Message Feature Map toward Effective Facilitation on On-line Discussions

Noriko Imafuji Yasui
University of Illinois at Urbana-Champaign
104 S.Mathews Ave., Urbana, IL 61801, USA
niyasui@illigal.ge.uiuc.edu

Shunsuke Saruwatari
University of Illinois at Urbana-Champaign
104 S.Mathews Ave., Urbana, IL 61801, USA
saru@illigal.ge.uiuc.edu

Xavier Llorà
University of Illinois at Urbana-Champaign
1205 W.Clark St., Urbana, IL 61801, USA
xllora@illigal.ge.uiuc.edu

David E. Goldberg
University of Illinois at Urbana-Champaign,
104 S.Mathews Ave., Urbana, IL 61801, USA
deg@uiuc.edu

Abstract

In these years, with the expansion of network-based communication tools, consumer based marketing strategies using on-line discussions have received much attention. To get the benefit from on-line discussions, methodologies for analyzing on various discussion elements are demanded. In this paper, we propose a message feature map, which is a visualization method for plotting discussion messages on a plane with two axes; centralities and expansiveness. This map gives us an intuitive understanding and quick grasp of discussion status. The experimental results with using the data collected in a real on-line focus group discussion shows that the message feature map helps us to observe the discussion status effectively.

1. Introduction

In the last decades, network-based communication has become more and more important for people to obtain information, exchange opinions and propagate new ideas or concepts more easily, more quickly and more effectively. People pay attention to the potential for on-line discussions with making the best use of the network-based communication. Since diverse people can easily join the on-line discussions, it becomes possible to build consumer-based marketing strategies. However, people at companies still have face-to-face discussions for brainstorming of new ideas on new products or services, and conceptualizations of new products. This is due to the complications in facilitation on on-line discussions.

Key-element of the success on off-line discussion is how skillfully the facilitators control the discussions. Typically, the facilitators always grasp discussion status accurately, and ask or talk to proper participants about proper topics in the best timing. One of the biggest advantages of on-line discussions is that we can have many discussions in parallel with various groups of diverse people. It is difficult to handle all the discussions by a limited number of facilitators. In order to reduce facilitators' task and pursue their role more effectively, methodologies for analyzing discussion status and visualizing them are required.

Suppose we have on-line discussions with a certain number of participants by posting and replying messages on on-line discussion boards. This paper proposes a message feature map, which is a visualization method for intuitive understanding of "*who said what, when and how*". *Who* means discussion participants, *what* means message contents, *when* means timing on discussion dynamics status, and *how* means how messages contribute to discussion going. We also introduce two metrics to measure *centrality* and *expansiveness* of each message. *Centrality* indicates how much each message was center (or conversely, peripheral) in the discussion. *Expansiveness* tells us how much each message contributed to expansion (or conversely, contraction) of the discussion. The message feature map plots each message on the plane of the axes with the two metrics.

The reminder of this paper is organized as follows. Section 2 reviews the KEE algorithm, which is an algorithm for finding key persons and key terms synchronously from discussions, and is a basic for identifying message centralities. In Section 3, we propose a message feature map with two metrics; centrality and expansiveness. Section 4 reports the experimental results with using real data collected in a real on-line focus group discussion. Finally, this paper concludes in Section 5 with summarizing and directions for future work.

2. KEE Algorithm

In [2], we proposed an algorithm called KEE (*Key Elements Extraction*). KEE is an algorithm for synchronously 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. This section reviews KEE algorithm.

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 [4] 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 denote an edge between $p_i \in V_P$ and $t_j \in V_T$ and its weight by (p_i, t_j) and $w(p_i, t_j)$, 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 [4], 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 synchronously 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 not be argued in this paper. Please refer to [2] in detail.

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

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

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 $A^T A$ and AA^T [4]. 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).

3. Message Feature Map

Visualization is a key element in successful facilitation on on-line discussions. Good visualization is simple, but makes it possible to observe the discussions in various points of views. This section proposes a visualization called a *message feature map* with two metrics; *centrality* and *expansiveness*.

Centrality: This metric measures how much messages are center of the discussion topic. For identifying centralities of messages, we use the mutual reinforcement relationship between messages and terms (instead of participants and terms) for KEE algorithm: a message is center of the discussion, if the message contains a lot of terms with high centralities, and a term is center of the discussion, if the term is contained in many messages with high centralities. If the centrality of the message is high, the message would be center of the discussion. Conversely, if the centrality is low, the message would discuss about peripheral topic.

Suppose a discussion consists of a sequence of messages $(m_1, m_2, \cdots, m_n)$. By using KEE algorithm, we calculate the score for each message k times with different sets of messages. Let M be a message score vector obtained by KEE algorithm, and $M(m_i)$ be the score for the message m_i. Denote the centrality of the message m_i by $c(m_i)$, $c(m_i)$ is defined as follows.

$$c(m_i) = \sum_{i \le j < i+k} M_j(m_i),$$

where $M_j(m_i)$ is a score of m_i with a set of k messages $\{m_{j-k+1}, \cdots, m_{j-1}, m_j\}$.

Expansiveness: This metric measures how much messages contribute to the discussion activeness. We assume that discussion is expanding when participants give a lot of new ideas and thoughts. Thus, expansiveness is based on number of the new terms in each message. We define a *new term* in a message as a term which is not appeared in some previous messages. If the expansiveness of the message is high, the message would lead the discussion to be more active. Conversely, if the expansiveness is low, the message would lead the discussion to be inactive.

Suppose a discussion consists of a sequence of messages $(m_1, m_2, \cdots, m_n)$. Denote the expansiveness of the message m_i by $e(m_i)$, $e(m_i)$ is defined as follows.

$$e(m_i) = N^l(m_i),$$

where $N^l(m_i)$ is a number terms which are not existed in a set of the messages $\{m_{i-l}, \cdots, m_{i-1}\}$.

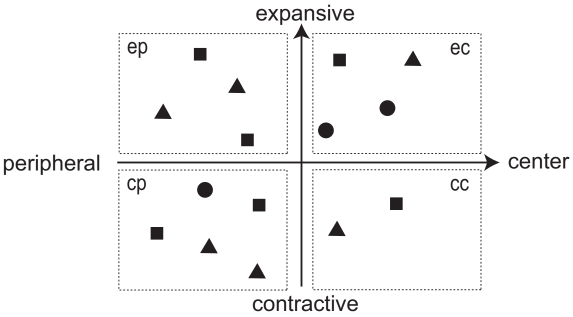

Figure 1. A sample of message feature map

Figure 1 shows a sample of a message feature map. Each message is plotted with different colors (or shapes) for each participant on a plane with two axes of centrality and expansiveness. The centrality and the expansiveness are normalized so that the values are in the range of -1 to 1. Let C and E be vectors which store centralities and expansiveness of massages, respectively. The normalized centrality $c'(m_i)$ and the normalized expansiveness $e'(m_i)$ are obtained as follows. Note that $maxC$ and $maxE$ indicate the highest values in C and E. Conversely, min means the lowest.

$$c'(m_i) = \frac{2 \cdot c(m_i)}{maxC - minC} - 1$$

$$e'(m_i) = \frac{2 \cdot e(m_i)}{maxE - minE} - 1$$

As seen in the Figure 1, the message feature map has four areas; ep (upper-left), ec (upper-right), cc (lower-right), and cp (lower-left). Followings are message types plotted in the four areas.

- Type ep: this type of messages is the message which brings new idea or opinion to the discussion, but any other participant didn't follow the idea nor talk about it any further.

- Type ec: this type of message is the message which brings new idea or thought to the discussion. Originating from this message, the discussion topic is shifted to the one in this message. The message is a topic starter, in a sense.

- Type cc: this type of message is the message which gives more idea or deeper insight on the current topic. The discussion topic is somehow converged by this message.

- Type cp: this type of message is the message which does not influence on the discussion going. The message content doesn't have any specific topic. For example, these are short messages such as yes/no answers, short comments and so on.

4. Experiments

This section reports preliminary experimental results of our approaches applying to actual discussion data. The discussion data was collected from a series of focus groups discussions held on March 2005 together with Hakuhodo Inc. (the second largest advertising company in Japan). We used the data obtained from one of the discussions. In this discussion, six participants (including a facilitator) posted and replied 76 messages in total. The data consists of a sequence of messages. A message consists of message id, title, author name, replying id, and message content. Please refer to [2] for the details of the data, and various experimental results including content analysis, time series analysis, and social network analysis.

Figure 2 shows the message feature map using the discussion data. The horizontal axis represents the centrality and the vertical axis represents the expansiveness. In this experiment, we used $k = 10$ for the centrality, and $l = 10$ for the expansiveness. Each message is colored differently for each participant. The messages plotted in the upper-left area were containing many variable minority opinions and ideas. For example, "*one of the biggest things I see for cell phones ... is money withdrawal. By linking your cell phone and debit card, you could buy stuff using cell phone and PIN number*". Typically, this type of message which has no salient features based on term frequency is difficult to detect without reading thorough all the messages. The messages plotted in the upper-right area indicated appearance of new topics. The four messages out of eight were by a participant P001. We could have an assumption that P001 was good at generating new topics. The messages plotted

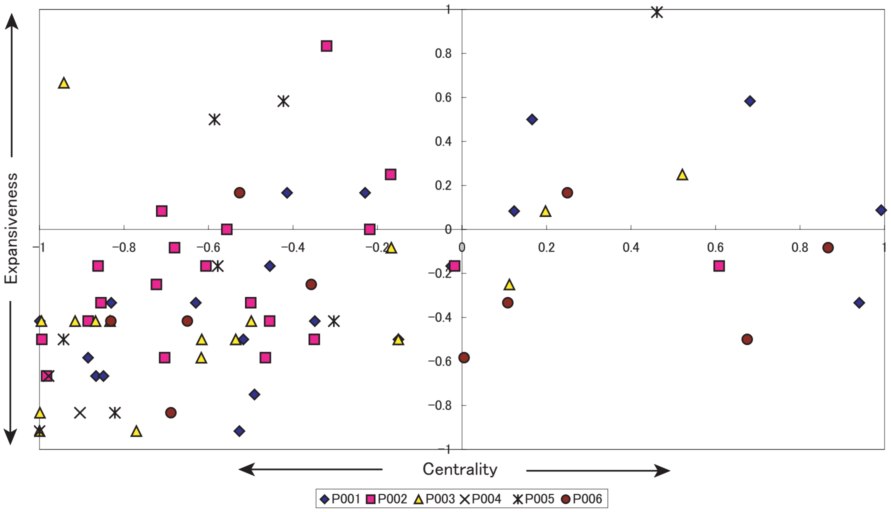

Figure 2. Message feature map

in the lower-right area were containing additional information, comments on an existing topic, or summarizing, in a sense. The participant P006 posted some messages of this type. Most of the messages plotted in the lower-left area were short replies, or quick answers, for example ”*I like that idea*”.

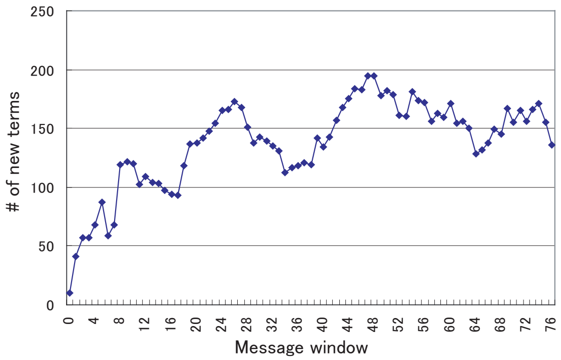

Figure 3. Discussion activeness

We also examined the relationship between discussion activeness and the message feature map. The line chart in Figure 3 shows how active the discussion went on. We assume that the more new terms used in the messages, the more active participants are discussing, and then define the discussion activeness as follows. The ith message window is defined as a set of k messages, and denoted by M_i, that is, $M_i = \{m_{i-k+1}, \cdots, m_i\}$, where m_i be a ith message. The activeness for M_i is defined a sum of number of new terms in the messages of M_i. The x axis in the chart represents message windows. In this experiment, a message window contains ten messages, that is, $k = 10$. The y axis shows the activeness.

As seen in the chart in Figure 3, the discussion went on with repeating active and inactive phases. We examined the differences between active and inactive phases on the message feature maps. Figure 4 shows the three message feature maps; from 18th to 27th messages (the left), from 28th to 37th message (the center) and from 38th to 47th (the right). Observed in the discussion activeness chart, we could assume that the discussion was quite active from 18th to 27th messages and from 38th to 47th, but not active from 28th to 37th message. One of the most significant differences in these maps is that in the inactive phase (in the middle map), most of the messages are plotted in the lower-left area.

5. Conclusions

In this paper, we proposed a message feature map, which was a visualization method for plotting discussion messages on a plane with two axes. We also introduced two metrics; *centrality* and *expansiveness*. *Centrality* indicates how much each message was center (or conversely, peripheral) in the discussion. *Expansiveness* tells us how much each message contributed to expansion (or conversely, contraction) of the discussion. This map gave us an intuitive understanding and quick grasp of discussion status. We showed the experimental results with using the data collected in a

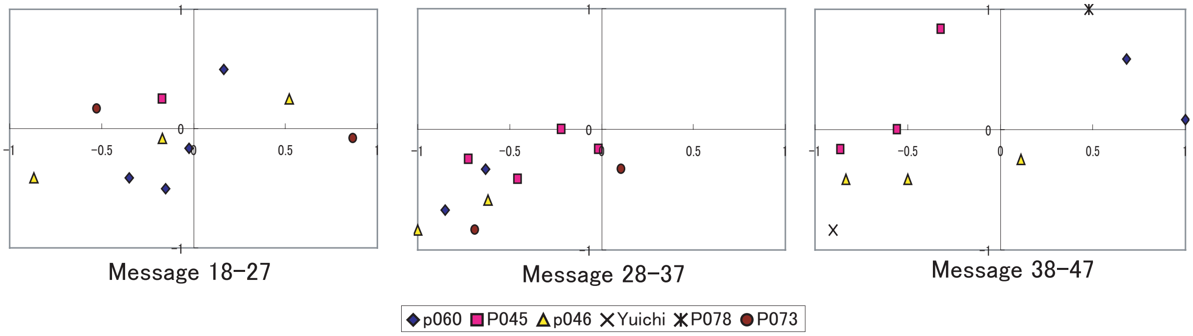

Figure 4. Message feature maps for selected messages.

real on-line focus group discussion. A lot of scenarios for facilitation usages can be assumed from the message feature maps. For example,

- Ask questions or talk to somebody when a lot of messages are plotted in lower-left area, because the discussion is going inactive.
- Pay attention to the participants whose messages are plotted in upper area, because they have potential for giving new ideas.
- In order to deepen discussion on a certain topic, use the participants whose messages are plotted in lower-right area, because they tend to give idea or thoughts on current topics.

Our future works include to build a message transition model which tells how message status transit on the message feature map, and to simulate discussion for a given set of participants. Discussion simulation will be a very useful tool for discussion planning - determining the discussion goal, grouping the people, building strategic facilitation scenarios.

6 Related Works

The DISCUS project targets on innovation support through network-based communication [1]. In addition to *KEE* methods, two chance discovery approaches: KeyGraph [7] and influence diffusion models (IDM) [6] are used in the DISCUS. Various methods have been proposed for finding significant terms from text (key phrases [8], topic words [5]). Some works have focused on finding persons in text-based communication [3]. These are effective methods for analyzing discussion in one aspect. The analysis results might help the discussion facilitators, but to handle multiple discussions easily and effectively, a simple visualization of analysis data with various aspects is required.

Acknowledgments

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).

References

[1] D. E. Goldberg, M. Welge, and X. Llorà. 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, 2003.

[2] N. Imafuji Y., X. Llorà, D. E. Goldberg, W. Y., and T. D. Delineating topic and discussant transitions in online collaborative environments. In *Proceedings of 9th International Conference on Enterprise Information Systems (ICEIS 2007)*, 2007.

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

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

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

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

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

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