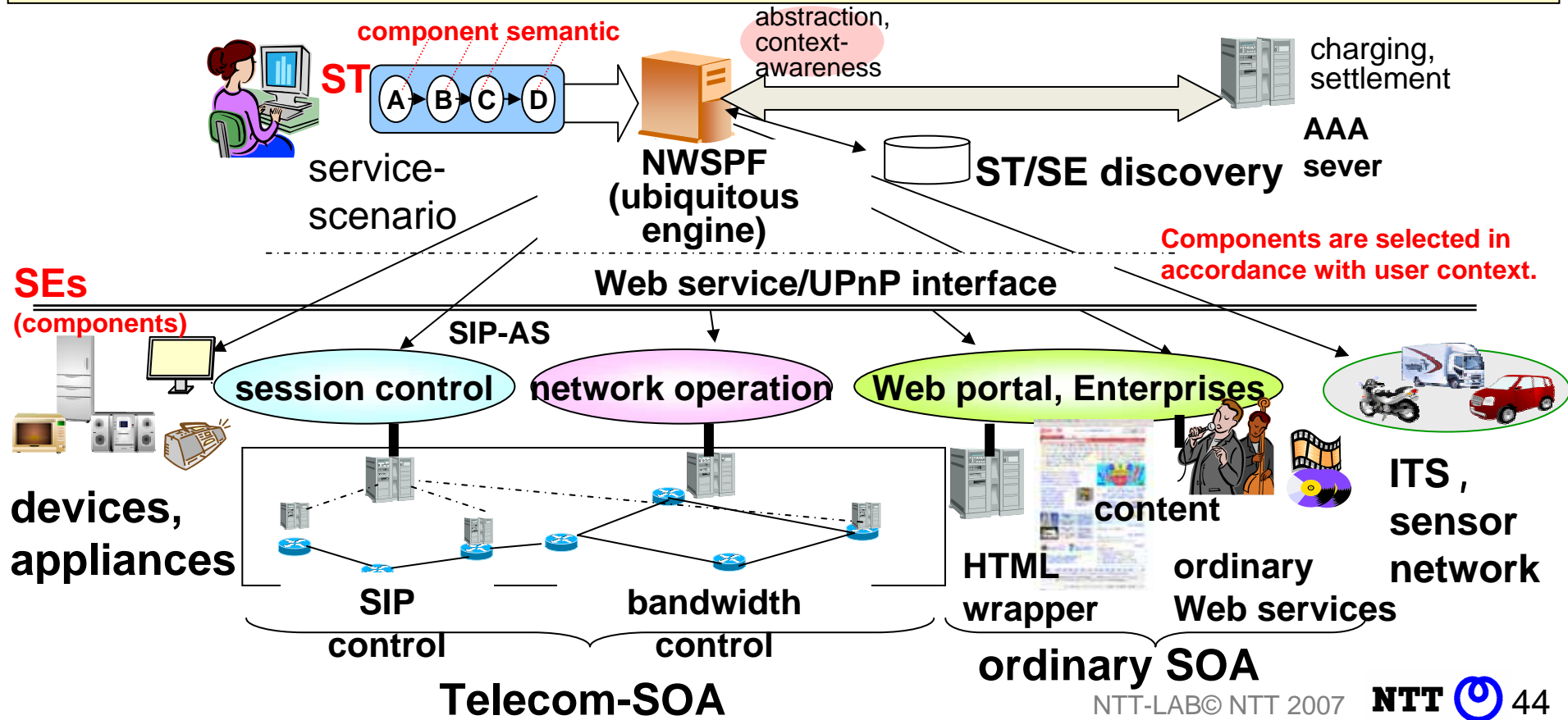
Context-awareness for coordinating services

Personalization based on context-awareness is a key technology to coordinate services in ubiquitous communication environment.



Towards Ubiquitous SOA

- We are proposing a new architecture by enhancing the ordinary SOA and the SDP to provide context-aware ubiquitous services. In this new architecture, named as the **Ubiquitous SOA**, elements such as appliances, sensors, telecom-SOA capabilities as well as ordinary SOA capabilities, are integrated, and composed services can act in context-aware ways.
- Application service providers, and even ordinary people, can easily develop and provide application services by integrating SEs in the network.





Old and innovative city, Kyoto.