DELINEATING TOPIC AND DISCUSSANT TRANSITIONS IN ONLINE COLLABORATIVE ENVIRONMENTS

Noriko Imafuji Yasui[1], Xavier Llorà[2], David E. Goldberg[1]
[1]*IllGAL, University of Illinois at Urbana-Champaign*
104 S.Mathews Ave., Urbana, IL 61801, USA
[2]*NCSA, University of Illinois at Urbana-Champaign*
1205 W.Clark St., Urbana, IL 61801, USA

Yuichi Washida and Hiroshi Tamura
Research and Development Division, Hakuhodo Inc.
Tokyo 108-8088, Japan

Keywords: Online collaborative environments, Innovation and Creativity Support, Topic transition, HITS.

Abstract: In this paper, we propose some methodologies for delineating topic and discussant transitions in online collaborative environments, more precisely, focus group discussions for product conceptualization. First, we propose *KEE* (Key Elements Extraction) algorithm, an algorithm for simultaneously finding *key terms* and *key persons* in a discussion. Based on *KEE* algorithm, we propose approaches for analyzing two important factors of discussions: discussion dynamics and emerging social networks. Examining our approaches using actual network-based discussion data generated by real focus groups in a marketing environment, we report interesting results that demonstrate how our approaches could effectively discover knowledge in the discussions.

1 INTRODUCTION

In the last decades, mainstream communication has been shifted to network-based ones. Network-based communications which enable a great number of diverse people to join collaborative discussions are rich repositories of innovative and creative ideas. The methodological approaches for modeling, measuring and analyzing the network-based communications have become key elements to success in the fields of decision making, problem solving, and total quality management.

A goal of this paper is to delineate topic and discussant transitions in online collaborative environments, more precisely, focus group discussions for product conceptualization. Toward the goal, first we propose *KEE* algorithm for simultaneously finding key terms and key persons in network-based discussions. A *key term* is a significant word- or phrase-indicative of innovative and creative ideas. A *key person* is a significant participant having innovative and creative ideas or potential for producing them. We suppose a network-based discussion is (1) held for enhancing innovation and creativity toward product conceptualization, (2) based on participants posting and replying messages (3) on online message boards or chat rooms. Those discussions are made several attempts with different focus groups.

One of the biggest advantages of the KEE algorithm is its high applicability. We propose and examine approaches based on the KEE algorithm for analyzing discussions with two important factors: *discussion dynamics* and *social network*. Observing topic transitions and detecting topic segmentations lead to making sense out of key terms, which helps us figuring out the building block of innovative and creative ideas (Llorà et al., 2006). Detecting relationship between participants based on their significances lead to grasping diffusion of key persons, which helps us figuring out who had the innovative and creative ideas. Both analyses are essential in planning strategies for further discussions, including discussion theme setting and re-grouping of participants.

The reminder of this paper is organized as follows. Section 2 proposes the KEE algorithm, which is a core algorithm for innovation and creativity oriented mining from discussions. In Section 3, we propose two approaches for analyzing discussions: discussion dynamics and social networks. Section 4 reports the experimental results with using real data collected in real focus groups. Finally, this paper concludes in Section 5 with summarizing and directions for future work.

Imafuji Yasui N., Llorà X., E. Goldberg D., Washida Y. and Tamura H. (2007).
DELINEATING TOPIC AND DISCUSSANT TRANSITIONS IN ONLINE COLLABORATIVE ENVIRONMENTS.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - AIDSS*, pages 14-21
DOI: 10.5220/0002381000140021

2 KEYS EXTRACTION BY KEE

Our goal is to find out *key terms* and *key persons* from network- and text-based discussions. Suppose several discussions are held with different groups of people. In this section, we propose KEE (*key elements extraction*) algorithm. KEE enables us to find key persons and key terms simultaneously. We also propose two metrics used for the term weight: a *generality* and a *particularity*.

2.1 Kee Algorithm

KEE (*Key Elements Extraction*) is an algorithm for finding *key persons* and *key terms* of a discussion by scoring participants and terms in the context of their *significance* in discussions. Higher scored participants are *key persons* having innovative and creative ideas or potential for producing them. Higher scored terms are *key terms* indicating or leading to innovative and creative ideas.

KEE is based on the idea of mutually reinforcing relationship between participants and terms: significant participants are the participants using many significant terms, and conversely, significant terms are the terms used by many significant participants. KEE uses HITS (Hyperlink-Induced Topic Search) algorithm (Kleinberg, 1999) in an unintended way. HITS is an algorithm for ranking web pages in terms of *hubs* and *authorities*. KEE is an algorithm applying HITS framework to text mining, and obtains scores for ranking participants and terms by an iterative calculation.

A discussion is represented by a weighted directed bipartite graph $G(V,E)$ where V and E are sets of nodes and weighted edges, respectively. Let V_P be a set of participants of the discussion, and V_T be a set of terms used by the participants. $V = V_P \cup V_T$, $V_P \cap V_T = \phi$. Let (p_i, t_j) and $w(p_i, t_j)$ denote an edge between $p_i \in V_P$ and $t_j \in V_T$ and its weight, respectively. $w(p_i, t_j) = m$, if the participant p_i used the term t_j m times.

Participants and terms are ranked by *key scores* of participants (or *participant scores* for short) and *key scores* of terms (or *term scores* for short). Let $s(p_i)$ and $s(t_i)$ denote the key score of participant p_i and the key score of term t_i, respectively. Similarly to HITS algorithm (Kleinberg, 1999), the mutually reinforcing relationship in KEE algorithm are as follows: If the participant p_i had used many terms with high key scores, then he/she should receive a high participant score; and if the term t_i had been used by many participants with high key score, then the term should receive a high term score.

KEE algorithm obtains participant and term scores simultaneously by an iterative calculation. Given participant score $s(p_i)$ and term score $s(t_j)$, $s(p_i)$ and $s(t_j)$ are updated by the following calculations. $\alpha(t_j)$ is a weighting factor for the term t_j, which will be argued in the next sub section.

$$s(p_i) \leftarrow \sum_{(p_i,t_j)\in E} s(t_j) \cdot w(p_i,t_j) \cdot \alpha(t_j) \qquad (1)$$

$$s(t_i) \leftarrow \sum_{(p_i,t_j)\in E} s(p_i) \cdot w(p_i,t_j) \cdot \alpha(t_j) \qquad (2)$$

KEE algorithm is as follows. A vector of participant scores and a vector of term scores are represented by S_P and S_T respectively. k in the below is a natural number.

```
KEE algorithm:
 1. Initialize $S_P^0 = 1,1,\ldots,1$, and $S_T^0 = 1,1\ldots,1$
 2. For $i = 1,2,\ldots,k$
    (a) $S_P^i$ is obtained using Equation (1) with $S_T^{i-1}$
    (b) Normalize $S_P^i$ so the square sum in $S_P^i$ to 1
    (c) $S_T^i$ is obtained using Equation (2) with $S_P^i$
    (d) Normalize $S_T^i$ so the square sum in $S_T^i$ to 1
 3. Return $S_P^k$ and $S_T^k$
```

Kleinberg proved theorems that S_P and S_T converge and the limits of S_P^k and S_T^k are obtained by the principal eigenvectors of A^TA and AA^T (Kleinberg, 1999). A is an adjacency matrix; (i,j) entry is 1 if $(p_i,t_j) \in E$, and is 0 otherwise. Empirically, S_P and S_T converge very rapidly ($k = 6$ on the average in our experiments).

2.2 Term Weight Assignment

In the previous subsection, we described how KEE algorithm obtains key terms and key persons simultaneously. This subsection covers how to assign the term weight $\alpha(t)$ (see Equation (1) and Equation (2)).

A term weight is a tuning parameter in order to avoid strong influences by frequent terms. KEE algorithm tends to give high score to frequent terms, if not using the term weights. However, frequent terms are not always suitable for key terms. For example, in case that the discussion theme is *cell phone* (as of our experiments described in the next section), the discussion participants tend to frequently use terms, such as cell, phone, talk, and call. Those terms would not be significant. Besides, it is only natural that those terms are detected from the discussion.

Key terms must be not too general in every discussion, but particular only to a focused discussion.

We propose an assignment of a term weight based on a *generality* and *particularity*. A *generality* measures overall importance of a term. If a term was frequently used in all discussions, the generality would be higher. A *particularity* measures local importance of a term. If a term was frequently used only in a focused discussion, the particularity would be higher.

Let M be a set of messages posted in all discussions, $M_G(t) \subset M$ be a set of messages in all discussions containing the term t, and $M_L(t) \subset M$ be a set of messages in the focused discussion containing the term t. Denote a logarithmic generality and a particularity of the term t by $w_g(t)$ and $w_p(t)$, respectively. The weight or term t is assigned by

$$\alpha(t) = w_g(t) \cdot w_p(t),$$

where $w_g(t)$ and $w_p(t)$ are given by the following equations.

$$w_g(t) = \log \frac{|M|}{|M_G(t)|}, (0 \leq w_g(t) \leq 1) \tag{3}$$

$$w_p(t) = \frac{|M_P(t)|}{|M_G(t)|}, (0 \leq w_p(t) \leq 1)$$

3 DISCUSSION ANALYSIS BY KEE

In this section, two approaches for knowledge discovery from network-based discussions are proposed. These approaches are based on key terms and key persons obtained by KEE algorithm.

3.1 Discussion Dynamics Analysis

This subsection proposes methods for analyzing discussion dynamics. Suppose that we have a discussion data stored as a sequence of messages. Our goal is to observe how key terms and key persons were changed as the discussion went on.

Key terms/persons transition : Transitions of key terms and key persons are observed with *sliding windows* over the discussion. A *sliding window* is a sequence of a certain number of messages. We observe transition of key terms and key persons obtained in each sliding window.

In order to detect subtle changes of the keys clearly, a *particularity for window*, instead of the particularity proposed in the previous subsection, is used for the term weights. A *particularity for window* is defined so as to increase term weight proportion to particularity of a term in a sliding window. Suppose $M_G(t)$ be a set of messages containing the term t in the focused discussion. Let $M_{PW_i}(t) \subset M_G(t)$ be a set of messages containing the term t in ith sliding window. Denote the particularity for ith sliding window of the term t by $w_{pw_i}(t)$. $w_{pw_i}(t)$ is obtained as follows.

$$w_{pw_i}(t) = \frac{|M_{PW_i}(t)|}{|M_G(t)|}, (0 \leq w_{pg_i}(t) \leq 1)$$

Key terms and persons dynamics over a discussion are observed by the following procedure.

```
Key-s transition:
1. Collect document and clean with typical text
   processing methods including noise filtering and
   term stemming.
2. Assign generality to each term.
3. For each window,
   (a) Assign particularity for the window to each
       term.
   (b) Calculate score by KEE algorithm with $\alpha(t) =$
       $w_g(t) \cdot w_{pw_i}(t)$.
4. Chart the score transitions of each term and participant.
```

Discussion dynamics : How topics were changing over discussions is observed by examining the key-term transitions, which are, more precisely, similarities between a set of key terms in a sliding window and in each of sets of key terms in some previous windows.

Let C_i be a term score vector for ith window obtained by the procedure **Key-s transition**. $C_i = \{c_i(t_1), c_i(t_2), \ldots, c_i(t_n)\}$. Suppose that we examine the similarity between C_i and C_j ($i-k < j < i$, k is a natural number). Each entry of C_j is the score in jth sliding window of each term $t_1, t_2, \ldots, t_n$(extracted in ith window), that is, $C_j = \{c_j(t_1), c_j(t_2), \ldots, c_j(t_n)\}$. Let $Sim(C_i, C_j)$ denote similarity between C_i and C_j. Many similarity measures have been proposed (Lin, 1998; Strehl and Ghosh, 2000). We use one of the most typical similarity measures, a *cosine similarity* (Salton and McGill, 1986) for obtaining $Sim(C_i, C_j)$, which is given by the following equation.

$$Sim(C_i, C_j) = \frac{\sum_{k=1}^{n} c_i(t_k) \cdot c_j(t_k)}{\sqrt{\sum_{k=1}^{n} c_i(t_k)^2 \sum_{k=1}^{n} c_j(t_k)^2}} \tag{4}$$

How topics of the discussion were converged, or conversely diverged, are measured by the differences of each $Sim(C_i, C_j)$, $i-k < j < i$ and their average. Let *diff_sim*(i) and *ave_sim*(i) be a difference and an average of $Sim(C_i, C_j)$, $i-k < j < i$, respectively. They are given by

$$\begin{aligned} diff\_sim(i) &= \max_{i-k<j<i} Sim(C_i, C_j) \\ &\quad - \min_{i-k<j<i} Sim(C_i, C_j) \qquad (5) \\ avg\_sim(i) &= \frac{\sum_{j=i-k}^{i-1} Sim(C_i, C_j)}{k}. \qquad (6) \end{aligned}$$

If *diff_sim*(i) is small and *ave_sim*(i) is high, the discussion may have converged into a certain topic around *i*th sliding window. If *diff_sim*(i) is small but *ave_sim*(i) is low, the topic may have changed into completely different topics. The large *diff_sim*(i) indicates diversity of key terms.

The procedure for observing discussion dynamics is summarized as followings.

```
Discussion Dynamics :
1. Obtain term score vector for each window by the procedure Key-s transition
2. For each term score vector $C_i$
   (a) For each term score vector $C_j$ $(i-k < j < i)$
       i. Obtain cosine similarity $Sim(C_i, C_j)$
   (b) Calculate $diff\_sim(i)$ and $ave\_sim(i)$
3. Chart the transitions of $diff\_sim(i)$ and $ave\_sim(i)$
```

3.2 Social Network Analysis

This subsection proposes a method for generating a map of social network. The social network is represented by a weighted directed graph, based on *post-reply* relationships between participants. The network shows that which pair of participants were *how significant* throughout the discussion. This social network gives us an intuitive grasp of how key terms were transiting over the participants.

A social network is represented by a weighted directed graph $G(V,E)$, where V and E are a set of participants and a set of weighted edges, respectively. Let $w(u,v)$ be a weighted edge from a participant u replying to another participant v. Edge weight $w(u,v)$ is measured by sum of the scores of common terms in the messages from u to v, which is given by

$$w(u,v) = \sum_{k=1}^{n} r_k(u,v),$$

where $r_k(u,v)$ is a *post-reply* relationship from a message m_u by a participants u replying to a message m_v by another participants v, and

$$r_k(u,v) = \sum_{t \in T} c(t) \cdot s, \qquad (7)$$

where T is a set of common terms in the messages m_u and m_v, and $c(t)$ is a term score given by KEE algorithm. s is a tuning parameter used for presenting $w(u,v)$ with a larger value. Since $c(t)$ is less than 1, $w(u,v)$ tends to be quite small.

4 EXPERIMENTS

This section reports experimental results of our approaches applying to actual discussion data. First, we show the key terms extracted by our method and make a comparison with terms by *TFIDF* (Salton and Buckley, 1987). Next, as a discussion dynamics analysis, we report key-term transitions, and discussion dynamics with examining differences with *TFIDF*. Then, for understanding relationships between discussion participants, the transitions of key persons and the extracted social network are reported.

The data was collected from a series of focus groups held on March 2005 together with Hakuhodo Inc. (the second largest advertising company in Japan). The goal of the workshop was to identify future scenarios for cell phone usages and the features that will make them popular among consumers. A several discussions were held during each focus group. The discussion data consists of a sequence of messages. A message consists of message id, title, author name, replying id, and message content.

In the experiments reported in the below, only words (not phrases) were used as terms, and each words is stemmed using the Porter algorithm (Porter, 1997). The detail description of Porter algorithm is beyond the scope of this paper. The proposed approach and the Porter's algorithm were implemented by Perl. Multi feature terms are left out for further research.

4.1 Key Terms Extraction

This subsection reports key terms extracted from the seven discussions by different groups and gives a comparison with the terms obtained by *TFIDF*, one of the typical and traditional methods for finding key terms. In order to have fair comparison, we used *IDF* given by the equation (3) in Section 3.1.

Table 1 and 2 show highest ranked ten terms of each discussion by our method and *TFIDF*, respectively. Our method extracts terms which are even less frequent but given by key persons, in addition to high frequent terms given by key persons. For example, calendar, AOL, Nextel, keyboard from Dis. 1, dvd, palm from Dis.3, and blackberry from Dis. 4 are terms that cannot be detected by *TFIDF*, but they are

Table 1: KEE algorithm induced key terms for each discussion.

Rank.	Dis. 1	Dis. 2	Dis. 3	Dis. 4	Dis. 5	Dis. 6	Dis. 7
1	definitely	toy	dvd	dissatisfaction	roam	claim	3g
2	gps	plane	palm	louis	private	radiation	bluetooth
3	calendar	nap	card	unhappy	telemarketing	ofcours	unit
4	qwerty	compose	pilot	key	unfortunately	emit	implement
5	aol	longrun	identification	blackberry	sign	harm	fix
6	nextel	clearly	water	stand	cellular	microwave	picture
7	simply	society	thousand	lock	junk	unable	visual
8	keyboard	surf	dedicated	package	battery	study	update
9	interact	earlier	fraud	usage	old	scientifically	landline
10	frequent	arent	steal	design	advertisement	even	ultimately

Table 2: *TFIDF.* induced key terms for each discussion

Rank.	Dis. 1	Dis. 2	Dis. 3	Dis. 4	Dis. 5	Dis. 6	Dis. 7
1	cellphone	people	cell	people	battery	cell	cell
2	cell	line	computer	phone	cell	people	phone
3	device	work	card	tool	longer	phone	bluetooth
4	battery	society	pay	internet	phone	talk	import
5	gps	cell	phone	sometimes	roam	feel	picture
6	people	reach	video	cell	people	radiation	camera
7	computer	future	connect	cellphone	cellular	right	message
8	phone	dont	camera	pay	old	even	software
9	internet	land	credit	user	scenario	annoy	communication
10	number	comps	cellphone	unused	sign	really	3g

all worth to examine as significant terms by key persons. Of course, *TFIDF* has high potentiality to detect the discussion topic, although it tends to extract high frequent but not significant terms, such as cell, phone, people, etc. Therefore, examining key terms by the KEE algorithm is essential to grasp what terms are worth to focus as possible clue of innovative and creative ideas.

4.2 Discussion Dynamics Analysis

This subsection reports the discussion dynamics for one of the test discussions. The analysis presents key-term transition, transition of key term similarity differences, and transition of key term similarity averages. We used the sliding window size set to ten, and $k = 5$ for similarity comparisons (see Equation (4)).

The stacked chart in Figure 1 shows how key terms were changing over the discussion. X and Y axes represent *i*-th sliding window, and the stacked scores of highest ranked five key terms. Each area indicates a key-term. Figure 2 and 3 show similarity differences and averages for key terms extracted by our method and *TFIDF*. X axes represent *i*-th sliding window. Y axes represent *diff_sim* and *ave_sim* for each sliding window given by the equations (5) and (6).

In the stacked chart, the parts that each term area is stacked in parallel (roughly, windows between 9th-16th, 23rd-32nd, 49th-55th, 67th-74th, and 75th-78th) are identical to high *avg_sim* appeared in the chart in Figure 3. Moreover, $diff\_sim$ of those windows are quite low. These are clear indication that topics were converged in those windows. Low $diff\_sim$ and *avg_sim* were observed around 64th window, which indicates a topic segmentation. Table 3 shows highest ranked five key terms extracted each listed window. Observing the discussion dynamics, and key-term transition, we could see that the topics shifted from general to specific as the discussion went on. Moreover, key terms of 64th are quite general compared with others, which indicates that the topic was segmented at this point.

Table 3: Key terms of the selected windows.

Wd.	Key terms
16th	talk, call, vibration, make, answer
32th	function, instant, Bluetooth, internet, phonebook
55th	computer, keyboard, screen, small, simply
64th	feature, technology, depend, problem, computer
74th	gps, navigation, service, battery, device
78th	battery, player, music, mp3, problem

As seen in the Figure 2 and 3, compared with the line by KEE, the line chart by *TFIDF* is quite flat. The differences of similarities for the key terms by *TFIDF* were within the range of 0 to 0.5 and the averages of

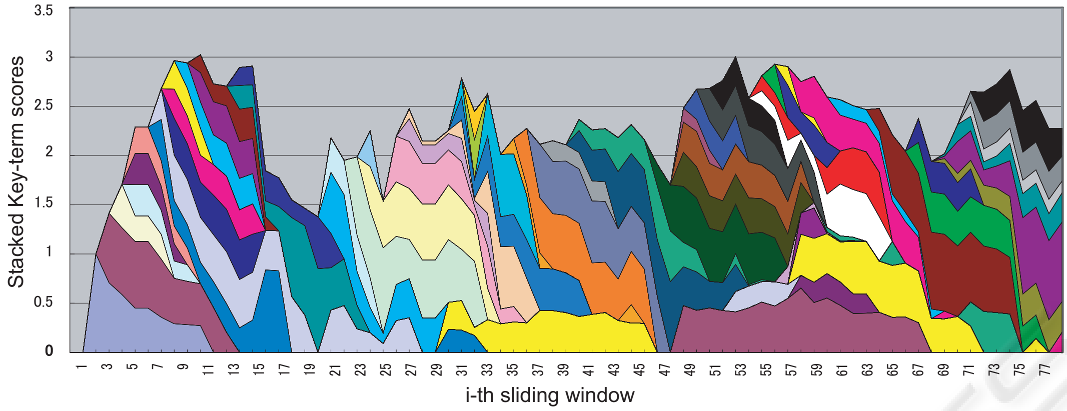

Figure 1: Discussion dynamics by key-term tansition.

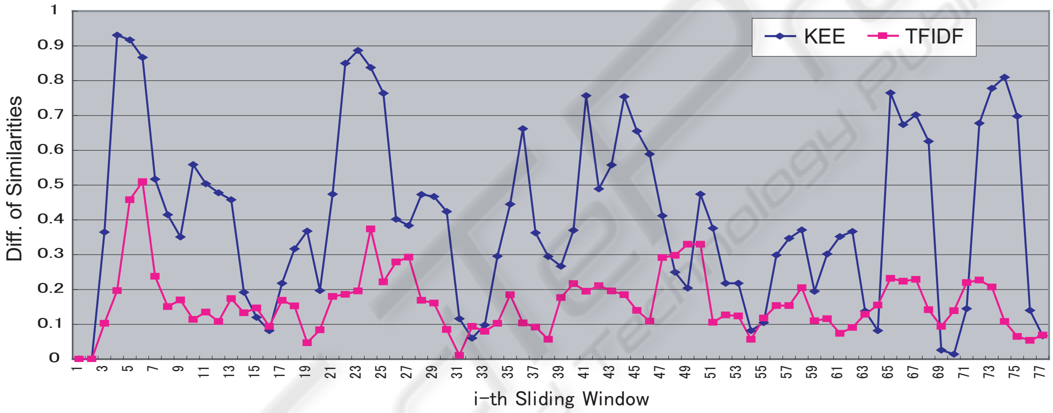

Figure 2: Discussion dynamics by key termsimilarity differences.

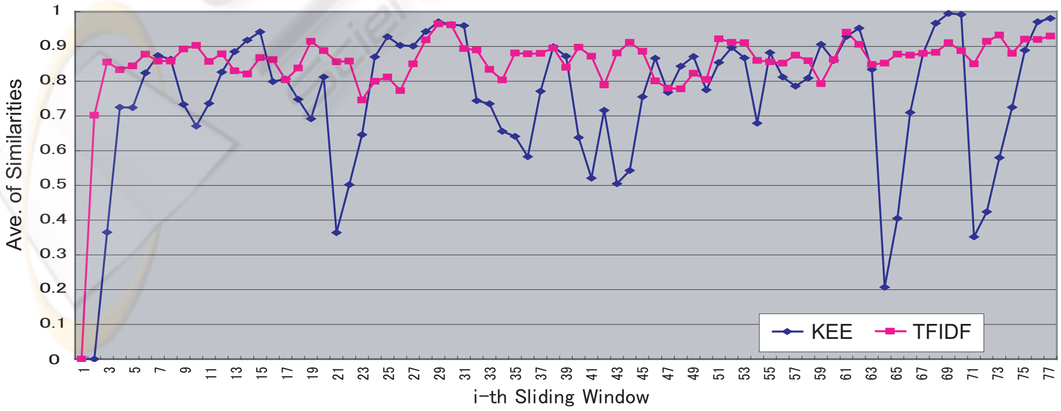

Figure 3: Discussion dynamics by key termsimilarity averages.

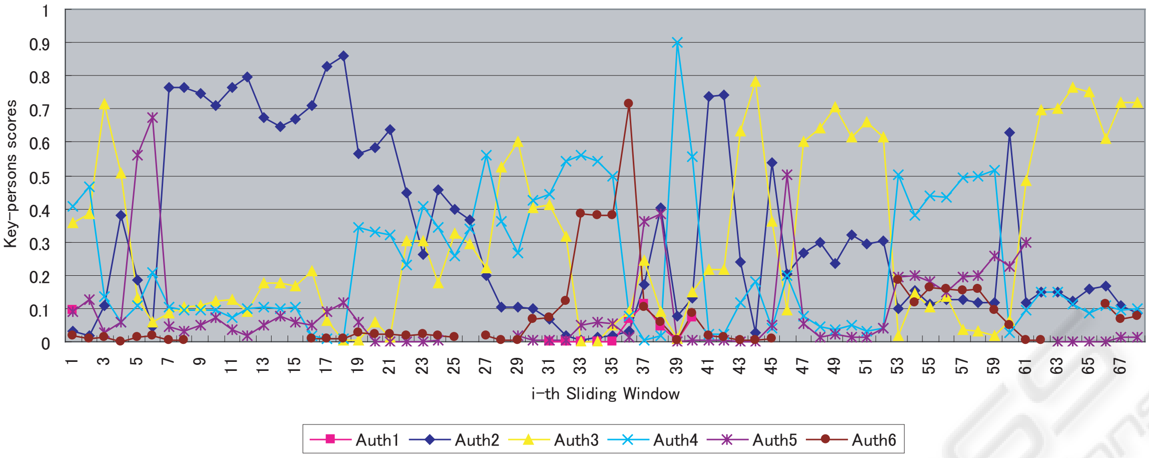

Figure 4: Key persons transition.

similarities are constantly high. It is very difficult to identify periods which are worth to examine.

4.3 Social Network Extraction

This subsection reports an observation for one of the experiment discussions from the view of participants. Seeing both key persons transition and social network will promote better understanding the relationship between participants and their significance.

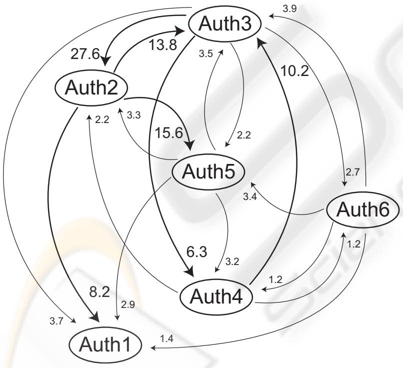

Figure 5: Social Network.

Figure 4 depicts key persons transition. X, Y axes represent *i*-th sliding window and key persons scores, respectively. Sliding window was set to be 10. Figure 5 shows a social network extracted by our method. Each number beside each edge represents the edge weight. We used a tuning parameter $s = 100$ (see the equation (7).)

As seen in Figure 4, Author2 and Author3 dominated the first and the last part of the discussion, respectively. Seen in Figure 5, the relationship between Author2 and Author3 is the most significant. However, each role seemed slightly different. While Author2 had a large connection only from Author3, Author2 had large connections to many authors, such as Author5, Author3, and Author1. This indicates Author2 would play a role as a replier. Conversely, Author3 had large connections both from and to other participants. This indicates that Author3 would play a role as an opinion generator.

5 CONCLUSIONS

This paper focused on delineating topic and discussant transitions in online collaborative environments, more precisely, focus group discussions for product conceptualization. We proposed *KEE* algorithm. Based on *KEE* algorithm, we proposed two approaches for analyzing discussions: discussion dynamics and social network. Our experimental results using real discussion data showed that key terms obtained by *KEE* algorithm gave us better understanding of participantsfidea than the terms obtained by a traditional method *TFIDF*. Moreover, since the key terms were from the key persons in the discussion, those key terms would be potential knowledge, that we had looked for. Both discussion dynamics analysis and social network analysis also gave us significant knowledge which is essential to decision support. These experimental results show the effectiveness of KEE algorithm for network- and text-based communication analysis.

As future work, we plan to use *KEE* algorithm for knowledge discovery in web-logs or web forums. *KEE* algorithm effectively works not only on the relationship between terms and people, but also for any relationship between terms and possible *conceptual packets* of terms, such as, sentences, messages, etc. Similar idea to *key terms/persons* extraction can be applied to these relationship; key terms are included in many key sentences (or messages), and key sentences (or messages) contain many key terms. We would like to apply our approaches to various data source, and lead to innovation and creativity support.

6 RELATED WORKS

The DISCUS project targets on innovation support through network-based communication (Goldberg et al., 2003). In addition to *KEE* methods, two chance discovery approaches: KeyGraph (Ohsawa and Yachida, 1998) and influence diffusion models (IDM) (Matsumura et al., 2002) are used in the DISCUS. Various methods have been proposed for finding significant terms from text (key phrases (Witten et al., 1999), topic words (Lawrie et al., 2001)). Many approaches have been proposed for analyzing text stream by topic detection, tracking, and segmentation (Allan et al., 1998; Beeferman et al., 1999). Some works have focused on finding persons in text-based communication (Kamimaeda et al., 2005; Reich et al., 2002). However, there had been no method for finding significant terms and persons simultaneously.

ACKNOWLEDGEMENTS

We would like to thank to Hakuhodo Inc. for their project collaboration. This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF (AF9550-06-1-0096 and AF9550-06-1-0370). The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

REFERENCES

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study: Final report.

Beeferman, D., Berger, A., and Lafferty, J. D. (1999). Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

Goldberg, D. E., Welge, M., and Llorà, X. (2003). DISCUS: Distributed Innovation and Scalable Collaboration In Uncertain Settings. IlliGAL Report No. 2003017, University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Urbana, IL.

Kamimaeda, N., Izumi, N., and Hasida, K. (2005). Discovery of key persons in knowledge creation based on semantic authoring. In *KMAP 2005*.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Lawrie, D., Croft, W. B., and Rosenberg, A. (2001). Finding topic words for hierarchical summarization. In *SIGIR '01: the 24th ACM SIGIR conference on Research and development in information retrieval*, pages 349–357.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

Llorà, X., Goldberg, D., Ohsawa, Y., Matsumura, N., Washida, Y., Tamura, H., Masataka, Y., Welge, M., Auvil, L., Searsmith, D., Ohnishi, K., and Chao, C.-J. (2006). Innovation and creativity support via chance discovery, genetic algorithms, and data mining. *New Mathematics and Natural Computation*, 2(1):85–100.

Matsumura, N., Ohsawa, Y., and Ishizuka, M. (2002). Influence diffusion model in text-based communication. In *WWW '02: Special interest tracks and posters of the 11th international conference on World Wide Web*.

Ohsawa, Y.and Benson, N. E. and Yachida, M. (1998). KeyGraph: Automatic indexing by co-occurencd graph based on building construction metaphor. In *Proceedings of Advances in Digital Libraries*, pages 12–18.

Porter, M. F. (1997). An algorithm for suffix stripping. pages 313–316.

Reich, J. R., Brockhausen, P., Lau, T., and Reimer, U. (2002). Ontology-based skills management: Goals, opportunities and challenges. *Universal Computer Science*, 8(5):506–515.

Salton, G. and Buckley, C. (1987). Term weighting approaches in automatic text retrieval. Technical report.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc. New York, NY, USA.

Strehl, A. and Ghosh, J. (2000). Value-based customer grouping from large retail data-sets. In *Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology II, 24-25 April 2000, Orlando, Florida, USA*, volume 4057, pages 33–42. SPIE.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G. (1999). Kea: practical automatic keyphrase extraction. In *DL '99: the fourth ACM conference on Digital libraries*, pages 254–255.