Discourse Analysis of Public Debates Using Corpus Linguistic Methodologies

Hayeong JEONG[1]
Department of Urban Management, Graduate School of Engineering, Kyoto University, Japan[1]
Email: hayeong@psa2.mbox.media.kyoto-u.ac.jp[1]

Shun SHIRAMATSU[2], Kiyoshi KOBAYASHI[3], and Tsuyoshi HATORI[4]
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Japan[2]
Department of Urban Management, Graduate School of Engineering, Kyoto University, Japan[3]
Department of Civil Engineering, Tokyo Institute of Technology, Japan[4]
Email: siramatu@kuis.kyoto-u.ac.jp[2], kkoba@psa.mbox.media.kyoto-u.ac.jp[3], hatori@plan.cv.titech.ac.jp[4]

***Abstract*—The aim of this study is to develop a computational method of discourse analysis based on corpus semantics. The objective is to achieve an accurate understanding of the debate content and structure through hypotheses generation. As for verifying the hypotheses, the topic extraction and semantic similarity evaluation from the public debate minute corpus is examined by using a multi-method which includes TFIDF, T-VSM, and MDS. The main issue of public debate and the inconsistency level between participants' utterance could be described by using the method. The methodology presented in this study is applied to a case example. Finally, the applicability of the proposed methodology to practical debates is discussed.**

***Index Terms*- Content and Structure of Discourse, Corpus Semantics, Computational Methodology**

I. INTRODUCTION

Nowadays, the trend towards the dispersion of information and knowledge among the members of society, such as residents, enterprises and governments leads to increasing social complexities, diversifications, intricacies and a move towards specificity. As a result, decision makers of public projects are now facing problems involving environments of high uncertainty and ambiguity. In this context, public debate of groups consisting of various stakeholders is needed to aid in the decision-making of public projects. Indeed, in many public projects, the government implements Public Involvement (PI) processes where various project of stakeholders communicate with each other. One of the significant roles of public debates is to understand the diverse perceptions possessed by the members of the society and make judgment related to the appropriateness of the projects.

Participants, however, have diverse concerns, values and expectations and as a result they have different cognitions of the public projects. Incompatibility between the cognitions of participants could be explained as follows: "*Each person has his/her own subjective definition of a social problem*". In this situation, it is possible to get improper communication between the participants. A proper public debate process helps the diverse participants of public debates to make sound communication. For a proper public debate process, it is necessary to clarify the cognition of participants and their inconsistencies. The process for clarifying the cognition of participants is crucial for a proper public debate process. However, there are not enough investigations on formulating the cognition of participants and studying the cognitive dissonance between them. Under this situation, public debate such as the PI process has been implemented.

This study aims to investigate the conflict structure of public debates and to aid in supportting the public debate process. For this purpose, working hypotheses are generated in order to understand the conflict contents and contexts which might be caused by cognitive dissonance. In order to verify the working hypotheses, a corpus based discourse analysis method, clarifying the debate content and structure of public debate on public projects, is proposed. The proposed method is based on corpus linguistics, natural language proceeding and corpus based techniques are adopted to extract the topic of public debate and to estimate the semantic similarity between the utterances of participants. Verification of the hypotheses of the conflict structure between the participants of a debate is examined by using the proposed corpus based discourse analysis method. The study is organized as follows: section 2 describes the basic idea in detail. In section 3, the working hypotheses are generated. In section 4, the outline of the method developed in this study is explained. In section 5, a case study is presented. Finally, in section 6, the application, significance and problems of this method in conjunction with the result of the study are discussed.

II. THE BASIC IDEA

A. Discourse Analysis of Public Debate

There are various attempting methodologies to understand people's concerns such as interviews,

hearings and surveys. So far, researchers usually analyze people's awareness through questionnaires. Public debates of public project, however, has made progress towards various communications between participants. They could speak and understand based on their subjective cognition. So, there is a development of debate based on a wide context. Due to the heterogeneous content and context, the investigators may face difficulties to carry out their mission. Content and Context analysis of qualitative data such as the descriptive information is required for debate analysis [8].

In this study, discourse analysis having the feature of the content analysis is adopted in understanding the debate content and the structure using public debate minutes [9] [10]. Discourse is a term used in semantics. It is generally defined as "*the structure of texts and utterances longer than one sentence.*" Discourse is characterized as the aspect of both language use and language in-use. Schiffrin proposed a definition of discourse as "*Utterances*" that sits at the intersection of structure and function [10]. He explained that formalists and functionalists define discourse differently. Formalists define discourse by considering the linguistic characteristics of sentences as clues to textual structures. Functionalists define discourse by considering an interrelationship between language and context. He compared these two different definitions of discourse from two paradigms and discourse as language above the sentence, and discourse as language use. Utterance refers to using a sentence in a specific context. Context here means "*Information of text or statement that surrounds text or utterance and determines its meaning*"[1]. Context has various forms such as the expression or gestures of a speaker and the cultural or social context.

Discourse analysis is a method to analyze the structure of discourse as well as both their linguistic content and their sociolinguistic context. In the public debate of a public project, participants having diverse concerns and values communicate in diverse contexts. Discourse analysis, therefore, is an important method to clarify the public debate circumstances and understand its contexts.

B. Related Studies

Based on a discourse analysis way of thinking, Hatori and others developed a protocol analysis method based on Facet theory for verifying the conflict and incompatibility between opinions of participants [6]. The method was applied to minutes of public debates analysis and clarified the pattern of discourse and conflict structure. The Facet classification, however, is left to the researcher's discretion so that there is exists a problem of replicability. Horita and Kanno developed an information system that supports the policy discourse by visualizing the discourse structure [7]. This system is only for the logical relevance of discourse between units of the discourse, rather then for content of discourse. Horita and Kanno did not consider the utterance content of participants. In this study, discourse analysis method is also based on content analysis as mentioned above, and is applied to public debate minutes for certifying the content and debate structure. Here, the computational approach, based on corpus linguistic [2], is developed to topic extraction and the calculation of semantic similarity between utterances of participants of the public debates. The approach presented in this study, i.e. discourse analysis of public debates using natural language corpus (debate minutes), excludes the problem of replicability in conducting discourse analysis. The primary objective of this method is to investigate the utterance content and the semantic similarity.

C. Corpus based Discourse Analysis

Corpus is generally defined as a text collection, a large-scale sample of written and spoken material on the way of using language. Corpus linguistics is a linguistic method mainly using the corpus as evidence for explaining the meanings of words and phrases. Corpus linguistic is characterized by a reconstructive method for analysis of language data using a computer. Saussure, Wittgenstein, and Austin identified two principles on corpus semantics; 1) Meaning is use 2) Meaning is relational [11]. The meaning of language is obtained or transformed by accepting the social and language context. So it could be said that the two principles could make people grasp the implication of language.

Firstly, the meaning of word is for the first time realized in the use of context of an actual situation. This is consistent with the definition of discourse and utterance which are described in the previous section. Secondly, the meaning of a word is captured in relation to other words coappearing in the context. That means, the meaning of a word is not fixed only to be described in dictionary, but it is changed due to the actual social context and linguistic context. Based on the principles, corpus linguistics evaluate the ways of using language with actual data of language use. Thus, it is possible to investigate the meaning of a word in context and also its implication. Therefore, corpus linguistics is consistent with the main idea of this study for discourse analysis.

Discourse analysis studies based on corpus linguistics are classified broadly into two approaches: the examination of the frequency and the distribution of words [18]. The study of frequency is a statistical way to explore how many times words and phrases appear in language use. The study of distribution is also a statistical way to verify contexts derived from words and phrases. In this regard, however, a lot of prior studies focus on specific word and phrase use in text. Only a few studies have attempted to compare the language use differences - content and context - between speakers in identical text.

The discourse analysis proposed in this study is based on corpus linguistics described above, and characterized to investigate the difference on content and context of participants in public debate. Using the proposed method, we examine and extract topics from corpus of public minute debate, and estimate the semantic similarity among the utterance of participants with the topic collocations, i.e. the way words usually co-occur with topics. Topic collocations may infer the regulation of discourse prosody based on an individual lexical system

which is closely connected with the individual background knowledge, belief, frame, fixed idea and so on. It is expected to infer the frame and the cognition of participants. Topic extraction falls under the studies of frequency and the semantic similarity is under the studies of distribution. It may be useful to evaluate the conflict structure in a public debate, as well as, the inconsistency between the cognitions of participants and the closing condition of a public debate process.

III. Generation of Working Hypotheses

The following are the generated working hypotheses in order to interpret conflict contents and structures of public debate, with regards to the participants' utterance.

Hypothesis 1. The more serious a conflict between any two participants is, the less semantic similarity of their utterances is.

This study assumes conflict situations as follows. Two individuals have their opinions on a discussed subject in their different contexts. For example, individual p_i reminds "A" of S. Conversely, individual p_j reminds "B" of S. In addition, two individuals have their opinions on a discussed subject based on their different values. For example, individual p_i say S is "good". Conversely, individual p_j say S is "bad".

Hypothesis 2. There exist clustering of opinions between participants, that tends toward the average of semantic similarity of the participants' utterances.

We define "cluster" where at least two individuals have high semantic similarity. That might mean that they express very similar opinions and values on a discussed topic. Here, we call two situations "high clustering" One situation is where there is a single cluster and the number of individuals is high. The other is where the number of clusters is high.

Hypothesis 3. The lower the aggregation and coherence of opinion between all participants is, the more dispersed their semantic similarity of utterances is.

We define "aggregation and coherence of opinion" when all participants have high semantic similarity with each other. The high semantic similarity and the less dispersion among all participants is better while the less semantic similarity and the high dispersion between not only two individuals but also clusters is worse.

These hypotheses imply that the individuals' cognition and lexical systems are different as much as cognitive dissonance on discuss subject is concerned. The level of cognitive dissonance is described by conflict, opinion aggregation and opinion coherence with 'semantic similarity'.

While we verify the level of cognitive dissonance, we expect that we might understand three important things. First, what are the issues that derive conflict among participants? Second is how differences in cognition affect the ways of thinking of subjects and their utterances. Third, what is the dynamic of cognitive dissonance according to debate situation.

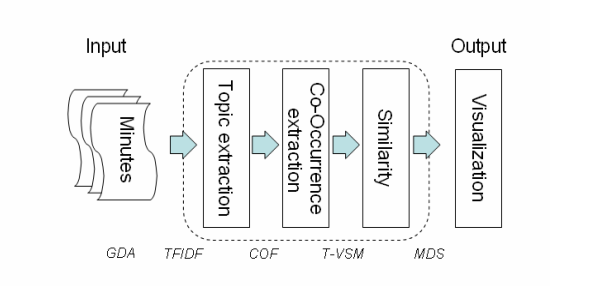

Figure 1. Methodology Outline

These hypotheses need to be verified coupling the methodologies of information science and statistics for analyzing discourse. A new methodology in order to validate the hypothesis is proposed and described in detail in the next section.

IV. Methodology

A. Outline of Methodology

The methodology is developed in order to analyze public discourse by using the discourse minutes instead of the testimony of investigators, which is based on corpus-based techniques. The debate minutes that recoded all the utterance of the participants are very useful for investigating the content and structure of a public debate. The methodology is applied to data mining from minutes for discourse analysis.

Fig. 1 shows the methodology outline. The methodology combines five different types of techniques for data mining and analysis. The natural language proceeding technique enable computers to understand human language is applied to the modification of minute. Statistical data mining techniques are combined in order to carry out discourse analysis of public debates . The methodology is available for an accurate understanding of not only the debate content, such as participants' interests, but also the debate structure such as those who have common interests or conflict interests and the variation of semantic similarity between individuals.

In detail, the Global Document Annotation (GDA), Term Frequency Inverse of Document Frequency Implementation (TFIDF), Co-Occurrence Frequency (COF), Topic-based Vector Space Model (T-VSM), and Multidimensional Scaling (MDS) are used to analyze the minutes.

First of all, a minutes corpus of natural language is made by natural language proceeding techniques. In this study, part-of-speech (POS) tagging for Japanese is done by the ChaSen (a Japanese morphological analysis system) module [12]. Meta-language is necessary to support computational linguistics or corpus linguistics. A meta-language is a language used to describe other languages. GDA is based on Extensible Markup Language (XML) and is used to get meta-language and minute corpus. The GDA initiative allows machines to automatically recognize the underlying semantic and pragmatic structures of documents. It has applications in the information retrieval, informative summary, the

anaphoric relation, morpheme analysis and so on. It is applied to this study to get the POS information of language in minutes [13].

Next, significant keywords in the minute corpus could be extracted using the TFIDF measure proposed by Salton *et al.* (Salton and McGill 1984; Salton and Buckley, 1988)[19][20][21]. This measure is used to understand both the debate context and participation's primary concerns. The co-occurrence set (co-occured terms and their frequencies) of keywords is specified. The co-occurrence set could express individual belief and knowledge via the keywords, and it is useful to understand the individual cognition related to the keywords in a public debate. Finally, the co-occurrence set similarity is evaluated using TVSM and the similarity is visualized using MDS. Following, the cognitive dissonance structure of public participants is described.

B. Topic Extraction

Terms are weighted using the TFIDF scheme proposed by Salton et al., and it is applied to keyword extraction from the minutes corpus to understand both the debate topic and the participants' primary concerns [3]. TFIDF is based on the term frequency of word appearance and is used to decide the significance of a term w in a document a. The definition of TFIDF is given below:

$$TFIDF_{w,a} = TF_{w,a} \times IDF_w$$
$$IDF_w = \log(\frac{N}{DF_w}) + 1 \quad (1)$$

$TF_{w,a}$ = Number of occurrences of term w in document a.

DF_w = Number of documents containing term w.

N= Total number of documents.

The total weight of significance $TFIDF_{w,a}$, named as TFIDF score, is calculated as "term w frequency in document a ; $TF_{w,a}$ " times the "inverse of document frequency containing term w ; IDF_w ". The term with a high TFIDF score implies a significant term w in document a. The terms that appear frequently in a document characterize the document. High frequency terms, however, are not necessarily important. The inverse of document frequency is, then, applied to decide whether a term is significant or not compared with relevant documents. The IDF value is defined as the logarithm of the value of the total number of documents, DF. DF_w stands for the document frequency or in how many documents the term w occurs. For example, if a term w_1 appears frequently in document *d* and only in a few documents, then IDF value of term w_1 is high and term w_1 has high TFIDF score. Nevertheless, if another term w_2 appears in many documents and appears infrequently in a document, then IDF value of term w_2 is low and term w_2 has low TFIDF score. This study has a minutes corpus of 100 relevant minutes on public debates. Five high TFIDF scores were selected as debate topics and the participants' primary concerns.

C. Topic based Co-occurrence and Semantic Similarity

As mentioned in previous section, we examined how to measure the semantic similarity for understanding the participants' cognitive dissonance around topics. For measuring the semantic similarity, we use "Topic based Co-occurrence" and "Cosine Distance Measure between different Topic based Co-occurrences".

Topic based Co-occurrence is a set of co-occurred terms of a topic and its frequency in an individual's utterances. Individuals have different co-occurrence restrictions on a topic due to different knowledge and belief. By using the Topic-based Co-occurrence, we can understand the lexical system difference between individuals and thus to infer the cognitive difference of individuals.

The topic-based co-occur term set is used to expand to VSM to present the semantic similarity between two different individuals' utterances. The VSM proposed by Salton (1968) is a conventional information retrieval (IR) model that represents documents and queries as vectors in a multidimensional space [16].

In this study, each term *w* is respectively weighted with the co-occurrence frequency $TF_{w,t,p}$, i.e., the frequency of co-occurrence between *w* and topic *t*. It is counted only in an individual *p*'s utterances. Where $W = \{w_1,...,w_n\}$ denotes a set of total *n* words in a document. The weighted co-occured term set is represented by the vector $\overrightarrow{TF}_{W,t,p}$ in an n-dimensional space.

$$\overrightarrow{TF}_{W,t,p} = [TF_{w_1,t,p},...,TF_{w_n,t,p}] \in R^n \quad (2)$$

Fig. 2 describes the weight of vector examples belonging to different individuals. Different individuals' topic based co-occurrence vectors $\overrightarrow{TF}_{W,t,p_i}$ and $\overrightarrow{TF}_{W,t,p_j}$ could be represented in the same n-dimensional space [4][5].

Perl regular expression syntax is used to calculate the co-occurrence vectors. We obtain the matrix S composed of $\overrightarrow{TF}_{W,t,p_1},...,\overrightarrow{TF}_{W,t,p_m}$ with all individuals $\forall p \in \{p_1,...,p_m\}$ as follows:

$$S = \begin{bmatrix} \overrightarrow{TF}_{W,t,p_1} \\ \vdots \\ \overrightarrow{TF}_{W,t,p_m} \end{bmatrix} = \begin{bmatrix} TF_{w_1,t,p_1} & \cdots & TF_{w_n,t,p_1} \\ \vdots & \ddots & \vdots \\ TF_{w_1,t,p_m} & \cdots & TF_{w_n,t,p_m} \end{bmatrix} \quad (3)$$

The semantic similarity between the cognitions of two individuals p_i and p_j on topic t is defined as the cosine angle distance between $\overrightarrow{TF}_{W,t,p_i}$ and $\overrightarrow{TF}_{W,t,p_j}$ as

given in the following equation (4).

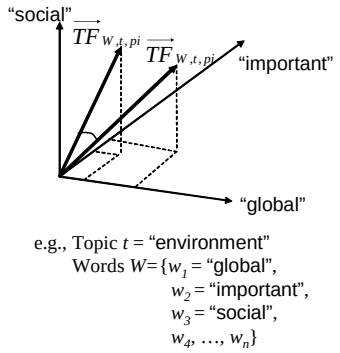

Figure 2. Co-Occrrence Vector Difference

$$sim_t(p_i, p_j) = \cos(\overrightarrow{TF}_{W,t,p_i}, \overrightarrow{TF}_{W,t,p_j}) = \frac{\overrightarrow{TF}_{W,t,p_i} \cdot \overrightarrow{TF}_{W,t,p_j}}{|\overrightarrow{TF}_{W,t,p_i}||\overrightarrow{TF}_{W,t,p_j}|} = \frac{\sum_{k=1}^{n} TF_{w_k,t,p_i} \cdot TF_{w_k,t,p_j}}{\sqrt{\sum_{k=1}^{n} TF^2_{w_k,t,p_i}} \sqrt{\sum_{k=1}^{n} TF^2_{w_k,t,p_j}}} \quad (4)$$

Notice $\cos^{-1}(sim_t(p_i, p_j)) \in [0, \frac{\pi}{4}]$, that is, the angle between $\overrightarrow{TF}_{W,t,p_i}$ and $\overrightarrow{TF}_{W,t,p_j}$ is equal or less than 90 degree.

In fact cosine distance is very sensitive to the variance of compared vectors. Thus, it is very useful in calculating the similarity between two factors in the feature quantity space [17][22][23]. Consequently, high cosine distance between two vectors means high similarity of cognition between two individuals in this study.

By using the cosine angle distance, the similarity of cognition is realized, thus the level of cognitive dissonance between different individuals can be inferred. The cosine angle distance is applied to visualize the semantic similarity between individuals in a two-dimensional space.

Consequently, cognitive dissonance on a topic t between individuals can be identified. Thus, the conflict debate content and its structure becomes clear.

D. Visualization

With the semantic similarities evaluated in the previous section, each individual could be arranged in a two-dimensional space to represent the observed cosine angle distances using Multidimensional Scaling (MDS) [14] [15].

For example, among individuals $p \in \{p_1, ..., p_m\}$, the semantic similarity between p_i and p_j, $sim_t(p_i, p_j)$ has a weak relation evident from the distance between p_i and p_j, $dis_{pi,pj}$ on a two-dimensional space.

If $sim_{pi,pj} > sim_{pi,pk}$ then $dis_{pi,pj} < dis_{pi,pk}$.

Using the MDS, the semantic dissimilarities between two individuals p_i and p_j, $dsim_{pi,pj}$ are evaluated from the semantic similarities $sim_t(p_i, p_j)$ by inversing the cosine function. [24][25][26]

$$dsim_{pi,pj} = \cos^{-1}(sim_t(p_i, p_j)) \quad (5)$$

The semantic dissimilarity between two individuals p_i and p_j, $dsim_{pi,pj}$ is congruent to the distance between two individuals p_i and p_j, $dis_{pi,pj}$

$$dsim_{pi,pj} \cong dis_{pi,pj} \quad (6)$$

All distances $dis_{pi,pj}$ are arranged in a correlation matrix D .

$$D = \begin{bmatrix} dis_{p_1,p_1} & \cdots & dis_{p_1,p_m} \\ \vdots & \ddots & \vdots \\ dis_{p_m,p_1} & \cdots & dis_{p_m,p_m} \end{bmatrix} \quad (7)$$

MDS attempts to reproduce the distance $dis_{pi,pj}$ on an n-dimensional space to the distance $dis^*_{pi,pj}$ on a two-dimensional space. The sum of squares of distance between $dis_{pi,pj}$ and $dis^*_{pi,pj}$ is desired to be minimized. By minimizing the gap between $dis_{pi,pj}$ and $dis^*_{pi,pj}$, an accurate coordinate value is ensured, hereafter referred to as "stress."

$$stress = \sqrt{\frac{\sum_{i=1}^{m-1}\sum_{j=2,j>i}^{m}(dis_{pi,pj} - dis^*_{pi,pj})^2}{\sum_{i=1}^{m-1}\sum_{j=2,j>i}^{m-1}(dis_{pi,pj} - \overline{dis})^2}} \quad (8)$$

The term, $\overline{dis}$ denotes the average of the distance between p_i and p_j

$$\overline{dis} = \sqrt{\frac{\sum_{i=1}^{m-1}\sum_{j=2,j>i}^{m} dis_{pi,pj}}{{}_mC_2}} \quad (9)$$

By changing the dimension value to a lower value, optimal arrangement for the coordinate value in a two-dimensional space can be defined. Consequently, the semantic similarity of all individuals could be visualized as the distances in a two-dimensional space.

V. ANALYSIS

A. The profile of case studies

The case study investigates the public debate discourse analysis of the minutes of Yodo-river committee in Japan. The Yodo-river committee can be considered as an

example of public debate on public project of the Yodo-river. The Yodo river committee established to obtain advice for the planning and policy handling of the river improvement project related to the building of a dam, and also to reflect the opinions of the representatives of the citizens and public organizations. The Yodo-river committee meeting consists of a general meeting, four regional meetings, five theme meetings, five working group meetings and three meetings of sub-working group. About 400 meetings were held since 2001 until today. From these meetings, three cases, two regional meetings and one sub-working group meeting, are included in this study. Only, regional meetings (case1 and case2) were related to different dams.

B. Participants Classification

Participants are classified into several categories based on their properties. In order to clarify the conflict structure between participants, we divide them based on two aspects. One is their role in society and the other is their opinion. Regarding the role, participants are divided into three groups, experts of committee members, citizens, and administrators (river managers). Regarding the opinion, participants are divided into two groups, Pros or Cons. The classification of the main content of Pros and Cons opinions are shown in Table Ⅰ.

In regional meetings, case 1 and case2, experts of committee members and citizens discussed concerns on dam building project. Case 1 consisted of 8 experts and 4 citizens. In the case, 1 citizen and 5 experts have contrary opinions on the promotion of project, and 3 citizens and 1 expert have approval opinions. The rest of them stand neutral. Case 3 consisted of 6 experts and 4 citizens. Among them, 2 citizens and 4 experts have contrary opinions on the promotion of project whilst. 2 citizens have approval opinions. The rest of them stand neutral.

In the sub-working group meeting, case 3, experts and administrators, who are river managers, discuss the promotion of building dam project focus on its capacity. Case 3 consisted of 13 experts and 5 administrators. Among them, 4 experts have contrary opinions on the promotion of project, and 3 administrators and 1 expert have approval opinions. The rest of them stand neutral.

C. Comparison of Topics

All the terms of the three cases were weighted with TFIDF score. Table Ⅱ shows the high-ranked ten terms of the three cases.

Initially, Case 1 and Case 2 are compared. On the contrary to the high-ranked terms such as “dam”, “flood control”, “basin”, “opinion”, low-ranked terms such as “water system”, “water”, “environment”, and “resident” in case 1 are different with low-ranked term in case 2. This means that the primary subject of debate content is

TABLE Ⅰ.

PARTICIPANT CLASSIFICATION

	Content of Pros	Content of Cons
Expert Citizen Administrator	Promoting project Necessity & Validity Merits of project Development	Finding alternation Unnecessity & Invalidity Demerits of project Expenses

TABLE Ⅱ.

THE RESULT OF TOPIC EXTRACTION BY TFIDF

Case 1		Case 2		Case 3	
TERM	TFIDF	TERM	TFIDF	TERM	TFIDF
DAM (dam)	930.36	DAM (dam)	545.27	SUII (water level)	633.38
TISUI (flood control)	409.17	KAWAKAMI(upper river)	252.23	DAM (dam)	538.45
RYUEKI (basin)	362.82	RYUEKI (basin)	236.25	WORKING (working)	289.10
IKEN (opinion)	270.18	TISUI (flood control)	190.57	TISUI (flood control)	218.60
IIN (committee)	222.57	KASEN (river)	180.97	SOSA (operation)	211.80
KASEN (river)	187.21	IKEN (opinion)	146.12	SE (rapid)	202.79
MIZU (water)	169.62	SUIBOTSU (submergence)	117.70	KASEN (river)	180.97
SUIKEI (water system)	117.70	IIN (committee	112.32	RYUEKI (basin)	164.53
KANKYO (environment)	107.21	MONDAI (problem)	101.97	RISUI (irrigation)	132.62
ZYUMIN (resident)	104.44	RISUI (irrigation	98.24	GIRON (debate)	118.53

TABLE Ⅲ.

COMPARISON OF CASE 1 AND CASE 2

Case 1			Case 2		
Ex-rank	Term	TFIDF	Ex-rank	Term	TFIDF
12	BIWAKO (Biwa Lake)	102.00	7	SYUBOTSU (submergence)	117.71
13	TEIBO (bank)	95.90	12	KAMIRYU (the upper reaches)	92.82
32	YOSUI (water)	63.28	13	KAI (gorge)	84.08
34	KAIZEN (repair)	60.94	14	IDEN (moving)	80.09
37	SE (rapids)	58.59	19	KARYU (the lower reaches)	67.26
42	KADOU (river road)	52.56	32	KEIKAKU (plan)	41.16
44	KARYU (lower)	50.45	38	SEIBUTSU (life)	35.26
53	KAIHATSU (development)	41.35	40	KAISAKI (excavating)	33.63
54	KEIKAKU (plan)	41.16	41	KOKUDOKOTSUSYO (MLIT)	32.63
56	SIGA (Siga)	39.96	45	OTAKA (Accipiter gentilis)	28.03
58	KOUJI (construction)	39.26	47	HIGAI (damage)	27.28
66	BASAI (deforest-ation)	35.96	50	DAITAI (changing)	26.61
67	IDEN (moving)	35.59	52	TYOSA (inquiry)	25.49
69	HUKURYU (under flowing)	34.38	53	RISUI (water supply)	24.56
72	KODOMO (kids)	33.41	58	TISUI	22.53

similar, but the detailed debate content is different. Case 3 and regional meetings are then compared. Case 3 included professional and technical terms rather than plain terms. For example, terms like “water level”, “operation”, “irrigation” were included in case 2. That means that committee members might debate on specific problems of the project. Moreover, High rank terms of case 1 included “basin”, “opinion”, “committee.” It is possible to infer that mainly the “debate process” and “reflecting citizen’s opinions” might be debated and the terms might occur frequently.

The debate contents of case 1 and case 2 are compared more precisely. Same terms of high TFIDF score between two cases were removed within rank 20. Table Ⅲ shows the results. Terms with high TFIDF score in case 1 included “bank”, “river road”, “repair.” Terms with high TFIDF score in Case 1 included “moving”, “life”, “accipiter gentiles.” From this result, it is possible to infer that the main issue of case 1 might be flood control works i.e. construction of a dam. On the other hand, the main issue of case 2 might be the relocation of residents and the conversation of the environment. The result from the topic extraction can demonstrate the characteristics of each debated subjects well.

D. Comparison of semantic similarity of utterance

Using the co-occurrence data of individuals on public debate and the T-VSM, semantic similarities between different individuals were calculated as explained in previous subsections Ⅳ.A and Ⅳ.B. Among the topics extracted in subsection Ⅴ.C., “DAM (dam)”, “TISUI (flood control)” and “KANKYO (environment)” appear to be the central terms as topics. Co-occured terms of the three topics are extracted. In this study, words empty of meaning such as parse, particle, ancillary-verb, prefix, suffix, conjunction, number, pronoun, indexical are ignored.

Following, the average and dispersion of semantic similarities of case 1 and case 3 are compared, except case 2. This is because the average and the dispersion of the semantic similarities between case 1 and case2 are very close. Fig.3 depicts the average and dispersion of semantic similarity of co-occurred terms on topic “Dam”, “Flood control” and “Environment” between case 1 and case 3.

In case 1, as for topic “Dam”,(C1(Dam) shown in Fig. 3), the average of semantic similarities is very high, the dispersion of semantic similarities is very low. That might mean that there are lots of common cognitions between participants on the topic “Dam”. On the other hand, regarding topic “Environment” and “Flood control”, the averages of their semantic similarities are somewhere in the middle and their dispersions are very high. That might mean that there are lower common cognition and higher cognitive inconsistency around topics “Environment” and “Flood control” (c.f. topic “Dam” in case 1).

In case 3, both the average and the dispersion of the semantic similarities on topic “Dam” is middle. That might mean that there are some common cognitions and some conflicts among participants related to the topic “Dam”. The average of semantic similarities on topic “Environment” and “Flood Control” is low and its dispersion is high. That might mean that there are few common and high conflict cognitions comparing on topics “Environment” and “Flood control” in case 3.

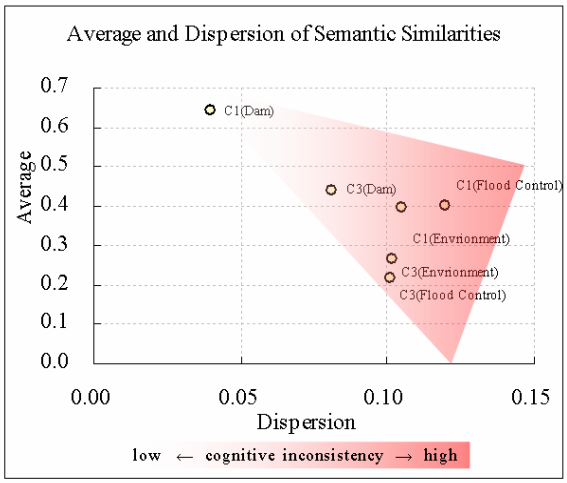

Figure 3. Average and Dispersion of Semantic Similarities

Regarding topics “Environment” and “Flood control”, the average of semantic similarities is low and the dispersion of the semantic similarities is high, which is commonly-observed in both case 1 and case 3. That might mean that these can easily cause conflicts among the participants as compared to “Dam”.

In this subsection, the semantic similarity of utterance between participants was described and compared. The result allows us to get an accurate understanding of the inconsistency level on each topic and the conflict contents in each case. Using the results, the debate structure such as the conflict structure partiality or in full can be analyzed in depth. This is presented in the next subsection.

E. Visualizing of debate structure with the semantic similarity

In this subsection, debate structures are visualized in a two-dimensional space by applying the MDS method which is described in the previous subsection Ⅳ.D. The MDS method represents two participants with high semantic similarity closely to each other in space.

Regarding representation, all participants are marked with colors and initials according to the participants classification as described in the subsection Ⅴ.B. The participant properties are described in Table Ⅰ. Regarding the colors, blue means a participant who has cons opinions of promoting the project. Red means a participant who has pros opinions of promoting the project. Green means a participant who stands neutral. As for the initials, Expert, Citizen, and Administrator are marked with E, C, and A respectively. To avoid cluttering on the graphs, E is omitted. Among participants, some of them did not express any opinion on some topic, in that

case the participants are not represented. We try to understand the debate structures with the distances between the participants.

Fig. 4, Fig. 5, and Fig 6 illustrate the representation of case 1. As mentioned in subsection Ⅴ.B, case 1 consisted of 8 experts and 4 citizens with 1 citizen and 5 experts have contrary opinions (cons opinions) on the promotion of project and 3 citizens and 1 expert have approval opinions (pros opinions). While the rest of them being neutral.

Fig 4 shows the debate structure on topic "Dam" in case 1. We can see the situation where almost all participants are located very closely to each other. Two participants are far from the main cluster and each other.

Comparatively, Fig 5 shows the debate structure on the topic "Environment" in case 1. It is shown that there is a cluster around pros citizens and natural experts. Participants, however, tend to be placed apart from each other and distributed impartially. Especially, there is a wide distance between citizens who have pros opinion and experts who have cons opinion on the project promotion. That might mean that there is a high level of cognitive inconsistency on the topic "Environment" between the pros citizens and cons experts.

Similarly, Fig 6 shows the debate structure on the topic "Flood Control" in case 1. It is shown that there is a cluster between several pros citizens and cons experts. It is however shown that there is an overall trend to have a wide distance between pros citizens and some cons experts on the project promotion.

The results allow us to get an accurate understanding of the conflict structure i.e. when the participants are likely to conflict and have high cognitive inconsistency among them. In case 1, regarding the topic "Dam", there is a particular conflict between two experts. Regarding topics "Environment" and "Flood Control", the pros citizens are likely to conflict with the cons experts. From results, we can infer that there is a cognitive dissonance between the citizens and some experts around the dam project in both aspects "Environment" and "Flood Control."

Fig. 7, Fig. 8, and Fig. 9 illustrate the representation of

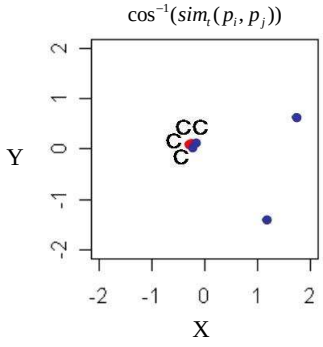

Figure 4. Topic "Dam" in case 1

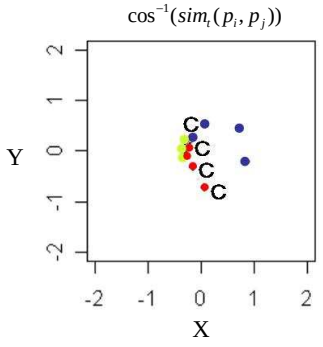

Figure 5. Topic "Environment" in case 1

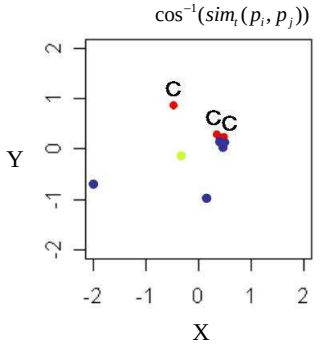

Figure 6. Topic "Flood Control" in case 1

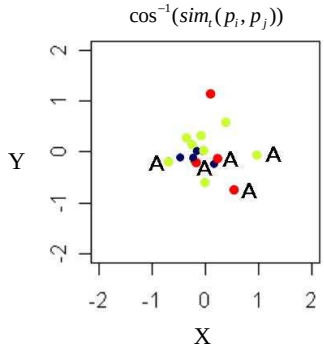

Figure 7. Topic "Dam" in case 3

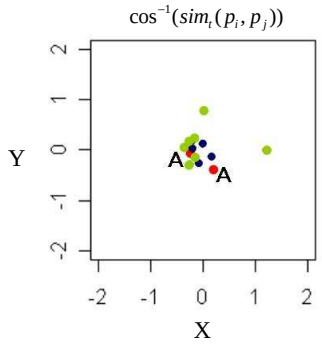

Figure 8. Topic "Environment" in case 3

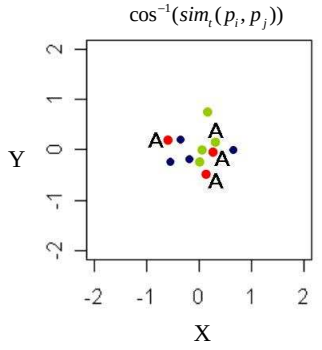

Figure 9. Topic "Flood Control" in case 3

case 3. Case 3 consisted of 13 experts and 5 administrators. Among them, 4 experts have contrary opinions on the promotion of project. While 3 administrators and 1 expert have approval opinions all. Others stand neutral.

Fig. 7 shows the debate structure of topic "Dam" in case 3. We can observe that the participants are dispersed widely . Even though the social roles of participants are same, there are wide distances. It worths nothing that the administrators are distributed most widely. That might mean that there is a serious cognitive dissonance on the topic "Dam" between them.

In detail, the pros administrators and a pros expert are far from each other. Conversely, the cons experts are very close to each other. In addition to that most participants are located near the cons expert cluster. That might mean that there is a common cognition around the opinion of cons experts. There is, however, a high cognitive inconsistency between the cons experts group and others.

Fig. 8 shows the debate structure on the topic "Environment" in case 3. There are no co-occured terms on the topic "Environment" in the utterances of the three administrators so that the three administrators are not presented in the space. Thus, the average of the semantic similarities is low and the dispersion is high as can be seen in Fig. 3. Fig. 8, however, shows that there is a cluster among the neutral opinions except two while the cons experts are far from each other. That might mean that there is a common cognition around the opinion of neutral experts. There is, however, some cognitive inconsistency between the neutral experts group and others.

Fig. 9 shows the debate structure of the topic "Flood Control" in case 3. It is shown that the participants are distributed very far apart. There is a distinctive trend of the participants who have the same social role to be located far from each other. The cons experts, the pros administrators and the neutral experts are far from each other so that there are no clusters. That might mean that the cognitive dissonance of the topic "Flood Control" between participants is very intense.

The result of case 3, regarding the topic "Dam" shows a cluster around the cons experts. Regarding the topic "Environment", there is a cluster around the natural experts. Regarding the topic "Flood Control", most participants are likely to conflict with each other. From the result, we can infer that there is a serious cognitive dissonance between all the participants around the dam project in regard of the term "Flood Control."

VI. Verification of Working Hypotheses & Discussions

Working hypotheses are verified with the results of the analysis using the proposed methodology in section Ⅴ.

First, Hypothesis 1: "The more serious conflict between any two participants, the less semantic similarity of their utterances," is supported in the analysis of case 1. Table Ⅳ shows the semantic similarities of utterances between the two groups in case 1.

It is shown that the semantic similarity of utterances between the groups is high in the following.

$$[C_{pros}\ E_{pros}] > [C_{pros}\ E_{cons}][E_{pros}\ E_{cons}]$$

In fact, groups of cons experts are strongly conflicting with pros groups, pros citizens and pros experts. Among the pros groups, pros citizens and pro experts, there is strong similar cognition. Therefore, the result supports hypothesis 1. The result, regarding the cons citizen, is very astonishing.

$$[C_{cons}\ C_{pros}] > [C_{cons}\ E_{pros}] > [C_{cons}\ E_{cons}]$$

Cons citizens have a high semantic similarities with pros citizens and pros experts rather than cons experts. That might be related to the lexical system rather their cognition on debated subject. That is, the cons citizens expressed their opinion with daily conversation words.

Secondly, Hypothesis 2 "The high cluster of opinion between participants, the more average of semantic similarity of their utterances," is supported in analysis of case 1 and case 2. The average semantic similarity is higher in the following order (See also Figure 3).

TABLE Ⅳ.

SEMANTIC SIMILARITIES BETWEEN TWO GROUPS

Dam			Environment			Flood control		
C_{cons}	C_{cons}	1.00	C_{cons}	C_{cons}	1.00	C_{cons}	C_{cons}	1.00
E_{pros}	E_{pros}	1.00	E_{pros}	E_{pros}	1.00	E_{pros}	E_{pros}	1.00
C_{pros}	C_{pros}	0.88	E_{neut}	E_{neut}	0.86	E_{neut}	E_{neut}	0.83
E_{neut}	E_{neut}	0.87	C_{pros}	C_{pros}	0.67	E_{pros}	E_{neut}	0.61
C_{cons}	C_{pros}	0.70	C_{cons}	E_{neut}	0.55	C_{pros}	C_{pros}	0.61
C_{pros}	E_{pros}	0.67	E_{pros}	E_{neut}	0.55	C_{cons}	E_{pros}	0.60
C_{cons}	E_{pros}	0.66	C_{pros}	E_{pros}	0.47	C_{cons}	E_{neut}	0.55
C_{cons}	E_{cons}	0.66	C_{cons}	E_{pros}	0.45	E_{cons}	E_{cons}	0.48
C_{pros}	E_{neut}	0.61	E_{cons}	E_{cons}	0.44	C_{pros}	E_{pros}	0.36
C_{cons}	E_{neut}	0.59	C_{pros}	E_{neut}	0.42	C_{cons}	C_{pros}	0.32
E_{pros}	E_{neut}	0.58	C_{cons}	C_{pros}	0.38	E_{cons}	E_{pros}	0.31
C_{pros}	E_{cons}	0.56	E_{cons}	E_{pros}	0.18	C_{cons}	E_{neut}	0.28
C_{cons}	E_{cons}	0.51	C_{cons}	E_{cons}	0.15	C_{cons}	E_{cons}	0.27
E_{cons}	E_{neut}	0.51	E_{cons}	E_{neut}	0.15	E_{cons}	E_{neut}	0.27
E_{cons}	E_{pros}	0.51	C_{pros}	E_{cons}	0.13	C_{pros}	E_{cons}	0.16

$$Case1_{dam} > Case3_{dam} > Case1_{floodcontrol}$$

$$> Case1_{environment} > Case3_{environment} > Case3_{floodcontrol}$$

The average of the semantic similarity is realized on the topic "Dam" in case 1. As mentioned in the previous section, on topic "Dam" in case 1, there is a very strong single group in which almost all participants are placed very close to each other. The second highest average of semantic similarity is on the topic "Dam" in case 3. The cons experts are very close to each other and most participants are distributed near the cons expert group. That is, there is a single group in which many participants are placed close to each other. The third highest average of semantic similarity is on the topic "Flood Control" in case 1. There is a grouping between several pros citizens and cons experts. The fourth highest average of semantic similarity is on the topic "Environment" in case 1. There is a grouping around the pros citizens and natural experts but they tend to be placed apart from each other, distributed impartially. That is, there is a weak grouping and conflict between cluster and the others. The fifth highest average of semantic similarity is on topic "Environment" in case 3. There is a cluster of neutral opinions but it is weak. The other participants are not close to the group. Finally, a low semantic similarity is realized on the topic "Flood Control" in case 3, where most participants are likely to conflict with each other. The level of grouping follows the order of the average of semantic similarity. These results support the hypothesis 2.

Finally, Hypothesis 3 "The worse unity and coherence of opinion between all participants, the more dispersion of semantic similarity of their utterances," is supported by the previous analysis results with Hypothesis 1 and Hypothesis 2 as follows.

(1) High Average & High Dispersion: There is a strong grouping and strong conflict between groups. It is demonstrable in the result of the topic "Dam" in case 3. There is a very strong grouping of pros experts and strong conflict between the pros expert group and other participants. As mentioned in section Ⅲ, there is a conflict so that the case of high average and high dispersion. That might mean the expressed opinions of participants are not well aggregated and coherent.

(2) High Average & Low Dispersion: There is a strong grouping and weak conflict between groups. That might mean that there is aggregation and coherence of opinion between all participants. This is demonstrable from the results of topic "Dam" in case 1. This case carries less conflict thus there is aggregation and coherence of opinion.

(3) Low Average & High Dispersion: There are few groupings and most participants strongly conflict with each other. That is, there is no unify and coherence of opinion. This is demonstrable from the results of topic "Flood Control" in case 1. That might mean that the expressed opinions of participants are not well aggregated and coherent due to strong conflict.

(4) Low Average & Low Dispersion: there are few groupings and most participant weakly conflict each other. This may be an unusual case. It is not correspond to any results in this study.

VII. Application, Significance and Problems.

A method of discourse analysis based on the corpus approach was effectively used to investigate the public debate and understanding the content and structure of a debate. The main issue of public debate and the inconsistency level with semantic similarity can be described using this method. As a result, the main concerns vary on different debates. There are diverse inconsistency levels on diverse debate issues. It is impossible to get consistency among all people. Finally, it is possible to infer the consensus time by getting the high level of consistency.

However, this method is not without constraints, i.e.; The utterance based approach has a limit in the understanding of the cognition of people. Most importantly, there are tacits and gestures that are not recorded in minutes. In addition, there are synonyms and homonyms that these terms can not be extracted automatically by using the computational method. Using that information imposes difficulties in the computational analysis. Finally, the semantic similarity is not necessarily same with the cognition similarity. If the semantic similarity of some people is constantly low related to other participants, the cognitive structures among the participants should be more carefully scrutinized.

References

[1] Y. Takubo., N. Yuji., H. Mitou., M. Kameyama., Y. Katagiri., *Discourse and Context,* Iwanami (in Japanese), 2004.

[2] M. Stubbs, *Corpus semantics*, Kenkyusya (in Japanese), 2006.

[3] S.E. Robertson and K. Spärck Jones, "Understanding Inverse Document Frequency: On theoretical arguments for IDF", *Journal of Documentation* 60 no.5, 2004, pp 503-520.

[4] J. Becker and D. Kuropuka, "Topic-based Vector Space Model", *Business Information Systems*, Proceedings off BIS 2003, Colorado Springs, USA

[5] S. Ikehara, J. Murakami, Y. Kimoto and T. Araki, "Vector Space Model based on Semantic Attributes of Words", *Natural Language Processing* ,Vol.1, No.1 , 1994.

[6] T. Hatori, K. Takahiro, K. Kawayoke, K. Kobayashi, Y. Natsume and E. Fujisaki, "Protocol analysis on public debate based on facet theory", *Journal of Infrastructure Planning and Management*, No.23, pp.91-102 (in Japanese), 2006.

[7] M. Horita and Y. Kanno, "Development an information-based CRANES for Public involvement Management", *Journal of Japan Society of Civil Engineers*, VI, Vol. 686,No. 52, pp.109-120 (in Japanese), 2001.

[8] K. Krippendorff, *Content Analysis; An Introduction to Its Methodology*, Beverely Hill, CA: Sage Publications, 1980.

[9] D.S. Schiffrin, *Approaches to Discourse*, Oxford: Blackwell, 1994.
[10] D.S. Schiffrin, D. Tannen and H.E. Hamilton, *The Handbook of Discourse Analysis*, Blackwell publishers, Malden, MA, 2003.
[11] M. Stubbs, *Words and Phrases: Corpus Studies of Lexical Semantics*, Blackwell Published Ltd, Oxford, 2002.
[12] M. Asahara and Y. Matsumoto, "Extended Models and Tools for High-performance Part-of-Speech Tagger", *In Proceedings of COLING*, 2000, pp. 21-27, 2000.
[13] K. Hasida, "Intellectual contents of all-round based on GDA meaning modification" *The transaction of Japanese Society for Artificial Intelligence*., Vol. 13, No.4, pp.528-535 (in Japanese), 1998.
[14] J.B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". *Psychometrica*, Vol. 29, pp.1-29, 1964.
[15] J.B. Kruskal, "Nonmetric multidimensional scaling: a numerical method". *Psychometrica*, Vol. 29, pp.115-129, 1964.
[16] X. Tai, F. Ren and K. Kita, "An information retrieval model based on vector space method by supervised learning", *Information Processing and Management* 38, pp.749–764, 2002
[17] B.S. Ong, "Towards Automatic Music Structural Analysis:Identifying Characteristic Within-Song Excerpts in Popular Music," *Universitat Pompeu Fabra Barcelona*, 2005
[18] D. Orpin, "Corpus Linguistics and Critical Discourse Analysis- Examining the ideology of sleaze", *International Journal of Corpus Linguistics* 10:1, pp37-61. 2005
[19] E. Amitay, "Using common hypertext links to identify the best phrasal description of target web documents", *Proc. SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval for the Web*, pp.271–276, 1998
[20] G. Salton, and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval" *Information Processing and Management*, 24(5): pp.513-523.1988.
[21] G. Salton, and M.J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill:NY. 1984.
[22] G. Qian, S. Sural, Y. Gu and S. Pramanik , "Similarity between Euclidean and cosine angle distance for nearest neighbor queries" *Proceedings of the 19th Annual ACM Symposium on Applied Computing*, 2004,.
[23] M. Rajapakse, J. Tan and J. Rajapakse, "Color channel encoding with NMF for face recognition", *IEEE ICIP*, Vol. 3,pp 2007- 2010, 2004
[24] T. Koreniusa, J. Laurikkalaa, and M. Juhola, "On principal component analysis, cosine and Euclidean measures in information retrieval", *Information Sciences*, Vol 177, Issue 22, 15, pp. 4893-4905. 2007
[25] W. Choi and S.K. Das, "A proxy based indirect routing scheme for ad hoc wireless networks", *IEEE, INFOCOM*, Vol. 3, pp. 1395-1404. 2002
[26] H. Chen and B. Bhanu "3D free-form object recognition in range images using local surface patches", *Pattern Recognition Letters*, Vol 28, Issue 10, pp.1252-1262, 2007.

Hayeong JEONG was born in Korea 1978. She received the B.S. and M.S. degrees in Engineering from Dong-A University, Busan, Korea, in 2001 and 2003, respectively.
She is currently working toward the Ph.D degree in Planning and Management Systems, Graduate School of Engineering at Kyoto University in Kyoto, Japan. Her current research interests include Public Involvement, Trust Formation, Data Mining, Facet Theory. Her major publications are "The Open Public Debate and the Impacts upon Trust Formation" (Tsuyoshi HATORI, Hayeong JEONG, Kiyoshi KOBAYASHI, *Journal of Japan Society of Civil Engineers D* Vol.64 No.2, pp.148-167, 2008) and "Discourse Analysis of Public Debates: A Corpus-based Approach"(Hayeong JEONG, Tsuyoshi HATORI, Kiyoshi KOBAYASHI, *IEEE Systems, Man, and Cybernetics Conference*, pp.1782-1793, 2007).

Shun SHIRAMATSU was born in Japan 1976. He received the Ph.D. degree in Informatics from Kyoto University, Kyoto, Japan, in 2008.
He is currently a JSPS Postdoctoral Fellow in Graduate School of Informatics at Kyoto University in Japan. His current research interest is focused on visualizing the dynamics of discourse context. His major publications are "Meaning Games" (Koiti HASIDA, Shun SHIRAMATSU, Kazunori KOMATANI, Tetsuya OGATA, and Hiroshi G. OKUNO, *New Frontiers in Artificial Intelligence, Lecture Notes in AI*, Vol. 4914, pp. 228-241, 2008) and "Meaning-game-based centering model with statistical definition of utility of referential expressions and its verification using Japanese and English corpora" (Shun SHIRAMATSU, Kazunori KOMATANI, Koiti HASIDA, Tetsuya OGATA, and Hiroshi G. OKUNO, *Proceedings of DAARC2007*, pp. 121-126, 2007)

Kiyoshi KOBAYASHI was born in Japan 1953. He received the Ph.D degree in Engineering from Kyoto University in Japan 1984.
He is currently a professor at Kyoto University. His major publications are "Structural Change in Transportation and Communications in the Knowledge Society," (Kobayashi, K., Lakshmanan, T.R., and Anderson, W.P., Edward Elgar 2007) and "The Management and Measurement of Infrastructure" (Karlsson, C., Anderson, W.P., Johansson, B. and Kobayashi, K., Edward Elgar, 2007). He is member of International Association of Regional Science, Japan Society of Civil Engineers. He was awarded JSCE (Japan Society of Civil Engineers) Research Award in 1993, 2001, and 2007. He currently works for professional committees as follows; Editor-in-Chief, Journals of JSCE, Vice Editor in chief, Journal of Infrastructure Systems Editorial board member, Annals of Regional Science and International Journal of Regional Science.

Tsuyoshi HATORI was born in Japan 1980. He received B.S., M.S. and Ph.D. degrees in Engineering from Kyoto University in 2002, 2004 and 2006, respectively.
He is currently an assistant professor at Tokyo Institute of Techonology, Tokyo, Japan. His major publications are "Social Capital and Development Trends in Rural Areas" (Ito, K., Westlund, H., Kobayashi K. and Hatori, T. (eds.), Vol.2, MARG, 2006), "Knowledge, political innovation, and referendum, in: Johansson, I., (ed.) Institutions for Knowledge Generation and Knowledge Flows - Building Innovative Capabilities for Regions" (Hatori, T. and Kobayashi, K., pp.471-488, University West, Sweden, 2007) and "Third party reviews and trust formation" (Hatori, T. and Kobayashi, K., *Proceedings of the 2006 IEEE Systems, Man, and Cybernetics Conference*, 2006).