# Multiple-text Summarization for Collective Knowledge Formation

**Tomohiro FUKUHARA*, Hideaki TAKEDA* and Toyoaki NISHIDA****

***Graduate School of Information Science,**
**Nara Institute of Science and Technology**
**8916-5 Takayama, Ikoma city, Nara 630-0101, Japan**
**E-mail: {tomohi-f,takeda}@is.aist-nara.ac.jp**

****School of Engineering, The University of Tokyo**
**7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan**
**E-mail: nishida@kc.t.u-tokyo.ac.jp**

## Abstract

Multiple-text summarization method for facilitating a process of the collective knowledge formation is proposed. There are enormous pieces of information represented by WWW pages which contain various topics in early stage of the community. To facilitate this process, organizing various unordered pieces of information is needed. We facilitate this process by multiple-text summarization. Proposed method consists of topic identification method and context-based summarization method. Topic identification method finds a topic including central information over the text set. We identify partialities of topics based on skewness and kurtosis of a word frequency. Context-based summarization method generates summaries by linking relevant topics. Summarization in our approach is to find a context that is an ordered set of sentences. We find a context based on theme and focus of a sentence which represent central information within a sentence. We developed a prototype system called Topic Showcase. Experimental results demonstrate the availabilities of proposed methods in identifying topics from each cluster and supporting users in forecasting contents of the texts.

## 1 Introduction

According to the wide spread of the network connectable electricities, a lot of network communities are arising. The network community (we refer network community as "the Community" below) here is an information resource that is consisted of human resources and information resources. The human resource is regarded as an implicit knowledge such as know-hows or experiences owned by members of the community who have expertise in a particular domain. The information resource is an explicit knowledge such as email, discussions on BBS, databases shared by members of the community.

One of the features of the community is the collective knowledge formation. The collective knowledge is a knowledge constructed and maintained among the member of the community, such as FAQs, community's knowledge bases and other documents shared within the community. The collective knowledge is formed by gathering, filtering and editing of the community's information resource.

One of the issues in a process of the collective knowledge formation is a complexity of the community's information resource. In early stage of the community, community's information resource is unorganized because each member add information to the community's information resource planlessly. Accordingly, members can't find out an overview of the community's information resource. Consequently, the process of the collective knowledge formation is delayed in early stage of the community. Accelerating the process of the collective knowledge formation, organizing information by summarization of unorganized information is needed.

We propose a multiple-text summarization method for facilitating a process of the collective knowledge formation. Summarizing community's information resource, each member of the community can find out what kind of information has been discussed or shared among the community. Proposed method consists of 2 sub-methods, (1)topic identification method and (2)context-based summarization method. Topic identification method identifies topics, which indicate crucial sentences among texts, based on statistical information of words in texts. Context-based summarization method generates a summary putting relevant sentences together as a context. A **summary** here is an ordered set of sentences to which each sentence relevant.

In the remain of this paper, we first analyze the process of the collective knowledge formation and discuss of the necessity of the summarization. We then describe the proposed method: topic identification based on statistical information and summarization based on the context. We show Topic Showcase: a prototype system for multiple-text summarization. We finally show the results of an evaluation and discuss the results.

## 2 Multiple-text Summarization for Collective Knowledge Formation

Summarizing multiple-text is needed for accelerating the process of the collective knowledge. We firstly analyze the process of the collective knowledge and then proposed multiple-text summarizing method.

### 2.1 Process of the Collective Knowledge Formation

The process of the collective knowledge formation consists of 3 steps.

**Gathering Step**
In this step, each member of the community gathers information or describes their own knowledge into documents. This process is mostly accomplished as personnel and independent process.

**Filtering Step**
Each pieces of information are gathered and filtered in this process. Some pieces of information are filtered and remains are shared among the community. This process feeds back to the previous process.

**Describing Step**
The collective knowledge is constructed according to the discussion. Core members or editorial board of the community are formed for describing knowledge as documents. The collective knowledge is described as FAQs, WWW pages and knowledge bases.

An issue that we focus in the process of the collective knowledge formation is how to organize a disordered community's information resource. Various kinds of information are included in the community's information resource in early stage of the community. These pieces of information must be sorted and organized in order to be utilized as the collective knowledge. However, it requires much time to organize these unsorted pieces of information. Accelerating the process of the community knowledge formation, ordering information is needed.

Related works of ordering the community's information is proposed. Matsubara proposed CommunityBoard which is a system for visualizing the structure of discussions between members of the community[1]. CommunityBoard facilitates users to know what kind of topics have been discussed or who is discussing currently.

Hirata proposed CoMeMo-Community which is a system for simulating discussions among members of a community[2]. CoMeMo-Community simulates discussions using agents who each have own memories which represent their each owner's knowledge. Discussions between agents are performed by linking relevant memories to other memories.

These systems have merits on knowing a process of the collective knowledge formation, but don't organize information directly. Organizing information such as classifying, extracting or summarizing information is needed for accelerating the process of collective knowledge formation.

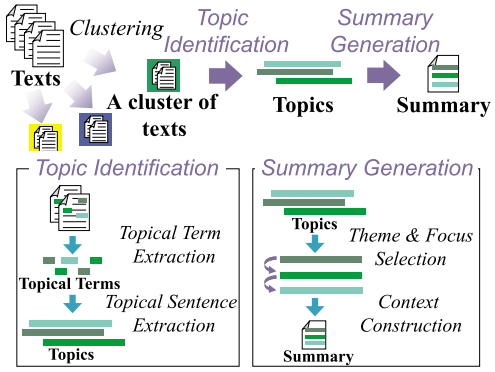


Figure 1: Overview of multiple-text summarization.

We propose a text summarization method. Our aim in the text summarization is to assist members of getting an overview of the community's information resource. Community's information resource in early stage of the community is regarded as a unorganized information pool where members of the community hardly grasp all of information resources shared in the community. Condensing an unorganized information pool into a summary, members can find out what kind of topics are discussed and focused in community's information resource.

### 2.2 Multiple-text Summarization

We propose a multiple-text summarization method for accelerating the process of the collective knowledge formation. One of the causes of delaying the process is to comprehend the community's information resource including various kinds of topics. We regard community's information resource as multiple-text including various kinds of topics. Summarizing multiple-text, the process of the collective knowledge formation is accelerated.

There are related works on multiple-text summarization. McKeon proposed a method for summarizing a series of news articles[3]. Proposed method is based on information extraction which extracts specified information using cue words. Summary is formed based on a comparison between extracted pieces of information.

Nanba proposed a method for summarizing a set of scientific papers[4]. Proposed method utilizes a reference of citation among papers. Identifying a purpose of citation, proposed method generates a summary according to the purpose of reference.

One of the merits of these methods is that their methods can product a detailed summary which considers the contents of text itself and relationships among texts. However, these methods restrict texts

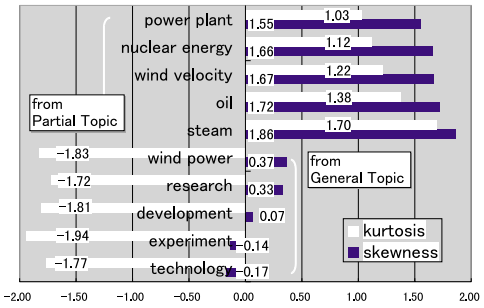


Figure 2: Comparison of skewness and kurtosis among words within a general or partial topic.

types because their methods rely on specific domain knowledge.

In the process of the collective knowledge formation, there are various types of texts from formal information such as news articles, research papers to informal information such as WWW pages, e-mails. Former methods have a limitation on summarizing texts because they rely on domain knowledge. In summarizing community's information resource which includes various types of information, domain independent summarization method is needed.

We propose a multiple-text summarization method being domain independent. We use statistical information of words in texts and identifies topics which indicate the points of texts. Figure 1 shows our approach.

Proposed method consists of 2 sub-methods.

**Topic Identification**
Topic identification method identifies topics. A **topic** is a sentence indicating the points of texts. We identify topics using statistical information of words from classified texts.

**Context-based Summarization**
Context-based summarization method generates a summary which has a context. A **context** here is a relationship between sentences and a **summary** is a set of topics which has sequential order originated from source topic. We generate a summary for each topic by linking relevant topics.

## 3 Topic Identification based on Statistical Information

We identify topics based on classification and statistical information of texts. We firstly describe that skewness and kurtsis indicate the partiality of topics among entire text set, and then we show our proposed method.

### 3.1 Skewness and Kurtosis

**Skewness** and **Kurtosis** are both statistical measure. **Skewness** indicates a distortion of a distribution and **kurtosis** indicates a centralization of a distribution. Formulas (1) and (2) show the definitions of skewness and kurtosis.

$$\alpha_3 = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{s^3} \quad (1)$$

$$\alpha_4 = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{s^4} - 3 \quad (2)$$

$x_1, x_2, \cdots, x_n$ are input data, $\bar{x}$ is an average and $s$ is a standard deviation.

Skewness and kurtosis show a difference between general topic and partial topic. **General topic** is a topic which indicates a common topic to the entire texts. General topic is specified from all of texts. **Partial topic** is a topic which indicates a unique topic to some parts of the entire texts. Partial topic is specified from each category of classified texts.

Figure 2 shows an example that skewness and kurtosis indicate a difference between general topic and partial topic. In figure 2, the words appeared in the upper part of the figure (i.e. "power plant" to "steam") are extracted from a general topic, and the words of the lower part of the figure (i.e. "wind power" to "technology") are extracted from partial topics. We selected these words manually from general topic and partial topic retrieved by the keyword "wind generator".

This figure shows that the words selected from partial topic indicate highly values against general topic. Thus we can measure the partiality of the topic using skewness and kurtosis. We use skewness and kurtosis of the words as measures for knowing the partiality of each topic.

### 3.2 Topic Identification

We identify topics by the following steps.

1. Classifying texts into categories
2. Calculating the partiality of each word
3. Identifying general topic and partial topic

At first step, we classify texts into some clusters in which each text similars to each other. We apply hierarchical clustering method to the text set. We use VSM(Vector Space Model) and cosine similarity measure[5].

Accordingly, we calculate the partiality of each word. The value of the partiality of each word is calculated in the next formula (3),

$$\mathcal{P}(term_i) = w_1\alpha_3 + w_2\alpha_4 \quad (3)$$

where $term_i$ is ıi-th term in a text, $w_1, w_2 (w_1 > 0, w_2 > 0)$ are weights for skewness($\alpha_3$) and kurtosis($\alpha_4$). $\mathcal{P}$ means the partiality of the word.

Finally, each topic is identified based on the partiality of the word. We identify a topic as a sentence. We calculate a score for each topic using the formula

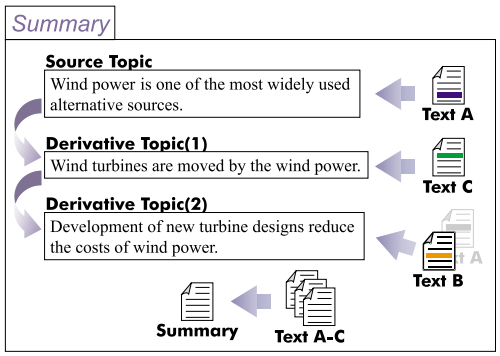


Figure 3: Context formation by linking theme and focus.

(4).

$$Score(S) = \sum_{i=1}^{m} \mathcal{P}(term_i) \quad (4)$$

where $S$ is a sentence in the text. We identify a topic by applying thresholds $(\tau_g, \tau_p(\tau_p > \tau_g))$ to the $Score(S)$. Formula (5) is an evaluation function.

$$S = \begin{cases} general & (Score(S) \leq \tau_g) \\ partial & (Score(S) \geq \tau_p) \end{cases} \quad (5)$$

We identify general and partial topic using $Score(S)$ and thresholds.

## 4 Context-based Summarization

For fascilitating the understandability of the summary, we think a coherence among sentences in the summary is important. In this paper, we form a summary which has a context where each sentence relevants to another sentence. **Context** here is an order of sentences. We form a summary context by linking the relevant topics(sentences).

We use theme and focus of a sentence for linking relevant sentences. **Theme** is a subject of a sentence and **focus** is an information which is emphasized in a sentence. For example, in the following sentence *Quick brown fox jumped over the lazy dog.*, "Quick brown fox" is the theme and "the lazy dog" is the focus.

We identify the theme and focus based on the case of each clause. In the case grammar, each clause in a sentence has a case such as subject and object which indicates a function against a verb. In the previous example, subject noun of the sentence is "fox" and object noun is "dog". We first specify cases of each clause in a sentence and then identify the theme and focus [1].

Context is formed by linking focus in a previous sentence and theme in a current sentence. An idea of context formation is shown in Figure 3. A summary consists of a source topic and derivative topics. **Source topic** is a topic which becomes first sentence in a summary and **derivative topic** is a topic which becomes second to the last sentence in a summary. Derivative topics are always relevant to the previous topic in which the focus of the previous topic relevants to the theme of the current topic.

Summarizing process consists of following process.

1. Specifying a source topic which becomes a first sentence in a summary.
2. Finding a relevant topic to the source topic from texts and linking the relevant topic as a derivative topic.
3. While the number of sentences included in a summary context is below N, repeat linking a topic which is relevant to the previous topic.

Algorithm for the context formation is shown in Table 1. In this algorithm, source topic $(c_i(i = 1))$ is given, and seeks a summary context $(C = \{c_1, c_2, \cdots, c_N\})$. Functions appeared in the algorithm are $\mathcal{F}$ and $\mathcal{S}$. $\mathcal{F}$ returns a focus which becomes a focus of the previous sentence. $\mathcal{S}$ returns a set of topics whose themes are equal to the focus of the previous topic. Algorithm returns the summary context when the number of the topics is equal to $N$.

Table 1: Algorithm for the context formation.

**Given:** $i = 1, c_i$ (source topic)
**Return:** $C = \{c_1, c_2, \cdots, c_N\}$ (summary context)

**step 1** Repeat while $i < N$. Return a context $C$ and exit the loop if $i = N$.
**step 2** Seek a set of focuses $\mathcal{F}(c_i) = \{f_1^i, f_2^i, \cdots, f_m^i\}$ of a sentence $c_i$. Select a focus $f_p$ from a set of focuses $\mathcal{F}(c_i)$.
**step 3**
Seek a set of topics $\mathcal{S}(f_p) = \{s_1^i, s_2^i, \cdots, s_n^i\}$ which take $f_p$ as a theme. Select a derivative topic $s_q^i$ from $\mathcal{S}(f_p)$. Select $s_q^i$ where $min(Score(s_q^i))) \wedge s_q^i \notin C$ for general topic. Select $s_q^i$ where $max(Score(s_q^i))) \wedge s_q^i \notin C$ for partial topic.
$i = i + 1, c_i = s_q^i$. Return to step 1.

## 5 Topic Showcase

We implemented the topic identification method and context-based summarization method to the prototype system called Topic Showcase. Topic Showcase is a multiple-text summarization system which summarizes the results of text retrieval. Screen image of the system is shown in Figure 4 and Figure 5.

[1] We specify cases using case analysis tool: KNP which analize cases in a Japanese sentence. Below is the URL.
http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html


Figure 4: Screen image of Topic Showcase: topic mode.

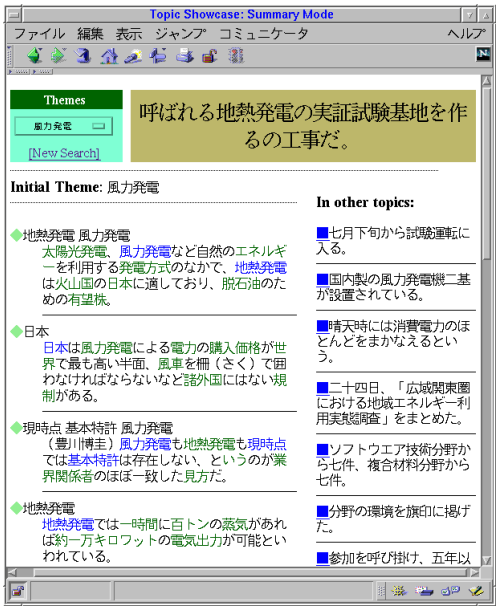


Figure 5: Screen image of Topic Showcase: summary mode.

System retrieves texts from news article[2]. In this paper, we regard news articles as a community's information resource.

Main features of Topic Showcase are followed.

**Retrieval and Clustering**
System can retrieve texts by keywords and classifies the retrieval results into clusters.

**Clustering Configuration**
Users can change the number of the cluster at anytime.

**Dynamic Topic Identification**
System can identify topics dynamically according to the change of the number of clusters.

**Topic Configuration**
Users can select a theme manually and browse theme-related topic.

**Dynamic Summarization**
System can generate summaries dynamically according to the change of the theme of a summary.

## 6 Experiment and Discussion

We evaluate topics and summaries produced by Topic Showcase. As a result, we got affirmative answers in the following evaluations.

- partial topics identified from each cluster
- forecasting the contents from a summary
- total evaluation of Topics Showcase

### 6.1 Description of the Experiment

We evaluated topics and summaries using Topic Showcase. Topics and summaries are produced by the system. Eight test subjects attended the experiment. All test subjects are graduate students of information science.

We use 5 grades as an evaluation measure from *Best* to *Worst*. We use given topics and user-selected summaries for evaluation. Given topics are generated from 28 texts which is retrieved by the keyword "wind generator" in the evaluation of topics. In the evaluation of summaries, we allow users to select summaries freely.

### 6.2 Topics

General evaluation of topics is shown in Figure 6. We got 51.4% of affirmative answers on partial topics. This result indicates the proposed method is available for identifying partial topic.

However, we got 50.0% of negative answers on general topics. We think the cause of the results is the size of texts in which topics are specified. We identified general topic from 28 texts which is retrieved by simple keyword matching. Finding general topic which is common to all texts is difficult because these texts include various topics. To improve the method, evaluating method for consistency of texts is needed.

### 6.3 Summaries

To Facilitate a process of the collective knowledge formation, providing an overview of community's information resource for members of the community is important. In this evaluation, we evaluated two criterion, i.e., (1)possibilities in forecasting the contents of texts from summaries and (2)agreement between forecasting and actual texts. As a result, we got affirmative answers in the question (1). Figure 7 shows the results.

We got 62.5% of affirmative answers in forecasting.

[2]Data of Topic Showcase is Japanese news articles.

This result shows that the proposed method is available in facilitating members to find what kind of information is in the community's information resource.

In the evaluation of the comparson between forecasts and the contents, we got 37.5% of affirmative answers and 25.0% of negative answers. This result shows that test subjects understood the actual texts wrongly from summaries. The main cause of this result is wrong context of summaries. Proposed method simply produces context by linking the theme and the focus. However, there are more relationships among sentences. To improve the problem, analisys on the relationship among sentences and computational context formation method is needed.

### 6.4 General Evaluation of Topic Showcase

In this evaluation, we evaluated the possibility of the Topic Showcase to support user's comprehension of texts. As a result, we got 75.0% of affirmative answers in general evaluation of Topic Showcase. Figure 7 shows the result.

This results show a possibility for facilitating the collective knowledge formation by providing topics and summaries of community's information resource.

To support user's comprehension on the text set, multiple-text summarization system should be developped in regarding the following points.

- information extraction from the texts which include various topics.
- identification of commonalities and differences among texts.
- media mixed summarization

First point is necessary for summarizing the text set including various topic such as retrieval results from WWW. Current summarization system limits the text for preciseness of the summary. However, there is a requirement for summarizing full-text retrieval results which include various topics. Summarizing various texts is one of the problem.

Second point is necessary for comparing among texts precisely. We proposed a multiple-text summarization system for supporting user's comprehension of the text set. However, users of our system have to read original texts for comparing the commonalities and differences among the texts. Identifying commonalities and differences among texts is needed.

Third point is necessary for suporting user's understanding of the texts. Current summarization system only uses text for summary. Adding images and sounds to simple text summary, user's understanding will be much fascilitated.

Multiple-text summarization system is needed for fascilitating the process of the collective knowledge formation. In this paper, we use news article as community's information resource and evaluated the system's

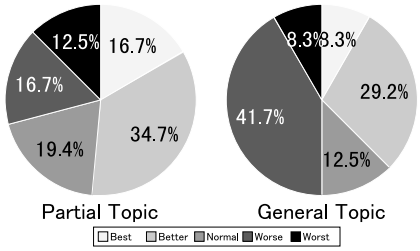


Figure 6: General evaluation for each topic.

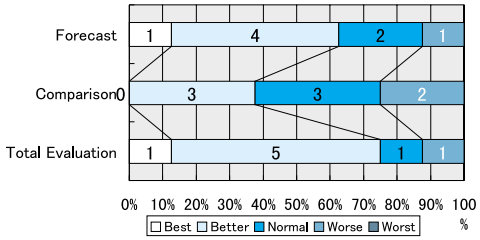


Figure 7: Evaluation for the summary.

possibilities. We'll use this system in the real community and evaluate the possibilities of the multiple-text summarization system.

## 7 Conclusion

We proposed a multiple-text summarization for facilitating the process of the collective knowledge formation. In early stage of the community, there are enormous pieces of information which include unorganized and various topics. To facilitate the process, ordering community's information resource is needed. Our aim is to facilitate the process by providing an overview of community's information resource for members of the community. Experiments showed an availability for ordering unorganized set of texts and a possibility for facilitating community's information resource.

## REFERENCES

[1] Matsubara,S., Ohguro,T. and Hattori,F.: CommunityBoard: Social meeting system able to visualize the structure of discussions; *Proceedings of Knowledge-based Intelligent Electronic Systems(KES'98)*, IEEE, pp.423-428(1998)

[2] Hirata,T, Maeda,H. and Nishida,N: Facilitating community awareness with associative representation; *Proceedings of Second International Conference on Knowledge-Based Intelligent Electronic Systems (KES'98)*, vol. 1, pp.411-416(1998).

[3] K.McKeown, D.R.Radev: Generating Summaries of Multiple News Articles; *Proceedings of ACM-SIGIR'95*, pp.74-82(1995)

[4] Nanba,H., Okumura,M.: Towards Multi-paper Summarization Using Reference Information; *The International Joint Conferences on Artificial Intelligence(IJCAI-99)(to appear)*, (1999).

[5] Salton,G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer; *Addison-Wesley*(1989).