

# Mining Messages in an Electronic Message Board by Repetition of Words

Yutaka Matsuo<sup>21</sup>, Yukio Ohsawa<sup>23</sup>, and Mitsuru Ishizuka<sup>1</sup>

<sup>1</sup> University of Tokyo, Hongo 7-3-1, Tokyo 113-8656, Japan

matsuo@miv.t.u-tokyo.ac.jp

<http://www.miv.t.u-tokyo.ac.jp/matsuo>

<sup>2</sup> Japan Science and Technology Corporation,

Tsutsujigaoka 2-2-11, Miyagino-ku, Sendai 983-0852, Japan

<sup>3</sup> University of Tsukuba, Otsuka 3-29-1, Bunkyo-ku, Tokyo 112-0001, Japan

**Abstract.** In this paper, we develop an algorithm to mine message boards, e.g., Yahoo! Clubs. Each message is evaluated by three indices: (i) the degree of inclusion of given information, (ii) the degree of inclusion of new information, and (iii) the degree of effect for the successive messages. We show these indices are useful for characterizing a message. We also show a prototype system of summarization applied to a message board, as an example of application of the indices.

## 1 Introduction

Recently, Web communities are receiving much attentions in the business world. Especially, message boards (or bulletin board systems) such as Yahoo! Clubs are gold mines of information for business people to make new strategic plans, because a large amount of textual data of chatting by business customers can be obtained freely.

Let us take Japanese Yahoo! BBS site (<http://messages.yahoo.co.jp/>) for example. In January 2001, there are more than 7000 message board categories, such as pregnancy (in Women, Health & Wellness directory), an Italian football league Seria A (in Soccer, Sports, Recreation & Sports directory), employment of the disabled (in Employment, Business & Finance directory), and so on. For each category, there are up to one hundred topics, such as “Let’s try fertility treatment”(3947 messages), “The analysis of Hidetoshi Nakata’s performance<sup>4</sup>”(12644 messages), and “I will make you shine!”(2284 messages). In Stock category in Business & Finance directory, there are more than 3700 message boards, each for a listed company from hospitality industry to manufacturing industry to financial industry. For example, a message board for one of chain restaurants includes opinions about taste of dishes, the company’s measure against BSE<sup>5</sup>, and the stock price. Such messages are very interesting sources of information. There is already an attempt to predict stock prices from a buzz on message boards [1].

However, due to a large number of messages, it is often very difficult to read through the messages. Furthermore, these messages are a different type of textual data than

<sup>4</sup> Hidetoshi Nakata is a famous Japanese football player who is currently playing in Italy.

<sup>5</sup> Bovine spongiform encephalopathy (or mad cow disease), which became an object of public concern in Japan in 2001.

common documents. Especially, a message is posted to reply the previously posted message. Thus a message board usually forms a reply-replied structure.

In this paper, we develop a new approach to characterize each message using the reply-replied structure. It is based on Halliday's given-new dichotomy, where given/new information is presented by the speaker as recoverable/not recoverable information to the listener. We take a very simple criteria to judge whether information is given or new; If a word in a message already appeared in the message replied, the word is considered given, otherwise new. By considering whether a word is given or new, we can assign three indices to each message; (i) the degree of inclusion of given information, (ii) the degree of inclusion of new information, and (iii) the degree of effect for the successive messages. We show these indices are useful for characterizing a message. Our approach is simple, and can be generally applied to messages in a message board.

The rest of the paper is organized as follows: In the following section, we first describe given and new information, and three indices to feature a message. Then, we show the relation between the indices and the classification of messages. A prototype of a summarization system for a message board is developed in Section 4. Section 5 is devoted to discussions and related works, and finally we conclude this paper.

## 2 Indices of Message Information Structure

### 2.1 Given and New information

Michael Halliday is one of the most famous linguists in the world, and his systemic functional model of grammar is recognized as one of the most powerful explanatory models of language. According to Halliday[2], the information unit in discourse is a structure made up two functions, the given and new.

**Given** information which the speaker assumes that the addressee can derive from a previous part of the text or the physical setting.

**New** information which the speaker presents in such a way that it is not derivable from the previous co-text or the physical setting.

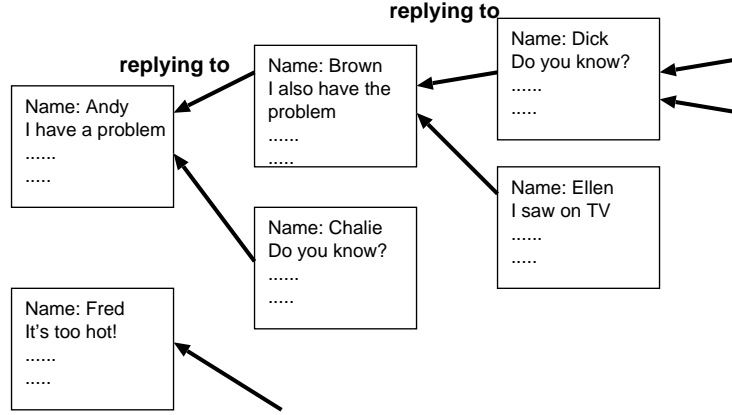
He says as follows:

Information ... is a process of interaction between what is already known or predictable and what is new or unpredictable. This is different from the mathematical concept of information, which is the measure of unpredictability. It is the interplay of new and not new that generates information in the linguistic sense. Hence the information unit is a structure made up of two functions, the new and the given.

A part of given information appears in many forms, such as reference, substitution, ellipsis, lexical cohesion, and repetition. For example, in the following dialogue,

“Yesterday, I attended a linguistic conference at University of Tokyo.” “Well, I have not been to University of Tokyo. Where is it?”

“University of Tokyo” is repeated and thus is a part of given information.



**Fig. 1.** A structure of messages.

## 2.2 Three characteristic indices

In a message board, a message is posted replying to a message previously posted. (A message can also be posted independently.) As a whole, a message board consists a reply-replied structure as shown in Fig. 1. In this sense, a message is considered to be an utterance to the message replied.

In order to apply the concept of given and new information, we regard each message as an information unit. A part of the message conveys given information and the rest conveys new information. In the simplest way, we can classify words in a message into two groups, given and new. Admitting that given information may appear in many forms, we employ the simplest criteria in order to enable automatic computation; *If a word in a message already appeared in the message replied, the word is considered given, otherwise new.*

Counting the number (or adding weights) of the words in given/new groups, we can calculate the degree of given/new information that a message conveys. The degree of given/new information is defined as follows.

**Definition 1.** *The degree of given information  $G$  in message  $d$  is defined as follows.*

$$G(d) = \sum_{w \in W(d) \cap W(R(d))} weight(w, d),$$

where  $W(d)$  is a set of words in message  $d$ ,  $R(d)$  is a message that message  $d$  replies to<sup>6</sup>, and  $weight(w, d)$  is a weight of word  $w$  in message  $d$ .

**Definition 2.** *The degree of new information  $N$  in message  $d$  is defined as follows.*

$$N(d) = \sum_{w \in W(d) \cap \overline{W(R(d))}} weight(w, d),$$

<sup>6</sup> In the experiment below, we use the expanded definition of  $R(d)$ ;  $R(d)$  is the message that  $d$  replies to (denoted as  $d_1$ ), and the message that  $d_1$  replies to.

where  $\bar{A}$  is the complement of a set  $A$ .

Furthermore, a part of new information is more important than other part, because it is used as given information in the subsequent messages. In other words, a part of information has large effect on the subsequent messages. We define the degree of the effect of new information as follows.

**Definition 3.** *The degree of effect  $E$  in message  $d$  is defined as follows.*

$$E(d) = \sum_{w \in W(d) \cap \overline{W(R(d))}} \sum_{\{d' | R(d')=d\}} weight(w, d'),$$

Though  $N$  and  $G$  depend on whether a word in the message appears previously or not,  $E$  depends on whether the word appears afterwards or not.

These definitions are easily expanded to measure the degree of given, new information and effect for each sentence, instead of each message. For example, the degree of given information in sentence  $s$  in document  $d$  is defined as

$$G(d, s) = \sum_{w \in W(s) \cap W(R(d))} weight(w, d),$$

where  $W(s)$  is a set of words included in sentence  $s$ .

### 3 Featuring Messages

#### 3.1 Classification and its Feature

In this section, we try to characterize a message by the indices. We take as an example a message board talking about one of chain restaurants in Japan in Yahoo!BBS Japanese site<sup>7</sup>. The message board consists of 262 messages, and we manually classify all the messages into the following 17 categories.

##### Question

**Q1: Question** E.g. Is there any new menu in the restaurant? Please tell me ...

**Q2: Question** E.g. What do you think the new menu? What is your opinion?

##### Answer

**A1: Answer** E.g. It is ...

**A2: Agreement** E.g. Yes. You are right.

**A3: Disagreement** E.g. I don't think so.

**A4: Thanks** E.g. Thank you for your answer.

##### Opinion

**O1: Positive evaluation** E.g. The new menu is good.

**O2: Negative evaluation** E.g. I don't like the menu.

**O3: Intention, action** E.g. I will buy the stock.

**O4: Wish** E.g. I hope the company will do this.

---

<sup>7</sup> Because Japanese is a agglutinating language, we use a morphological analyzer to separate a sentence into words.

## Information

**I1: General information** E.g. I saw this on TV.

**I2: Experience** E.g. It was very crowded.

**I3: Article or URL** Pasted articles, or URLs.

## Exceptions

**E1: Greeting, E2: Correction, E3: Abuse, E4: Deleted, E5: Others**

Each message is classified into one of these categories, but is permitted to be classified into two categories if necessary, e.g. I2 and O2. These annotations are made by two persons, and if two annotations are different, the consensus is made through discussion.

The result of classification and the averages of  $G$ ,  $N$  and  $E$  are shown in Table 1. We set  $weight(w, d) = tf(w, d)$ , where  $tf(w, d)$  is a frequency of word  $w$  in message  $d$ . On average 2.82 words which are used in the message replied are used, and 24.77 words are used newly. Messages belonging to Answer category has as high as 6.28 words recognized as given information, which is 2.23 times larger than the average. New information  $N$  is high in messages in Information category. This matches well with our intuition. Messages in Answer categories and Question 1 category don't convey much new information. Lastly, messages in Question category have high effects. Interestingly, messages in Disagreement category also have high effects. (People might heat up by objection.) Messages in Thanks and Article or URL have extremely low effects.

## 3.2 Weight of Words

Though we don't include function words for counting, such as "and" and "do," words with specificity, such as proper nouns, have more information than general words. Thus, we assign a weight to each word by utilizing the index  $tfidf$ [3], which is commonly used in the context of information retrieval. This measure is defined as follows.

$$tfidf(w, d) = tf(w, d) \times (\log n/df(w) + 1),$$

where  $tf(w, d)$  is the frequency of word  $w$  in message  $d$ ,  $n$  is the number of all the messages, and  $df(w)$  is the number of messages which include  $w$ .

By using  $tfidf(w, d)$  as  $weight(w, d)$ , the indices  $G$ ,  $N$  and  $E$  for each message is now refined. The average of each category of messages are shown in Table 3, where the values are normalized so that the total average is to be 1.0. (We also show Table 2, which is the normalized version of Table 1.) Though this weighing is important in that it takes into considerations the quantity of information, the averages are not changed dramatically as a whole.

## 4 Summarization of a Message Board

We show a prototype system of summarization applied to a message board, as an example of application of given, new information and effect indices. It is very essential to show given information besides new information to make understand discourse easily. And a message with large effect includes sometimes the trigger of discussion.

Firstly, we decide which messages to summarize among a large number of messages. Effect  $E$  is used to determine a message to be included in the summary.

**Table 1.** Classification and  $G$ ,  $N$  and  $E$ .

	Num. of Mes.	Ave. Given $G$	Ave. New $N$	Ave. Effect $E$
Question	32	1.54	23.66	6.19
Question(Tell me ...)	16	1.98	19.90	7.42
Question(How do you think ..)	17	1.03	26.93	5.32
Answer	56	6.28	20.06	2.68
Answer(It is ...)	20	7.86	20.49	1.85
Agreement(Yes.)	20	5.71	19.41	2.67
Negation(No.)	12	4.63	22.63	4.77
Thanks	4	6.17	13.47	0.56
Opinion	126	2.39	24.73	2.35
Positive evaluation	61	3.35	27.10	2.58
Negative evaluation	41	1.64	25.76	2.17
Intention, Action	20	1.89	19.02	1.90
Wish	11	2.44	27.47	2.32
Information	91	2.27	31.88	2.18
General information	19	3.15	32.38	3.03
Experience	50	2.08	27.16	2.39
Article or URL	25	1.71	40.42	0.99
Average	(total 262)	2.82	24.77	2.70

**Table 2.** Classification and  $G$ ,  $N$  and  $E$  (normalized).

	Num. of Mes.	Ave. Given $G$	Ave. New $N$	Ave. Effect $E$
Question	32	0.54	0.96	2.29
Question(Tell me ...)	16	0.70	0.80	2.75
Question(How do you think ..)	17	0.37	1.09	1.97
Answer	56	2.22	0.81	0.99
Answer(It is ...)	20	2.78	0.83	0.69
Agreement(Yes.)	20	2.02	0.78	0.99
Negation(No.)	12	1.64	0.91	1.77
Thanks	4	2.18	0.54	0.21
Opinion	126	0.85	1.00	0.87
Positive evaluation	61	1.19	1.09	0.96
Negative evaluation	41	0.58	1.04	0.81
Intention, Action	20	0.67	0.77	0.70
Wish	11	0.86	1.11	0.86
Information	91	0.80	1.29	0.81
General information	19	1.11	1.31	1.12
Experience	50	0.74	1.10	0.88
Article or URL	25	0.60	1.63	0.37
Average	(total 262)	1.0	1.0	1.0

**Table 3.** Classification and  $G$ ,  $N$  and  $E$  by *tfidf* measure (normalized).

	Num. of Mes.	Ave. Given $G$	Ave. New $N$	Ave. Effect $E$
Question	32	0.55	0.94	2.38
Question(Tell me ...)	16	0.80	0.80	2.99
Question(How do you think ..)	17	0.28	1.07	1.91
Answer	56	2.33	0.82	1.04
Answer(It is ...)	20	3.05	0.82	0.66
Agreement(Yes.)	20	1.93	0.79	0.89
Negation(No.)	12	1.80	0.97	2.20
Thanks	4	2.34	0.56	0.13
Opinion	126	0.80	0.99	0.81
Positive evaluation	61	1.13	1.07	0.85
Negative evaluation	41	0.57	1.07	0.82
Intention, Action	20	0.54	0.76	0.59
Wish	11	0.78	1.08	0.78
Information	91	0.78	1.30	0.81
General information	19	1.21	1.29	1.12
Experience	50	0.68	1.11	0.88
Article or URL	25	0.55	1.69	0.37
Average	(total 262)	1.0	1.0	1.0

- Select a given number of messages with the highest effect  $E$ . These messages are called *root messages*.
- For each root message, pick up the subsequent messages recursively, whereas a message where  $G$  is less than a given threshold<sup>8</sup> are excluded, and a message which already picked up are also excluded.

Then, we make a summarization of a root message and its subsequent messages as follows<sup>9</sup>;

- From the root message, pick up 4 sentences with the highest effect  $E$ .
- From each subsequent message, pick up one sentence with the highest  $G$ , one sentence with the highest  $E$ , and one sentence with the highest  $N$ . If no message replies to the message, (which means  $E = 0$ ,) then pick up one sentence with the highest  $N$  instead of the highest  $E$ .

A sentence with given information helps understand new information. A sentence with high effect is also necessary to show given information of subsequent messages. The most important part of a message is extracted based on the index of new information.

Furthermore, if the value  $G$  of a root message is high enough<sup>10</sup>, we include the message which the root message replies to, and pick up two sentences with high effect.

<sup>8</sup> We use 5.0.

<sup>9</sup> Note that the number of sentences to pick up depends on the number of sentences in a message. We show an ordinary configuration for a message with more than 4 sentences.

<sup>10</sup> We use 10.0 as a threshold.

Each root message and its subsequent messages are entitled by 4 words which have the highest effect in total throughout the messages.

Figure 2 shows a screen shot of our system. This example is a part of a summarization result of a message board about a clothing company in Japan, which contains 1235 messages. The topic here is about the company's new product, AIRTECH jacket and its size. We can easily grab the contents of the discourse although more than half the sentences are eliminated. The values  $G$ ,  $N$  and  $E$  are also shown at the right of each sentence<sup>11</sup>.

Though our approach seems to work very well, further research is needed to evaluate summarization results and to derive some conclusions.

## 5 Discussion and Future work

In this paper, we show the applicability of Halliday's given-new dichotomy to mine a message board. Messages are manually classified, and each category has different characteristics of the indices. We don't show how we can automatically classify a message into these category, because in this experiment, we classify each message into up to two categories. To learn the classification, we have to annotate all the categories which a message should be classified into.

Besides given-new structure, Halliday relates another type of structure; theme-rheme. The theme is the starting point of the communication chosen by the speaker, while the rheme is the remaining part which develops the theme. In English, the theme-rheme structure is conveyed by word order, and in Japanese, it is conveyed also by postpositional particles. The theme-rheme structure and the given-new structure are semantically interconnected.

There is a close semantic relationship between information structure and thematic structure. Other things being equal, a speaker will choose the Theme from within what is given and locate the focus, the climax of the new, somewhere within the Rheme. ([2])

In other words, theme and rheme are determined by the speaker, while given and new information are determined by the hearer.

One question is "Who is a hearer in a message board?" A message is posted to reply to the previous message. In this sense, the hearer is the person who wrote the message. However, a message is often posted with conscious of being observed by many others. In this sense, the hearers are readers of the message board. There is a gap between the communication model of given-new structure and the communication in a message board.

There has been a large amount of research related to summarization [4]. Recently, multidocument summarization (for example [5]) is receiving much attention. In our summarization system, we are targeting at message boards, which is a new attempt as far as we know.

---

<sup>11</sup> Our system can summarize a message board with about 1000 messages in less than 5 minutes except the time to download all the messages. However, this system is for Japanese Yahoo! BBS site, and the screen shot is the translated version of the original one.

## Topic 5 (AIRTECH, jacket, corny, women's)

[Previous topic](#)  
[Next topic](#)

### I'm going to buy a AIRTECH jacket

Name: sakanatun5143

Message 748/1235

[Full text Reply](#)

I also tried AIRTECH jacket, but it was corny...<sup>1</sup> G 3.37 N 7.63 E 15.83

I am a man but 165cm high, so I tried women's, which fits very well, and the outline was very cool, and I like the brown one.<sup>2</sup> G 0.00 N 25.27 E 8.86

Given: 6.83、New: 38.47、Effect: 35.06

Sentence 3, Replies 2

### Let's buy. Only 3 days left!

Name: ytmama

Message 753/1235

[Full text Reply](#)

I bought the black one, and it was so cool that I also want the white one.<sup>5</sup> G 2.60 N 0.76 E 0.00

But I am looking for a beige duffel coat, so I will not buy a second AIRTECH.<sup>6</sup> G 6.73 N 6.78 E 0.00

Is there a duffel that is above-the-knee?<sup>7</sup> G 0.00 N 3.74 E 0.00

Given: 6.07、New: 8.72、Effect: 0.00

Sentences 7, Replies 0

### Exactly!!

Name: haz24

Message 755/1235

[Full text Reply](#)

I am 179cm high, but very thin, and I also felt AIRTECH was corny.<sup>2</sup> G 12.82 N 6.65 E 4.84

So... I bought women's AIRTECH!!<sup>3</sup> G 9.71 N 2.79 E 3.55

It is Large size and just fit.<sup>5</sup> G 2.04 N 4.45 E 0.00

Given: 21.26、New: 12.51、Effect: 7.35

Sentences 7, Replies 1

### I went again.

Name: sakanatun5143

Message 756/1235

[Full text Reply](#)

On my way home, I went to the UNIQLO at the Meiji street.<sup>1</sup> G 0.00 N 7.33 E 0.00

I wanted but didn't buy a AIRTECH jacket, a crew-neck sweater, and a Henly neck fleece, just used the toilet.<sup>3</sup> G 10.62 N 15.24 E 0.09

Given: 20.47、New: 33.12、Effect: 0.19

Sentences 4, Replies 1

Fig. 2. A sample screen shot.

Our research is related also to text mining and Web content mining [6]. For example, Lazarinis proposes the use of information extraction techniques for the domain of calls for papers [7]. Kameyama et al. attempted to extract information from Japanese spoken dialogues and make a summary information [8]. Matsumura et al. analyzes a message boards, and try to find influential words, interesting messages, and opinion leaders [9]. A commercial software developed by Opion is to explore how chat-room banter affect stock prices [1]. It calculates the number and order of citations to determine the importance (or rank) of a person.

## 6 Conclusion

Electronic message boards have an abundance of useful information. In this paper, we have presented an approach to characterize a message in a message board by given, new information and effect. These indices are useful for featuring a message. We have also showed a prototype of a summarization system using these indices.

We show here the possibility of applying the linguistic theory to mine message boards, by focusing only on the repetition of words. We will next take considerations into other types of given information, such as references and lexical cohesion; if a message has a reference marker or lexically related words, the message can be considered to presuppose given information.

## References

1. Wakefield, J.: Catching a buzz: Software from Opion aims to turn Internet buzz into solid marketing science. *Scientific American* (November 2001)
2. Halliday, M.: *An Introduction to Functional Grammar*. Edward Arnold, London (1985)
3. Salton, G.: *Automatic Text Processing*. Addison-Wesley (1988)
4. Mani, I.: *Automatic Summarization*. John Benjamins Pub. Co., Amsterdam (2001)
5. McKeown, K.R., Klavans, J.L., Hatzivassiloglou, V., Barzilay, R., Eskin, E.: Towards multidocument summarization by reformulation: Progress and prospects. In: *Proc. of AAAI-99*. (1999) 453–460
6. Kosala, R., Blockeel, H.: Web mining research: A survey. *ACM SIGKDD Explorations* **1** (2000) 1–15
7. Lazarinis, F.: Combining information retrieval with information extraction for efficient retrieval of calls for papers. In: *Proc. of IRSG98*. (1998)
8. Kameyama, M., Arima, I.: A minimalist approach to information extraction from spoken dialogues. In: *Proc. of International Symposium on Spoken Dialogue*. (1993) 137–140
9. Matsumura, N., Ohsawa, Y., Ishizuka, M.: Influence diffusion model in text-based communication. In: *Proc. of WWW-2002*. (2002) to appear.