International Journal of Smart Computing and Artificial Intelligence
International Institute of Applied Informatics
2017, Vol.1, No.1, P.39 – 57

Automatic Summarization considering Thread Structure and Time Series in Electronic Bulletin Board System for Discussion

Ryota Kitagawa * , Katsuhide Fujita †

Abstract

On electronic bulletin board systems for discussion, a topic the users argue diversifies into multiple subtopics, and the entire structure becomes complicated. It is helpful to show users summarizations of the arguments because they can help in understanding the contents more easily without looking over from beginning to end of the discussion forum. The purpose of this paper is to propose an automatic summarization method of a single thread considering time series, reply relationships and user information. In the proposed method, a thread is restructured in several clusters by hierarchical clustering, and important sentences compressed with linguistic relationship of predicate argument structures are selected within each cluster using LexRank, which is a stochastic graph-based method for computing the relative importance of textual units. Finally, we conducted quantitative and qualitative analysis, comparing the proposed method with MMR. Both experimental results demonstrate that the proposed method can reduce redundancies more and extract fewer sentences unrelated to the whole context of the summary than the baseline. However, the proposed method included fewer important words than the baseline.

Keywords: Automatic Summarization, Electronic Bulletin Board System for Discussion, Clustering and LexRank.

1 Introduction

In recent years, rapid-growing technological advancement on the Web has stored enormous amounts of various information available to us in our lives. It is impractical to look over all of the information through brainpower alone because this requires an enormous amount of time. Therefore, automatic summarization approaches involving Natural Language Processing (NLP) techniques are expected to satisfy the demand of integrating information as a summary. Many approaches for the text summarization have been proposed based on centroid-based algorithm [16] and graph-based algorithm [4]. They aim to generate summaries that include no redundant sentences based on latent information. On the other hand,

* Department of Computer and Information Sciences, Graduate School of Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan
† Division of Advanced Information Technology & Computer Science, Institute of Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan

most of the document summarization methods focus on the documents written by a single or a few authors such as newspapers and academic papers.

There have emerged some kinds of electronic Bulletin Board Systems (BBS) from small-scale ones operated by a particular individual such as microblogs to large-scale ones such as Social Networking Services (SNS) that have massive unspecific users. In particular, BBS configured with discussions with other participants online will be called BBS for discussion in this paper. Such kind of BBS tends to have several sorts of topics simultaneously. For example, some users post new subtopics and others reply to other users' comments. Because the overall structure is constructed by a wide variety of posts with time, it becomes more complex as discussion is being developing. Assuming such situations, it is very difficult for users to always keep track of all changes on BBS for discussion. Thus, we consider that it is effective to present users with a summary of arguments.

In this paper, we propose a method of automatically generating a summary of a single thread in a bulletin board system for discussion to help users understand the contents of an argument easily. Our target is *Collagree* [5], which is one of the BBS for discussion and it structures threads whose posts reply to referential post as their own posts in its tree structure as well as generic bulletin board systems. First, all posts in a single thread are preprocessed and divided into several clusters consisting of similar ones based on characteristic elements of BBS for discussion: time series, reply relationship and user information. Second, some sentences constructed by unnecessary representations for the summary are removed and semantic relation called predicate argument structures extracts predicate clauses and the corresponding object clauses in each necessary sentence for shortening it. Then, the important sentences are extracted from each cluster based on LexRank [4], which is a multiple document summarization algorithm that expanded PageRank [15] to give the ranking score of each sentence. Finally, the candidate sentences are sorted from the oldest to the newest sequentially.

To evaluate the effectiveness of our approach, we conducted two evaluation experiments: qualitative and quantitative evaluations. In this paper, we compared the proposed method with the baseline which extracted sentences not overlapping the same information in a resultant summary by calculating Maximal Marginal Relevance (MMR) [3]. The experimental results show that the proposed method can reduce more redundancies and extract fewer sentences unrelated to the entire context of the summary than the baseline. However, the proposed method includes fewer important words than the baseline.

The remainder of this paper is organized as follows. In section 2, the related works on the automatic summarization method are shown. In section 3, the approach considering thread structure and time series is proposed. In section 4, we demonstrate the result of the preliminary experiment, and in section 5, we describe the results of evaluation experiments and discussions. Finally, we conclude this paper.

2 Related Work

The existing studies on automatic summarization focused on single or multiple documents written by a single or a few authors such as newspaper articles and academic papers. In other words, few studies deal with text data available on wide variety of the Web service like Bulletin Board Systems (BBS) and Social Networking Services (SNS). These generic approaches feature extracting important textual units (e.g. clause, sentence, word and etc.) so that they can integrate informative pieces of arbitrary units in original documents as a

summary. Extractive summarization approaches have been tackled and improved, applied to another research field such as mathematical programming, graph theory and etc. One of graph theoretic approaches is TopicRank proposed by Kitajima et al. [10] that extracts sentences from multiple documents based on latent topics estimated by Latent Dirichlet Allocation (LDA) [2]. They substitute distribution of the topics for traditional bag-of-words when calculating similarity between two sentences at the stage of sentence extraction by LexRank [4], which is an expansion algorithm of PageRank to multiple document summarization. In addition, penalizing the sentences in proportion as similarity with already retrieved sentences according to MMR enables to generate summaries including fewer redundant information. posting time Hatori et al. [6] proposed a graph-based model using PageRank algorithm. In their proposed method, clue words, reply relationship and lexical chain are incorporated with the fundamental concept of PageRank and important sentences in each remark and the topic of the thread are extracted. Another approach was to attempt to summarize the contents of a BBS, focusing on thread structure in which users post their remarks to other participant's remarks that have been already posted [13]. If a word in a certain remark appears in the parent remark, it is regarded as a word inheriting old information, otherwise as a word including new information. They proposed a summarization method introducing three-types indices of old information, new information and effect of new information. Finally, salient remarks and sentences can be extracted according to analytic result of an existent BBS based on those indices.

Text data available on the Web involving BBS often contains time information. For example, posts from participants store their posting time and drift to other matters with time. Kikuchi et al. [9] focused on document stream of Electronic Program Guide (EPG) with time series and proposed a hierarchical clustering method considering the closeness between average occurrence times of documents in each cluster. Combining the clustering method with C-value method for constructing compound words, keyword groups representing the transition of the topic are extracted. However, the aim of this approach is not to summarize the contents of document stream but to extract keywords expressing topic transition.

However, most existing works have not yet focused on summarizing multiple documents particularly like text stream on BBS for discussion. In this paper, we focus on the text written by wide variety of participants to deepen discussions and propose an automatic summarization method in BBS for discussion.

3 Automaric Summarization Method considering Time Series and Thread Structure

A BBS for discussion such as *Collagree* [5] is given a main theme and composed of several threads that deal with a single topic related to it. The thread described above refers to a set of posts related to a particular topic or issue. Fig. 1 shows the interface of *Collagree* with several functions; the first post of the user who made a thread becomes a parent post, and other users reply to it as a child post. There is a another case where some users reply to a child post as a grandchild post as you can see from Fig. 2.

Our goal is to propose a method to automatically generate summary of a thread in BBS for discussion. The proposed method follows a set of processes described below.

Step 1: Preprocessing

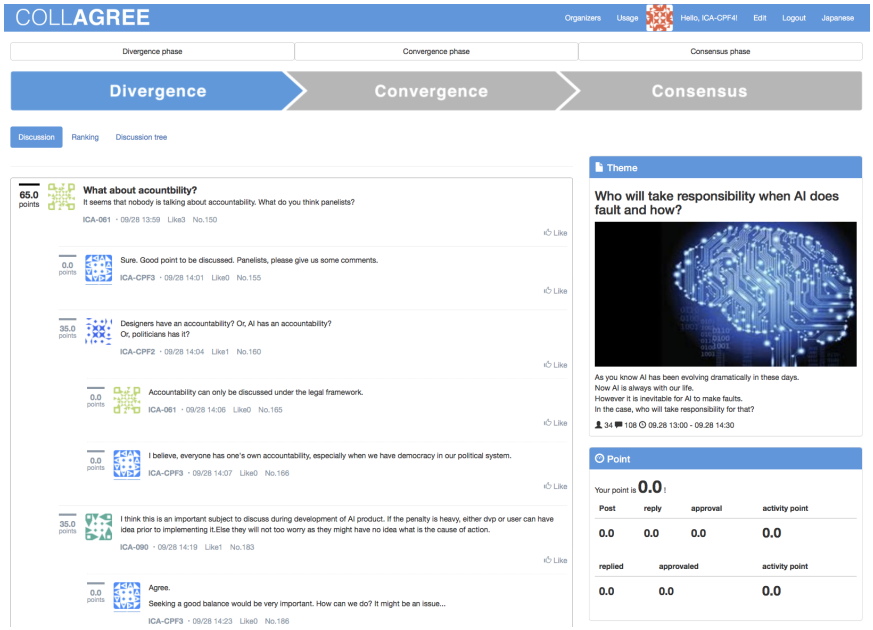

Figure 1: Interface of *Collagree*

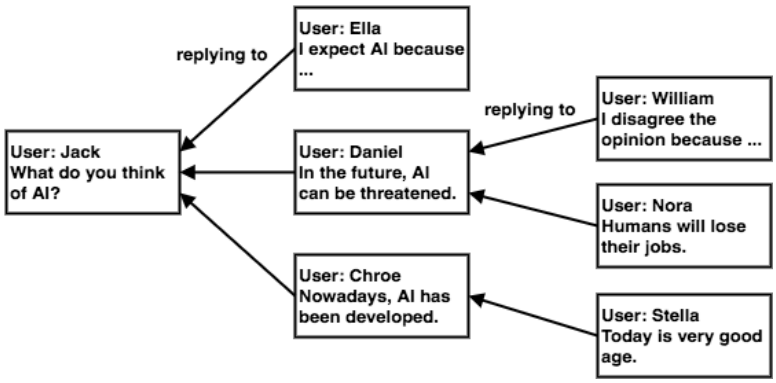

Figure 2: Overview of Thread Structure

Step 2: Hierarchical Clustering

Step 3: Unnecessary Sentence Removal

Step 4: Sentence Compression

Step 5: Sentence Extraction

Step 6: Postprocessing

3.1 Preprocessing

For arranging text data in Japanese of the BBS for discussion, the parentheses and URL are removed in body text of posts. These representations are likely to be unimportant information which causes the summary generated to be redundant and should be deleted before

summarization. It is also essential to clarify anaphors in the original document. In reference to the result of anaphora analysis using KNP [8], only anaphors that refer to a named entity are complemented. Furthermore, only nouns and verbs are extracted from each post or sentence based on the result of morphological analysis using MeCab [11] and are converted to the corresponding vectors using Paragraph Vector [12].

3.2 Hierarchical Clustering

We employ agglomerative hierarchical clustering to classify all posts in a single thread into several clusters on subtopics which is a subdivision of the thread. This results in discovering of posts that refer to similar matters and we consider that it leads to improve of qualitative coverage under the condition of the number of letters in the summary. In hierarchical clustering, the similarity between two vectors of documents can be calculated by cosine similarity defined as follows.

$$\cos(u,v) = \frac{\vec{u}\cdot\vec{v}}{|\vec{u}|\cdot|\vec{v}|} \tag{1}$$

where $\vec{u},\vec{v}$ are vectors of documents u and v, respectively. We adopt the group average method for the calculation of distance between clusters, which are defined as the average of all distances (i.e. non-similarity) between a couple of data belonging to two clusters, respectively. The number of clusters is determined based on the stop rule called Upper Tail [14]. This rule utilizes the fact that there are $N-1$ criterial distances when combining two clusters in regard to set of size N samples and proceeds to define an α of the distribution. Then, α_1 expresses a criterial distance that forms one cluster from last two clusters and α_{N-1} shows a first criterial distance. In this case, the clustering starts with $j=1$ and increments j until the following condition is satisfied.

$$\alpha_j \leq \bar{\alpha} + ks_\alpha \tag{2}$$

where $\bar{\alpha}$ is the average and s_α is the square root of the unbiased variance of the distribution. k is a constant and set as 1.

The initial post of a thread is often formed by two parts: introducing the topic to be argued in the thread and raising an issue which participants should consider. Therefore, all summaries should contain some representations included in the root post at their beginning. To prevent hierarchical clustering from classifying it into a cluster where other posts gather together, it forms a single cluster from only itself.

BBS for discussion consists of numerous posts with reply relationships, and the topics which users are interested in change as time passes. Thus, assuming that every participant has a tendency to post more and less biased posts to it, we consider three elements when calculating cosine similarity of the hierarchical clustering stage: the closeness of posting time, reply relationship and the degree of use by the same user.

Closeness of Posting Time: In a thread, posts whose posting times are close to each other are more likely to have the same topic. Thus, the attenuation function referring to the method by Kikuchi et al. [9] is applied to weighting posts.

$$w_{\text{time}}(u,v) = \exp\left(-\alpha_{\text{time}}(t_u - t_v)^2\right) \tag{3}$$

where $t_u,\ t_v$ are posting times of posts u,v on the second time scale and the absolute difference value is divided by 3,600 seconds (i.e. a hour) and α_{time} is a constant. Then, the

similarity between two posts can be obtained by Eq. (1) as follows:

$$\mathrm{sim}(u,v) = \cos(u,v) \cdot w_{\mathrm{time}}(u,v) \tag{4}$$

Reply Relationship: Threads consist of replies to each post on a particular topic or issue, and common topics are discussed in the thread. To perform clustering with retaining as much of the original relation as possible, the weighting method of the replies can be defined as follows:

$$w_{\mathrm{reply}}(u,v) = \begin{cases} \alpha_{\mathrm{reply}} & \text{if reply relationship} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where α_{reply} is a constant. Then, Eq. (4) can be updated as the following equation:

$$\mathrm{sim}(u,v) = \big(\cos(u,v) + w_{\mathrm{reply}}(u,v)\big) \cdot w_{\mathrm{time}}(u,v) \tag{6}$$

User Similarity: Posts by the same user are more likely to have similar ideas than those by different users. In addition, personal background such as a profession, a position or an experience affects their posts. Therefore, posts by the same user should be classified into the same cluster as much as possible. Accordingly, the weighting method considering it can be defined as follows:

$$w_{\mathrm{user}}(u,v) = \begin{cases} \alpha_{\mathrm{user}} & \text{if the same user} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where α_{user} is a constant. Then, Eq. (6) can be updated as follows:

$$\mathrm{sim}(u,v) = \big(\cos(u,v) + w_{\mathrm{reply}}(u,v) + w_{\mathrm{user}}(u,v)\big) \cdot w_{\mathrm{time}}(u,v) \tag{8}$$

We mention that in Eq. (8), $w_{\mathrm{reply}}(u,v), w_{\mathrm{user}}(u,v)$ are added to the cosine similarity between posts u and v in order to prevent the resulting similarity value from becoming too small to compute.

3.3 Unnecessary Sentence Removal

In BBS for discussion, some sentences not to be included in the summary are often seen: "Hello, [name]." and "I agree with [name]'s post." The unnecessary sentences classified as greetings, agreement or disagreement with other users' posts should be removed from each cluster before extracting important sentences, which prevent them from being selected as the important sentence for the summary.

Support vector machine (SVM), which is a binary linear classifier, can judge whether a sentence is required or not for the summary. A total of 400 sentences (200 positive examples and 200 negative examples) are selected from the past data of *Collagree* as training data. In addition, unnecessary sentences for the summary tend to be of short length, which can improve classification accuracy. The string-length of each sentence should be added to the corresponding document vector converted by Paragraph Vector, which means original vectors that increase by one dimension. The SVM model is learned using sentence vectors defined above as training data, and it classifies all sentences included in each cluster into two groups of sentences required for the summary or not.

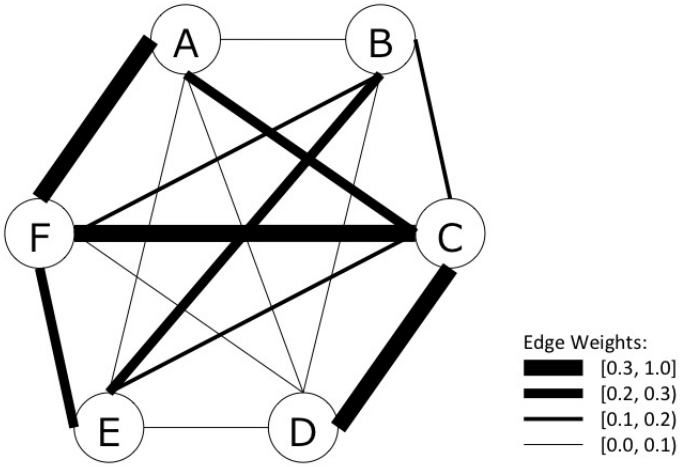

Figure 3: Graph Representation in LexRank

3.4 Sentence Compression

As a result of unnecessary sentence removal, the sentences must be compressed using the predicative argument structure, which refers to the semantic relation between a predicate that represents some kind of situation and arguments that are essential for it. It is preferred that a summary generated has a large amount of information. Therefore, compression of sentences in advance leads to a decrease in the number of characters in a sentence and enables us to extract more sentences from each cluster. JUMAN / KNP [7][8], a tool for syntax analysis, can extract the argument structure from sentences in clusters. Only clauses that are predicators in a sentence and arguments are extracted, and the new sentence is built from them.

3.5 Sentence Extraction

The purpose of this section is to select important sentences from all clusters containing no unnecessary sentences and compressed sentences. We consider that LexRank, which is a graph-based multiple document summarization technique, is suitable for this. In the proposed method, the scores among sentences contained in the same cluster are significantly different at any time even if the number of sentences in a cluster is extremely small. Therefore, we employ the Continuous LexRank using a weighted graph that applies normalized similarity between sentences to the corresponding edges as in Fig. 3.

LexRank: LexRank is a summarization technique proposed by Erkan et al. [4] and expanded from PageRank to multiple document summarization. For instance, the nodes mean the sentences in the documents, and the edges mean the similarity between two nodes (i.e. sentences) as in Fig. 3. The importance of a node can be obtained by the adjacent nodes based on the concept of eigenvector centrality in the graph representation. LexRank u is

defined as follows:

$$p(u) = \frac{(1-d)}{N} + d \sum_{v \in adj[u]} \frac{\mathrm{sim}'(u,v)}{\sum_{z \in adj[v]} \mathrm{sim}'(z,v)} p(v) \tag{9}$$

where N is the number of sentences in the target document, $adj[u]$ is a set of adjacent nodes. u means the sentence, d is damping factor for jumping to a non-adjacent node with a constant probability, whose value is set to 0.85 as determined by the reference [4]. $\mathrm{sim}'(u,v)$ is modified similarity between sentences defined as following section.

Similarity based on Relationship between posts: Important sentences for the summary should have connections between sentences as long as the summary isn't verbose to enable users to figure out the meaning more easily. Thus, we employ the similarity between sentences in LexRank that clusters obtained by hierarchical clustering consist of sentences extracted from multiple posts. This multiplies similarity between two sentences defined as Eq. (1) to the value of considering the relationship between two posts encompassing each sentence as follows, which refers to Hatori et al. [6].

$$\mathrm{sim}'(u,v) = \cos(u,v) \cdot \mathrm{rel}(u,v) \tag{10}$$

where $\mathrm{rel}(u,v)$ is set to 2.0 when both sentences u,v are included in the same post, $\mathrm{rel}(u,v)$ is set to 1.5 when both sentences in two different posts have a replying relationship, and $\mathrm{rel}(u,v)$ is set to 1.0 when both sentences in two different posts have no relation.

3.6 Postprocessing

The summary needs to rearrange important sentences extracted from each cluster compatibly with the context. Since each remark of BBS for discussion has a posting time, it is possible to arrange important sentences extracted from all clusters sequentially. When some sentences contained in the same post are selected at the stage of sentence extraction, they should all be given the same posting time so that they can be arranged in order of their initial position in the sentence at the posting time. This prevents time series and the positional relationship between sentences from being disturbed and enables us to present a summary that reflects the entire flow of the thread.

4 Preliminary Experimental Result

To evaluate the effectiveness of the proposed method in BBS for discussion, we conduct the preliminary experiments by subjective questionnaires. We use the datasets discussing five themes (human rights, environment, disaster, attractiveness and town planning) in large-scale social experiments carried out over a period of 2013 to 2015 in Aichi prefecture and Nagoya city, Japan using *Collagree* [5]. In particular, we selected one thread from one discussion theme. Table 1 shows the information of each thread to evaluate the summaries by the proposed method. All threads have at least 10 posts. 9 students in Tokyo University of Agriculture and Technology to evaluate by a five-point scale (5 is very good; 1 is very poor) in the five evaluation items in the subjective questionnaire. We decided the subjective evaluation items based on Asahara et al. [1].

- *Readability:* The summary doesn't include incomprehensible sentences
- *Non-redundancy:* The same information isn't repeated

Table 1: Statistical Information of Datasets in Preliminary Experiments

Thread No.	Discussion Theme	#User	#Post
No. 1	Human rights	6	13
No. 2	Environment	5	16
No. 3	Disaster	4	10
No. 4	Attractiveness	5	14
No. 5	Town planning	7	10

Table 2: Preliminary Results of Evaluating Summaries by Our Proposed Method

Evaluation Item	50%	25%	10%
Readability	3.91	3.87	4.36
Non-redundancy	4.36	4.47	4.64
Comprehension	3.98	3.24	2.76
Coverage	4.13	3.20	2.27
Structure	4.07	4.11	4.51

- *Comprehension:* The summary provides an outline of the original document
- *Coverage:* The summary contains enough important words from the original document
- *Structure:* The timeline of the summary is correct

The parameters of our proposed method are set as follows as a result of trying various values in range empilically: $\alpha_{\text{time}} = 0.01$, $\alpha_{\text{reply}} = 0.1$, $\alpha_{\text{user}} = 0.1$. We introduce summarizing rate that denotes the proportion length of a summary to that of original documents. To evaluate the influences of the summarizing rate to the quality of the summaries, we set three different summarizing rate: $50\%, 25\%$ and 10%.

Table 2 shows the averages of the subjective evaluation items in the questionnaire when the summarizing rates are set to 50%, 25% and 10%. The averages of the comprehension and the coverage get higher as the summarizing rate becomes large. On the other hands, the averages of non-redundancy and structure in the small summarizing rate (10%) get higher as the summarizing rate becomes small. This is because that the summaries of the high summarizing rate can contain more important sentences and words. In addition, comprehension becomes small in the low summarizing rate (10%), despite that the averages of readability and structure are the highest. In fact, readability and structure of the summaries don't have much relationship with summarizing rates because some sentences are compressed using the predicative argument structure in sentence compression and post-processing sections in our proposed method.

5 Experimental Result

We conducted additional experiments using summaries automatically generated from the text to evaluate the effectiveness of the proposed automated summarization method considering the thread structure and time series for electric BBS for discussion. These experiments demonstrated the quantitative evaluations about the contents of the summary and qualitative evaluations based on subjective evaluation by some subjects.

5.1 Baseline Method: Maximal Marginal Relevance (MMR)

In the experiments, the summaries produced by the proposed method are compared with those produced by the baseline using Maximal Marginal Relevance (MMR). MMR, an index to decrease redundancy in a document, is used especially in query-oriented document summarization. Details of MMR are as follows:

$$\mathrm{MMR} = \arg\max_{D_i \in R \setminus S} \left[\lambda \mathrm{Sim}_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} \mathrm{Sim}_2(D_i, D_j) \right] \tag{11}$$

where Q is a query and R is a set of sentences retrieved by Q. S is a set of important sentences in R which have been already selected and $R \setminus S$ is a set of sentences in R which have not been selected yet. $\mathrm{Sim}_1(D_i, Q)$ is similarity between a sentence in a target set of sentences and Q and $\mathrm{Sim}_2(D_i, D_j)$ is also similarity between two sentences, one of which is a sentence in a target set of sentences and the other has been already selected as one of the summary constituents. These two similarities consider the relevance to the query and the redundancy reduction at the same time. We can determine how much weight either similarity is attached to by adjusting the parameter λ in Eq. (11).

In the general online discussion forums with thread structures, the root post has many child posts in the certain thread which refers to the same topic like the tree structure. Therefore, in this experiments, we decide the first sentence of the root post in the target thread as Q and the set of sentences other than it in the same thread as R in Eq. (11). First, the first query sentence is included in S as 0-th important sentence. Second, the sentence with the highest MMR value is added to S after calculating MMR values of all remaining sentences in R until the summarizing rate is achieved. The parameter λ is set to 0.50 because the relevance to the query and the redundancy reduction are considered, equally.

The baseline utilizes preprocessing in section 3.1 and postprocessing in section 3.6 to rearrange the raw text and get the output format which is same as the proposed method. Under this condition, the baseline converts sentence to the corresponding vectors based on Paragraph Vector like the proposed method and the similarity between two sentences are evaluated by the Eq. (1).

5.2 Quantitative Evaluation

We used the datasets of 6 discussion themes (human rights, environment, disaster, attractiveness, town planning and spread of smart phones (additional one)) carried out from 2013 to 2015 through *Collagree*, which is same datasets in the preliminary experiments. Table 3 shows the statistical information of each dataset in the experiments.

All threads in the datasets have more than 5 posts and are summarized in both the proposed method and the baseline when the rates of summarizing the text are 50%, 25% and 10%, respectively. Then, three parameters of the hierarchical clustering are set as

Table 3: Statistical Information of Datasets in Quantitative Evaluation

Dataset	Discussion Theme	#User	#Post
Nagoya	Human rights, Environment, Disaster, Attractiveness	266	1,151
Aichi	Town planning	75	355
InLab	Spread of smart phones	41	205

follows: $\alpha_{\text{time}} = 0.01,\ \alpha_{\text{reply}} = 0.1,\ \alpha_{\text{user}} = 0.1$ used in the preliminary experimental results of the subjective questionnaire. The summaries generated by both methods are compared by two evaluation metrics defined in the following section.

Non-redundancy: Similar information is repeated or not

To examine how similar information is repeated in a summary generated, when D is a set of N sentences including the summary generated, the similarity of the summary is defined as follows:

$$\text{Sim}(D) = \frac{1}{N}\frac{1}{N}\sum_{d_i \in D}\sum_{d_j \in D}\cos(d_i, d_j) \tag{12}$$

where $d_i,\ d_j$ are sentences in a document D. We assign the corresponding vector of each sentence employing Paragraph Vector to sentences $d_i,\ d_j$ in Eq. (12) and calculate non-redundancy of summaries in each method and summarizing rate.

Coverage: Important words in original document are included

Firstly, we applied BM25 to the weights of all nouns extracted from a particular thread in terms of relativeness to overall context of the thread. Although BM25 originally calculates the degree of relativeness between a document D and a set of query $Q = \{q_1, q_2, \ldots, q_n\}$, we can also utilize the improved BM25 [17] for weighting a word w to a set of documents $D = \{d_1, d_2, \ldots, d_n\}$ defined as follows. The improved BM25 is implemented in *Collagree* as the following function of remarkable keywords.

$$\begin{aligned}\text{score}(w, D) &= \sum_{i=1}^{n}\text{IDF}(w) \times \frac{f(w, d_i) \times (k_1 + 1)}{f(w, d_i) + k_1 \times \left(1 - b + b \times \frac{|d_i|}{avgdl}\right)} \\ \text{IDF}(w) &= \log\frac{N - df(w) + 0.5}{df(w) + 0.5}\end{aligned} \tag{13}$$

where $f(w, d_i)$ is the frequency of the word w in the document $d_i, df(w)$ is the number of documents in which a word w appears, $|d_i|$ is the number of characters in a document d_i, $avgdl$ is the average number of document sets D, and $k_1,\ b$ are constant parameters. D is applied to a thread and d_i is an post in the thread, and $k_1 = 2.0, b = 0.75$, which values are often used in BM25.

Secondly, the set of nouns ranked by BM25 are divided into ten groups every additional ten percent in order of their own scores. After that, we tried to examine how many nouns belonging to each groups the generated summary contains and analyzes the distribution of nouns reflecting the ratio. For instance, assuming a certain case that there are 15 nouns

Table 4: Non-redundancy of The Proposed Method and The Baseline

Summarizing Rate	Proposed Method	Baseline
50%	0.083	0.075
25%	0.140	0.141
10%	0.298	0.343

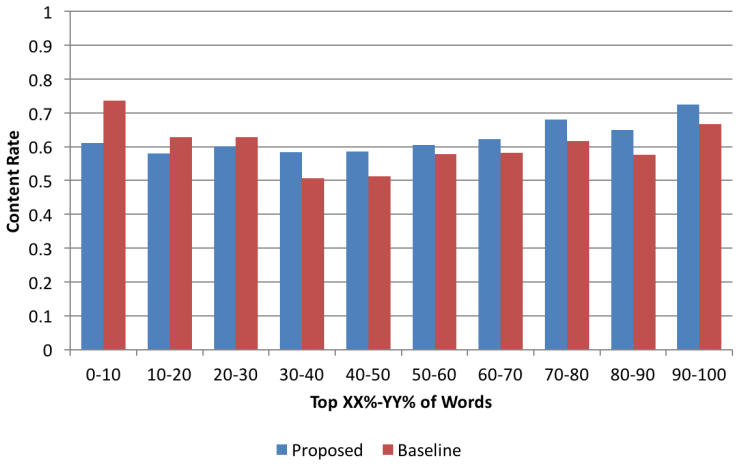

Figure 4: Coverage (summarizing rate: 50%)

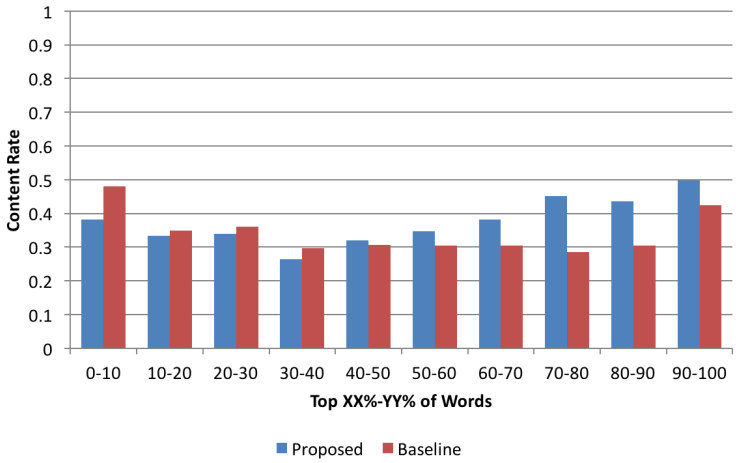

Figure 5: Coverage (summarizing rate: 25%)

belonging to a group1 (e.g. a group of nouns ranked among top 0 to 10 percent) and 9 nouns out of group1 contained in the generated summary, containing-ratio of group1 is calculated as the following: $9/15 = 0.60$.

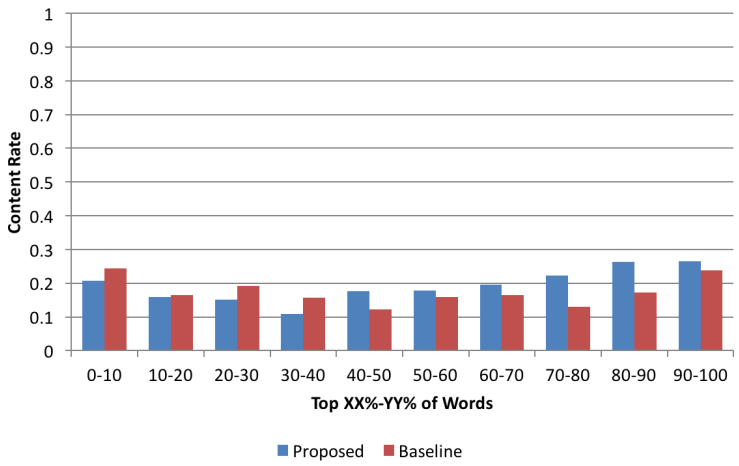

Figure 6: Coverage (summarizing rate: 10%)

Table 5: Statistical Information of Datasets in Qualtative Evaluation

Thread No.	Discussion Theme	#User	#Post
No. 1	human rights	9	24
No. 2	environment	5	14
No. 3	disaster	3	12
No. 4	attractiveness	8	19
No. 5	town planning	5	6
No. 6	spread of smart phones	7	10

Experimental Results

Table 4 shows the average similarity of all resultant summaries by the proposed method and baseline at every summarizing rate. When the summarizing rate is either 50% or 25%, our proposed method doesn't have a significant difference from the baseline. However, the results of our proposed method have less redundancy than the baseline when the summarizing rate is 10%. Figs. 4-6 show the histograms that represent how many nouns are extracted from the thread before the summarization in some distributions. The baseline has more words with top-ranked groups only compared with our proposed method. Therefore, our proposed method extract fewer essential words but more general words.

One of the reason that the coverage of the proposed method was poor is the parameters of calculating similarity between two sentences at the stage of sentence extraction in section 3.5. We aimed to automatically generate a summary whose sentences have coherence along the outline with each other by extracting sentences from the same post or replying/replied posts. It is likely that unimportant sentences are extracted because they are weighted heavily as well as important sentences in the same posts or replying/replied posts.

Table 6: Result of Evaluating Summaries by the Proposed Method

Evaluation Item	**50%**	**25%**	**10%**
Readability	3.55	3.59	3.95
Non-redundancy	4.06	4.42	4.67
Comprehension	4.00	3.35	2.67
Coverage	3.97	3.26	2.56
Focusing	3.67	3.62	3.26

Table 7: Results of Evaluating Summaries by the Baseline

Evaluation Item	**50%**	**25%**	**10%**
Readability	4.42	4.33	4.50
Non-redundancy	4.06	4.35	4.44
Comprehension	4.08	3.68	2.97
Coverage	4.05	3.48	2.71
Focusing	3.41	3.56	3.39

5.3 Qualitative Evaluation

We used the same datasets with 6 discussion themes as the quantitative evaluation and selected one thread from each dataset (i.e. one thread is associated with one discussion theme). Table 5 shows the statistical information of each thread in the experiment. Subjective evaluation items for summaries proposed by human subjects [1] were improved so as to be appropriate for the experiment. From the preliminary experiments, an evaluation item *Structure* is replaced to *Focusing* because a significant difference between the structure of summaries which rearranged the order of its component sentences by the proposed method and the baseline isn't expected. 11 students and research staff evaluated the summaries by the proposed method and baseline for the evaluation questionnaires on a five-point scale (5 is very good; 1 is very poor) for the five items below.

- *Readability:* the summary does not include incomprehensible sentences
- *Non-redundancy:* the same information is not repeated
- *Comprehension:* the summary provides an outline of the original document
- *Coverage:* the summary contains enough important words from the original document
- *Focusing:* the summary does not contain information unrelated to the context

Tables 6 and 7 show the averages of each evaluation item by the proposed method and the baseline at each summarizing rate. Table 8 shows the result of t-test, which is one of the statistical tests to evaluate whether there is a significant difference between two different groups or not. In this experiment, we evaluated each evaluation item between the summary by the proposed method and the one by the baseline. The red-colored values mean that the

Table 8: Result of t-test between Summaries by the Proposed Method and the Baseline

Evaluation Item	50%	25%	10%
Readability	0.000	0.000	0.000
Non-redundancy	1.000	0.388	0.010
Comprehension	0.470	0.009	0.017
Coverage	0.470	0.042	0.191
Focusing	0.034	0.654	0.338

∗The red-colored values means that the proposed method has a significant difference to the baseline and the blue-colored values means that the baseline has a significant difference to our proposed method.

proposed method has a significant difference to the baseline and the blue-colored values means that the baseline has a significant difference to our proposed method.

In general, the summaries by the proposed method and the baseline show better performance at the higher summarizing rate (50%). In addition, the proposed method is high non-redundancy, though the other items except for readability become small rapidly compared with the baseline. In non-redundancy and focusing, the proposed method outperforms the baseline at all summarizing rates except for focusing at 10%. The proposed method has low redundancy and can remove representations with unnecessary information. However, in readability, comprehension and coverage, the baseline outperforms the proposed method at all summarizing rates. Thus, compression of sentences can generate the shorter sentences from unnecessary clauses in the original sentences that are hard to read, and the proposed method cannot provide a sufficient amount of essential information.

The baseline is a simple automated summarizing method that calculates the MMR values and extracts sentences with high-score without compressing the sentences. It is natural that the baseline outperformed the proposed method in readability at all summarizing rates because MMR uses the sentences for summarizations without compressing the sentences.

In addition, the baseline considers the first sentence of the root post as the query in Eq. (11). It helps users to understand the topics of discussions and follow the next posts. Therefore, users are easy to understand the topics of discussions since the baseline consider the head sentence of the thread in the summaries. The summary should consider both the introduction and the problem from the root post in the thread, even if the summarizing rates prompt them to understand the contents easily.

The results of non-readability demonstrate that the unnecessary sentence removal part and correctness of sentence extraction in our proposed method are effective to generate the summaries. Furthermore, we confirmed that the proposed method is higher than the baseline using MMR in selecting non-overlapped informative sentences for summarizing thread.

Finally, we discuss about the evaluation results of each thread. In non-redundancy, the proposed method outperformed over baseline on overall evaluation. Although in town planning thread with the least number of posts, all averages of the proposed method were the best values, in human rights thread with the best number of posts, the averages were the worst scores when summarizing rates were 50% and 25%. In contrast, in coverage and focusing, the opposite tendency was discovered. The proposed method can extract important information, focusing on diverse information included in threads with many posts, whereas in order to eliminate more redundant representations we need to reconsider effective ap-

proach of sentence compression.

6 Conclusion

The aim of this paper was to help users understand the contents of electric BBS for discussion, and we proposed an automatic summarization method considering the thread structures and posting time, etc. We demonstrated the effectiveness of our proposed method on *Collagree*. In the proposed method, all threads were structured in several clusters by hierarchical clustering, and critical sentences were selected from each cluster using LexRank, which is a stochastic graph-based method for computing the relative importance of textual units. In the quantitative evaluation, the proposed method outperformed the baseline in non-redundancy at the lower summarizing rate. On the other hand, the coverage of the proposed method was inferior to the baseline. In addition, the result of the qualitative experiment showed that the proposed method outperformed the baseline in non-redundancy and focusing. However, it was revealed that the proposed method has been scored lower than baseline in comprehension because sentence compression can make readability lower markedly and sentence extraction can miss crucial information in the root post of the thread at lower summarizing rate. There are some limitations about the proposed method. Text data with rich structured features (e.g. reply relationship, user information and posting time) in BBS for discussion is given to it as input, thus another BBS for discussion gives text data to the same features can embed our approach in the own system, but raw text data without sufficient features cannot be applied. Because it cannot cover important information in original text under tight restriction of character length of summary at extremely lower summarizing rate, bad summaries can be generated.

Possible future work involves improvements of the sentence compressing to improve the coverage and readability of the summaries. In addition, we will perform more detailed analysis of each phase of our proposed method. It is necessary to investigate how the parameters in hierarchical clustering affect the quality of summary, and we intend to employ subjects to create referential summaries from datasets of *Collagree* manually. Now, *Collagree* is introducing a mechanism for automatically building discussion tree representing reply relationship and agreement/disagreement, which helping users and facilitators to discuss. We also consider incorporating our system into it.

Acknowledgments

This work was supported by CREST, JST.

References

[1] M. Asahara, M. Sugi, and S. Yanagino. Bccwj-summ: A summarization corpus of the ‘balanced corpus of contemporary written japanese’. In *Proceedings of The 7th Wordshop of the Japanese Corpus Linguistics*, pages 285–292, 2015.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal Machine Learning Research*, 3:993–1022, 2003.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of The 21st Annual International Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR) Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM.

[4] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479, 2004.

[5] T. Ito, M. Okumura, T. Ito, and E. Hideshima. Implementation of a large-scale discussion support system collagree - large-scale discussion support based on a weakly structured discussion process -. *Journal of Japan Industrial Management Association*, 66(2):83–108, 2015.

[6] H. Jun and A. Murakami. Extraction of important sentences and topics from online discussion using the thread structure and lexical chain. In *Proceedings of The 16th Annual Meeting of The Association for Natural Language Processing*, pages 290–293, 2010.

[7] D. Kawahara and S. Kurohashi. Japanese morphological analyzer juman, 2012. http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN.

[8] D. Kawahara and S. Kurohashi. Japanese systax / case structure / anaphora analyzer knp, 2012. http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP.

[9] M. Kikuchi, M. Okamoto, and tomohiro Yamazaki. Extraction of topic transition from document stream based on hierarchical clustering. *Database Society of Japan (DBSJ) Journal*, 7(1):85–90, 2008.

[10] R. Kitajima and I. Kobayashi. Graph based multi-document summarization with latent topics. *Intelligence and Information*, 25(6):914–923, 2013.

[11] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of The 2004 Conference on Empirical Methods on Natural Language Processing*, pages 230–237, 2004.

[12] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1188–1196, 2014.

[13] Y. Matsuo, Y. Ohsawa, and M. Ishizuka. Minig and summarizing conversational data on electrical message boards. *The 16th Annual Conference of the Japan Society of Artifitial Intelligence*, 16:1–4, 2002.

[14] R. Mojena. Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal*, 20(4):359–363, 1977.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[16] D. R. Radev, H. Jing, M. Styś, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938, 2004.

[17] K. Yoshioka and M. Koeda. Extraction of related words and classification of words using bm25. In *Proceedings of The 78th National Conversation of Information Processing Society of Japan*, volume 6, page 1, 2012.

Appendix

Fig. 7–10 show the original thread of town planning that has the least number of posts and some examples of the summaries in Japanese generated by the proposed method at summarizing rate 50%, 25% and 10%. Compared with summaries of other threads, they were highly evaluated in non-redundancy, but were poorly evaluated in coverage and focusing.

username	body
サウスライト	今は、ネット通販で物が買える時代。街中に店を構えるとしたら、物販ではなく、飲食の方がいいのでしょうか？物販店だと試着して、購入はネットでというお客さんも出てきますね。
NAO	体験、交流とかネットじゃ無理なものという視点はいいですね。
村人	ネットスーパーの普及も見逃せませんよね。イオンとかアピタとかヨーカドーとかが何でも宅配してくれるやつ。駅前商店の競争相手は、郊外のショッピングセンターだけでなく、これからはネット販売だと思います。高齢者の利用も増えるでしょうし。駅前の物販店はますます厳しくなるでしょうね。
ととろ	超高齢社会に向けてこらからは大型スーパーよりは近くのコンビニやネットスーパーに時代が来るように思います。
NAO	宅配業界も運ぶ担い手の確保があやしくなってきてるみたいです
S.hattori	確かに、買う側にとっては、選択肢が増えることはいいことだと思います。お店側にとっては、競争相手が増えたということになると思いますが、結局は、そのお店どうブランディングしたいか、によると思います。ネット上で、その商品の魅力を伝え切れるのであれば、いいと思いますが、、。

Figure 7: Original Thread of Town Planning

username	body
サウスライト	今は、ネット通販で物が買える時代
サウスライト	物販店だと試着して、購入はお客さんも出てきますね
村人	アピタとかヨーカドーとかが宅配してくれるやつ
村人	競争相手は、郊外のショッピングセンターだけでなく、これからはネット販売だと
NAO	宅配業界も運ぶ担い手の確保があやしくなってきてるみたいです
S.hattori	買う側に選択肢が増えることはいいことだと
S.hattori	お店側に競争相手が増えたとお店どうブランディングしたいか、に
S.hattori	ネット上で、商品の魅力を伝え切れるのであれば、いいと
ととろ	こらからは大型スーパーよりは近くのコンビニやネットスーパーに時代が来るように

Figure 8: Summary of Town Planning by the Proposed Method (Summarizing Rate: 50%)

username	body
サウスライト	物販店だと試着して、購入はお客さんも出てきますね
村人	アピタとかヨーカドーとかが宅配してくれるやつ
村人	競争相手は、郊外のショッピングセンターだけでなく、これからはネット販売だと
S.hattori	買う側に選択肢が増えることはいいことだと
S.hattori	ネット上で、商品の魅力を伝え切れるのであれば、いいと

Figure 9: Summary of Town Planning by the Proposed Method (Summarizing Rate: 25%)

username	body
村人	アピタとかヨーカドーとかが宅配してくれるやつ
S.hattori	買う側に選択肢が増えることはいいことだと

Figure 10: Summary of Town Planning by the Proposed Method (Summarizing Rate: 10%)