Profiling of Participants in Online-Community

Naohiro Matsumura
PRESTO, JST
The University of Tokyo
Tokyo, 113–8656 Japan
matumura@miv.t.u-tokyo.ac.jp

Yukio Ohsawa
PRESTO, JST
University of Tsukuba
Tokyo, 112–0012 Japan
osawa@gssm.otsuka.tsukuba.ac.jp

Mitsuru Ishizuka
The University of Tokyo
Tokyo, 113–8656 Japan
ishizuka@miv.t.u-tokyo.ac.jp

Abstract

Promoting interactions among participants in an online-community is catching attention of web sites' managers. In this paper, we first introduce Influence Diffusion Model (IDM), a method for discovering influential comments, participants and terms from threaded online discussions, and evaluate the performance by precision and recall measurement. Then we propose a new method for profiling of participants in an online-community by expanding the idea of IDM. The positioning maps derived from the profiles show the relations among participants as well as their characteristics.

Introduction

Communication places on the Internet, such as BBS, chat room etc. are designed to gather people into web sites by promoting the interactions among participants (Ishikawa 2001). Previously we had proposed Influence Diffusion Model (IDM) (Matsumura, Ohsawa, & Ishizuka 2002) which can discover influential participants, comments and terms from threaded online discussions. In this paper, we propose a method for profiling of participants in an online-community by expanding the idea of IDM. Here, we regard a set of influential terms of a participant as his/her profile. The profiles help us understand their characteristics by which we can manage the participants for aiding their interaction in the community.

Influence Diffusion Model

Firstly, we introduce IDM which is designed to measure the influence of comments, participants and terms by the degree of text-based relevance of comments under the situation that interactions among participants are done by exchanging comments, i.e., posting new comments or replying to the comments. The threaded comments, called *comment-chain*, show the flow of influence. For example, if a comment C_y replies to a comment C_x, it is considered that C_y is affected by C_x. Similarly, if a participant P_y replies to a comment of a participant P_x, P_y is considered to be affected by P_x. In these cases, the influence diffuses from C_x to C_y / from P_x to P_y. Here, IDM defines the process of diffusion of influence as follows.

Definition 1 *In text-based communication, influence diffuses along the comment-chain by medium of terms, i.e., words or phrases.*

On the basis of *Definition 1*, the influence is defined by the degree of terms propagating throughout the comment-chain. For example, If C_y replies to C_x, the influence of C_x onto C_y, $i_{x,y}$, is defined as

$$i_{x,y} = \frac{|w_x \cap w_y|}{|w_y|}, \tag{1}$$

where w_x and w_y are the set of terms in C_x and C_y respectively, and $|w|$ denotes the count of w.

In addition, if C_z replies to C_y, the influence of C_x onto C_z through C_y, $i_{x,z}$, is defined as

$$i_{x,z} = \frac{|w_x \cap w_y \cap w_z|}{|w_z|} \cdot i_{x,y}, \tag{2}$$

where w_z are the terms in C_z.

It is considered that the more a comment affects other comments, the more the influence increases. And the same can be applied to the influence of participant/term. The influence of a subject (including comment, participant or term) then comes to be measurable.

Definition 2 *The influence of a subject (comment, participant or term) to the community is measured by the sum of influence diffused from the subject to all other members of the community.*

Applying *Definition* 2 to C_x, the influence (note that IDM ignores "to other members of the community") is measured by the sum of influence diffused from C_x, i.e., $i_{x,y} + i_{x,z}$ if the community has three members x, y and z. Likewise, the influence of a participant P_x is measured by the sum of influence of P_x's comments. The influence of a term t is also measured by the sum of influence mediated by t.

In the followings, we show some examples of measuring influences of comments, participants, and terms.

Measuring the Influence of Comments

For example, we consider a sample comment-chains illustrated in Fig. 1 where C_1 is replied to by C_2 and C_3, and C_2 is replied to by C_4. In this case, term A, C are propagating from C_1 to C_2, term B is propagating from C_1 to C_3, and term C is propagating from C_2 to C_4. Here, the influence of C_1 is calculated as follows.

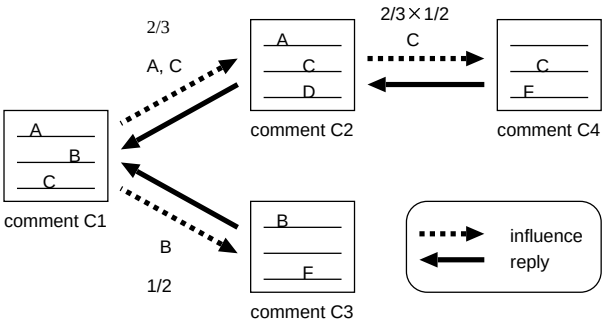

Figure 1: An example of comment-chains.

The influence of C_1 onto C_2: The count of propagated terms from C_1 to C_2 is two (A, C), and the count of terms in C_2 is three (A, C, D). Then, the influence from C_1 to C_2 becomes $2/3$.

The influence of C_1 onto C_3: The count of propagated terms from C_1 to C_3 is one (B), and the count of terms in C_2 is two (B, F). Then, the influence from C_1 to C_3 becomes $1/2$.

The influence of C_1 onto C_4 through C_2: The count of propagated terms from C_1 to C_4 via C_2 is one (C), and the count of terms in C_2 is two (C, F). Considering that the influence of C_1 onto C_2 is $2/3$, the influence of C_1 onto C_4 via C_2 becomes $2/3 \times 1/2 = 1/3$.

According to *Definition* 2, the influence of C_1 in Fig. 1 is calculated as *(the influence from C_1 to C_2) + (the influence from C_1 to C_3) + (the influence from C_1 to C_4)* $= 2/3 + 1/2 + 1/3 = 3/2$. Similarly, the influence of C_2, C_3 and C_4 are calculated as $1/2$, 0 and 0 respectively. Therefore, C_1 is selected as the most influential comment in Fig. 1.

Measuring the Influence of Participants

Next, let us measure the influence of participants, by assuming that C_1, C_2, C_3 and C_4 in Fig. 1 are posted by P_1, P_2, P_3 and P_3 respectively. The relations of participants, called *human network*, is illustrated in Fig. 2.

Here, the influence of P_1 onto P_2 is equal to the influence of C_1 onto C_2, and the influence of P_1 onto P_3 is the sum of the influence of C_1 onto C_4 via C_2 and of C_1 onto C_3. Referring to the above results, the influence of P_1 onto P_2 becomes $2/3$, and the influence from P_1 to P_3 becomes $2/3 \times 1/2 + 1/2 = 5/6$. Then, the influence of P_1 is calculated as *(the influence from P_1 to P_2) + (the influence from P_1 to P_3)* $= 2/3 + 5/6 = 3/2$. Likewise, the influence of P_2 and P_3 are calculated as $1/2$ and 0 respectively. From these calculations, we can understand that P_1 is the most influential participant in Fig. 2.

Measuring the Influence of Terms

IDM assumes that all terms equally mediate the influence, and the influence of a term is calculated by the sum of influence mediated by the term throughout comment-chains. Referring to Fig. 1, the inluence of A becomes $2/3 \times 1/2 = 1/3$ because A mediates $2/3$ influence together with C. In the same way, the influence of B becomes $1/2$, and the influence of C becomes $(2/3 \times 1/2) + (2/3 \times 1/2) + 1/2 = 7/6$. The influence of other terms (D, E, F) becomes 0. Then, the most influential term in Fig. 1 becomes C.

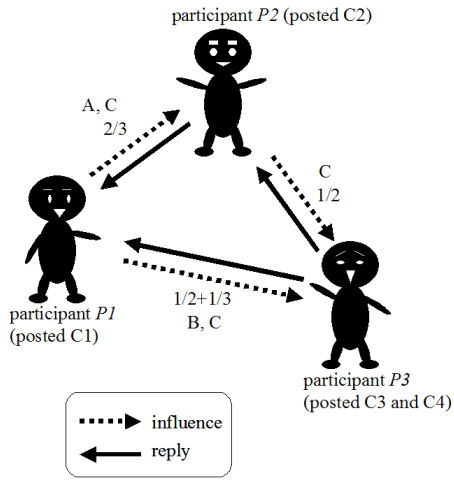

Figure 2: Human network in Fig. 1.

Evaluation of IDM

The notion of IDM was proposed in (Matsumura, Ohsawa, & Ishizuka 2002), however, the experimental evaluations were somewhat immature. Here, we evaluate the performance of IDM by using *precision* (the ratio of the correctly extracted targets to the extracted targets) and *recall* (the ratio of correctly extracted targets to the targets that should be extracted) which are the standard measurement for information retrieval. The precision and recall results are also plotted as precision/recall (P/R) curves (Buckland & Gey 1994) in order to analyze the retrieval performancs. The better the retrieval performance, the more convex the curve.

We analyzed a Bulletin Board Service (BBS) [1] in which participants were exchanging local information such as about a good coffee shop, a cherry blossom-viewing picnic etc. The number of comments was 250, and the number of participants was 47. Note that the comments were written in Japanese, there was no space between words. We in advance converted all the comments into morphemes and remained only nouns by using Morphological Analyzer ChaSen (Matsumoto *et al.* 1999). Then, we applied IDM. In the following, experimental results are translated from Japanese into English as the case may be.

Before experiments, we read all the comments of the BBS thoroughly. Then, based on our intuition, we picked up 28 influential comments, 9 influential participants and 56 influential terms that should be extracted.

[1] http://www.machibbs.com/

Evaluation of Influential Comments

We extracted 20 comments by IDM, and as a comparison, we also extracted 20 comments by Reply-Index (RI) that extracts comments having much replies. RI is a widely used approach for ranking popular comments in BBS. The precision and recall values were shown in Table 1, and the P/R curves were shown in Fig. 3. The top curve corresponds to the retrieval performance of IDM, and the bottom curve corresponds to the retrieval performance of RI. From Fig. 3, we can clearly understand that IDM was superior to RI.

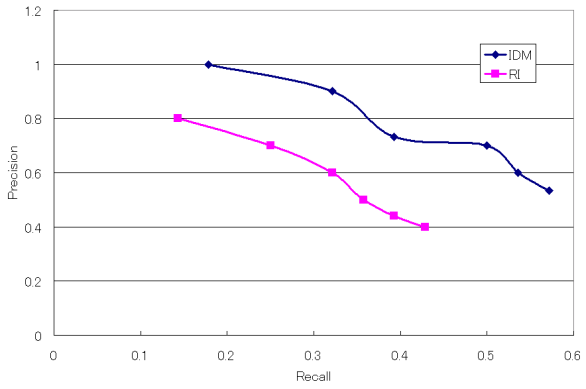

Figure 3: Precision/Recall curves for extracted comments.

Evaluation of Influential Participants

Next, we extracted 10 participats by IDM. As a comparison, we also extracted 10 participants by RI and 10 participants by Post-Index (PI) that extracts frequently posting participants. PI is a conventional approach to extract talkative participants. The precision and recall values were shown in Table 2 and the P/R curves are shown in Fig. 4. The curves were wavy, however, the curve of IDM were keeping the top. This mean that IDM is better than RI and PI.

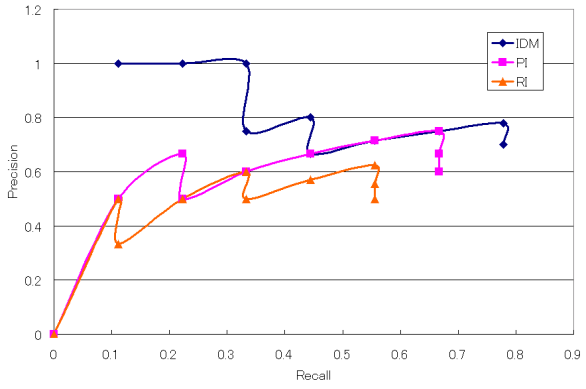

Figure 4: Precision/Recall curves for extracted participants.

Evaluation of Influential Terms

Finally, we extracted 100 terms by IDM, TF (Term Frequency) (Luhn 1957) and TFIDF (Term Frequency Inverse Document Frequency) (Salton & McGill 1983). TF and TFIDF are the widely used approaches for information retrieval. The corpus used by TFIDF was made from the electronic articles of Mainichi newspapers in 1998 and 1999. A total 236600 articles composed of 164790 kinds of words were collected. The results of precision and recall values were shown in Table 3, and P/R curves were shown in Fig. 5. As you can see from the Fig. 5, the curve of IDM is obviously upper than the curves of TF and TFIDF. This means the superiority of IDM.

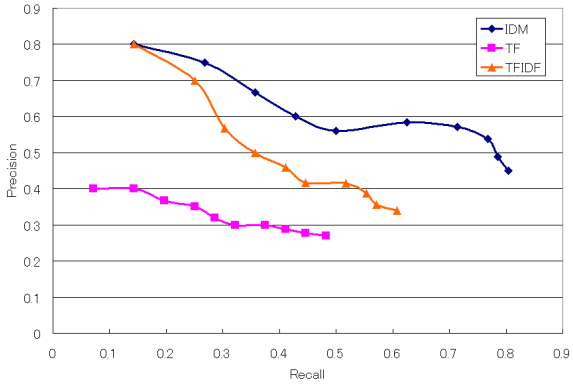

Figure 5: Precision/Recall curves for extracted terms.

Profiling of Participants

IDM measures the influence of comments, participants and terms. By expanding the idea of IDM, we propose a method for profiling of participants in an online-community. Here, we regard a set of influential terms posted by a participant as his/her profile. Referring to the examples of Fig. 1 and Fig. 2, term A of participant P_1 affected the participant P_2, whereas term A of participant P_2 did not affect anyone. In this case, A could be an element of P_1's profile. We can make the profiles like this.

In the followings, let us show three examples of profiling of P_1, P_2 and P_3.

The profile of P_1: The terms used by P_1 were A, B and C. The influence of A of P_1 was $2/3 \times 1/2 = 1/3$. The influence of B of P_1 was $1/2$. The influence of C of P_1 was $2/3 \times 1/2 + 2/3 \times 1/2 = 2/3$. Then, the profile of P_1 becomes $(\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}, \mathsf{E}, \mathsf{F}) = (1/3, 1/2, 2/3, 0, 0, 0)$.

The profile of P_2: The terms used by P_2 were A, C and D. The influence of C was $1/2$. The influence of A and D was 0. Then, the profile of P_1 becomes $(\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}, \mathsf{E}, \mathsf{F}) = (0, 0, 1/2, 0, 0, 0)$.

The profile of P_3: The terms used by P_3 were B, F (in comment C_3) and C, F (in comment C_4). However, the influence of these terms were 0. Then, the profile of P_1 becomes $(\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}, \mathsf{E}, \mathsf{F}) = (0, 0, 0, 0, 0, 0)$.

Table 1: Precision and recall values for extracted comments.

Num. of comments	*IDM*		*RI (Reply-Index)*	
	Precision	*Recall*	*Precision*	*Recall*
5	1.0	0.18	0.80	0.14
10	0.90	0.32	0.70	0.25
15	0.73	0.39	0.60	0.32
20	0.70	0.50	0.50	0.36
25	0.60	0.54	0.44	0.39
30	0.53	0.57	0.40	0.43

Table 2: Precision and recall values for extracted participants.

Num. of participants	*IDM*		*PI (Post-Index)*		*RI (Reply-Index)*	
	Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
1	1.0	0.11	0.0	0.00	0.00	0.00
2	1.0	0.22	0.50	0.11	0.50	0.11
3	1.0	0.33	0.67	0.22	0.33	0.11
4	0.75	0.33	0.50	0.22	0.50	0.22
5	0.80	0.44	0.60	0.33	0.60	0.33
6	0.67	0.44	0.67	0.44	0.50	0.33
7	0.71	0.56	0.71	0.56	0.57	0.44
8	0.75	0.67	0.75	0.67	0.63	0.56
9	0.78	0.78	0.67	0.67	0.56	0.56
10	0.7	0.78	0.60	0.67	0.50	0.56

Table 3: Precision and recall values for extracted terms.

Num. of terms	*IDM*		*TF*		*TFIDF*	
	Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
10	0.80	0.14	0.40	0.07	0.80	0.14
20	0.75	0.27	0.40	0.14	0.70	0.25
30	0.67	0.36	0.37	0.20	0.57	0.30
40	0.60	0.43	0.35	0.25	0.50	0.36
50	0.56	0.50	0.32	0.29	0.46	0.41
60	0.58	0.63	0.30	0.32	0.42	0.44
70	0.57	0.71	0.30	0.38	0.41	0.52
80	0.54	0.77	0.29	0.41	0.39	0.55
90	0.49	0.79	0.28	0.45	0.36	0.57
100	0.45	0.80	0.27	0.48	0.34	0.61

Algorithm

The algorithm of profiling of participants are as follows.

The influence of comment C_i diffuses along the comment-chains by the medium of terms (*Definition* 1), and the influence is measured by the sum of influence diffused from C_i (*Definition* 2). Here, let $\xi_{i,z}$ be the comment-chain which starts from C_i, i.e., $\xi_{i,z} = \{C_i, C_j, C_k \cdots C_q, C_r \cdots C_y, C_z\}$ $\{i < j < k \cdots q < r \cdots y < z\}$, and the influence of C_i onto C_r be $i_{i,r}$. Then, $i_{i,r}$ is described as

$$i_{i,r} = \frac{|w_i \cap w_j \cap \cdots \cap w_r|}{|w_r|} \cdot i_{i,q}, \tag{3}$$

where $|w_r|$ denotes the count of terms in C_r, and $|w_i \cap w_j \cap \cdots \cap w_r|$ denotes the count of propagated terms from C_i to C_r. Eq. (3) means that $i_{i,q}$ affects $i_{i,r}$ in proportion to the count of propagated terms from C_i to C_r in the count of terms in C_r.

IDM assumes that all terms equally mediate the influence of comments. In case of $i_{i,r}$, the influence is propagated by the medium of $|w_i \cap w_j \cap \cdots \cap w_r|$. Here, the sender of C_i be P_x, the influence of $t \in \{w_i \cap w_j \cap \cdots \cap w_r\}$, j_{i,r,t,P_x}, is described as

$$j_{i,r,t,P_x} = \frac{1}{|w_i \cap w_j \cap \cdots \cap w_r|} \cdot i_{i,r}. \tag{4}$$

j_{i,r,t,P_x} means the influence of t of P_x from C_i to C_r in ξ. Here, let J_{ξ,t,P_x} be the influence of t of P_x in ξ. Then, J_{ξ,t,P_x}, which is measured by the sum of j_{i,r,t,P_x} in ξ, is described as

$$J_{\xi,t,P_x} = j_{i,j,t,P_x} + j_{i,k,t,P_x} + \cdots + j_{i,y,t,P_x} + j_{i,z,t,P_x}. \tag{5}$$

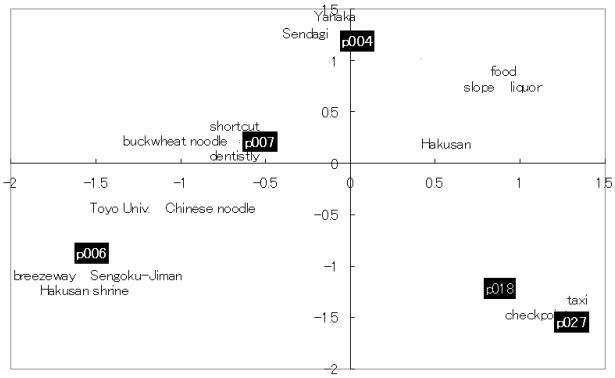

Figure 6: Positioning map of top 5 influential participants and their profiles (For ease of understanding, only limited number of terms were plotted).

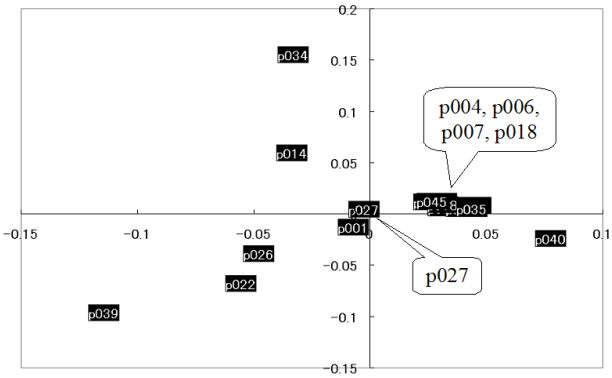

Figure 7: Positioning map of all the participants.

The influence of t of P_x is defined as the sum of J_{ξ,t,P_x} for all the comment-chains including t. Let the influence of t of P_x be D_{t,P_x}, and all the comment-chain including t be ξ_t. Then, D_{t,P_x} is defined as

$$D_{t,P_x} = \sum_{\xi \in \xi_t} J_{\xi,t,P_x}. \quad (6)$$

The profile of P_x can be made by extracting a set of terms of high D_{t,P_x} values.

Case Study

Profiles

We applied the profiling algorithm to the same BBS used for the evaluation of IDM. Here we extracted 20 terms of high D_{t,P_x} values for each participants. In case that the number of extracted terms was less than 20, i.e., only no more than 20 terms were propagated, we made up the deficient terms by TFIDF. The extracted profiles were shown in Table 4. Note that we extracted the profiles of only 24 participants because the comments of other participants were too poor to be analyzed.

Positioning Map of Top 5 Influential Participants

For understanding the relations of influential participants and their characteristics, we employ correspondence analysis (Miyagawa 1997) to visualize the relations as a two-dimentional positioning map. We skip the details of correspondence analysis because it is beyond the scope of the paper. Fig. 6 shows the positioning map of top 5 influential participants and their profiles. By seeing Fig. 6, we can clearly understand the relations of influential participants as well as their characteristics.

For example, `p007` is in the central position among them. This mean that `p007` can follow various topics such as `Chinese noodle`, `Hakusan` etc. effectively in the community. On the other hand, each of `p004` and `p006` has specific characteristics like `Hakusan shrine`, `Yanaka` etc. Also, `p018` and `p027` have the similar characteristics but thier characteristics are specific compared to others.

Positioning Map of All the Participants

Next, we analyzed all the participants and their profiles in Table 4 by correspondence analysis to understand the roles of top 5 influential participants in the community. The positioning map of participants, shown in Fig. 7, revealed the relations that participant `p004`, `p006`, `p007` and `p018` had similar characteristics. Whereas the characteristics of participant `p027` was rather different from above participants. While, we can understand the lack of influential participants around most of the participants.

As a plan for promoting the interactions among participants, we can realize that the manager of the community should hunt some influential people who are familiar with the topic around participants `p014`, `p022`, `p026`, `p034` and `p039` because there were no influential participants around them.

Conclusion

In this paper, we have first confirmed the good performance of IDM by precision and recall measurement. Then, we have proposed a new method for profiling of participants in an online-community by expanding the idea of IDM. The performance of profiles has been evaluated by interpreting the positioning maps derived from the profiles. The positioning maps have clearly showed the relations among participants as well as their characteristics. We are currently considering an application for promoting interaction in an online-community by managing influential participants according to coming topics.

Understanding the mechanism of human behaviors in the community is one of the main topics of Chance Discovery(Ohawa 2002). We believe that profiling of participants in an online-community will have a great impact on the field of Chance Discovery.

Table 4: Profiles of all the participants (24 participants).

Ranking	*Participant*	*IDM*	*Profile*
1	p006	1.722	Chinese noodle, breezeway, Hakusan shrine, · · ·
2	p007	1.689	dentistry, Chinese noodle, buckwheat noodle, · · ·
3	p004	1.647	Hakusan, bike, Yanaka, Sendagi, slope, liquor, · · ·
4	p018	0.731	Hakusan, bar, liquor, shop, slope, checkpoint, · · ·
5	p027	0.607	Odawaraya, checkpoint, Hakusan, time, Taxi · · ·
⋮	⋮	⋮	⋮
24	p039	0.00	hobby, Sendai, Miyagi, Kanagawa, Yokohama, · · ·

References

Buckland, M., and Gey, F. 1994. The relationship between recall and precision. *Journal of the American Society for Information Science* 45:12–19.

Ishikawa, N. 2001. *Internet Community Strategy (In Japanese)*. SoftBank Publishing.

Luhn, H. 1957. A statistical approach to the mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 1(4):309–317.

Matsumoto, Y.; Kitauchi, A.; Yamashita, T.; and Hirano, Y. 1999. Japanese morphological analysis system chasen version 2.0 manual. In *NAIST Technical Report, NAIST-IS-TR99009*.

Matsumura, N.; Ohsawa, Y.; and Ishizuka, M. 2002. Influence diffusion model in text-based communication. In *Posters of the Eleventh International World Wide Web Conference (WWW2002)*.

Miyagawa, M. 1997. *Graphical Modeling (In Japanese)*. Asakura Publisher.

Ohawa, Y. 2002. Chance discovery for making decisions in complex real world. *New Generation Computing* 20(2).

Salton, G., and McGill, M. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.