

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 collective knowledge formation process is proposed. Collective knowledge formation in early community is limited by a volume of disordered information. To accelerate a collective knowledge formation, facilitating community members to know an overview of information is needed. We propose a multiple-text summarization method for facilitating community members to know an overview of information in a community. Proposed method consists of topic identification method and context-based summarization method. Topic identification method finds a topic indicating important information in a set of texts. We identify topics based on skewness and kurtosis of a word frequency. Context-based summarization method generates a summary by linking relevant topics. A context is formed based on theme and focus in a sentence. We developed a prototype system called "Topic Showcase". Experimental results show availabilities of the proposed methods in (1)identifying topics from classified texts, (2)facilitating to imagine the contents of texts.

1 Introduction

According to the wide spread of the computer network, a lot of network communities (for short "community") are arising. There are various communities sorted by its size, purpose, member and so on. Java¹, Perl² and Linux³ communities are examples of large communities. In these communities, members discuss on various topic such as software development, bug report, trouble shooting, documentation and so on. Variety of information are exchanged in a community.

One of the features of the community is the collective knowledge formation. The collective knowledge is a knowledge constructed and maintained by members of the community. FAQs, documents and knowledge bases are examples of the collective knowledge. The collective knowledge is formed through a process of in-

formation gathering, discussions and thinking by each member.

We regard a community as information resource consisting of human resources and information resources.

Human resource is a group of community members. Human resource includes an implicit knowledge such as know-hows or experiences owned by each community member who has an expertise in a particular domain.

Information resource is an information repository used for sharing information in a community. An archive of mailing lists or discussions on BBS are examples of information resource. Information resource includes an explicit knowledge such as documents written by community members, discussion logs of mailing lists, databases constructed by community members and so on.

A problem in a collective knowledge formation is a volume of information. In early stage of a community formation process, there are many discussions on various topic. These discussions are stored in a mailing list or BBS server, but are not organized. The stored information includes a large volume of disordered pieces of information. Even if one can retrieve archives of discussions, it is difficult to find an overview of information resource in a community. Retrieval result doesn't facilitate to know each topic discussed in discussions, relations between topics and an overview of information resource in a community. Consequently, a process of collective knowledge formation is delayed.

To accelerate a collective knowledge formation process, facilitating community members to know an overview of information resource is needed. By Summarizing information resource in a community, community members can know what kind of information has been discussed or shared in information resource in a community.

We propose a multiple-text summarization method for facilitating to know an overview of information resource in a community. We regard information resource in a community as a set of texts. Proposed method consists of two sub-methods, i.e., (1)topic identification method and (2)context-based summarization method. Topic identification method identifies topics

¹<http://www.javasoft.com/>

²<http://www.perl.com/>

³<http://www.linux.com/>

which indicate topical sentences in texts. Context-based summarization method generates a summary which has a relation between sentences.

In the remain of this paper, we first analyze a process of the collective knowledge formation and show an overview of multiple-text summarization method. We describe the proposed methods, i.e., topic identification method and context-based summarization method. We show a prototype system for multiple-text summarization called Topic Showcase. We finally show the results of an evaluation and discuss the results.

2 Multiple-text Summarization for Collective Knowledge Formation

We analyze a process of a collective knowledge formation. Then we describe the proposed multiple-text summarizing method.

2.1 Process of the Collective Knowledge Formation

We assume that a process of collective knowledge formation consists of the following three steps.

Information Collecting Step

In this step, each community member collects information. Some members collect information from network resource and others describe their own knowledge into documents. This process is an individual process.

Discussion and Filtering Step

Community members discuss on a common topic based on collected information. A part of collected information is utilized for discussion. Other information are discarded. This process feeds back to the previous step. Community members collect information again according to the result of discussions.

Knowledge Formation Step

Collective knowledge is formed based on the result of discussion. Collective knowledge is described as FAQs, WWW pages, knowledge bases and so on.

An issue in the collective knowledge formation process is a difficulty in knowing an overview of information resource. In early stage of community, collected information and discussions are not organized. Thus, it is difficult to know what kinds of topics are discussed in a community for community members. Accelerating the process of the community knowledge formation, facilitating to know an overview of information resource is needed.

Related works of ordering the community’s information is proposed. Matsubara proposed Community-Board 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.

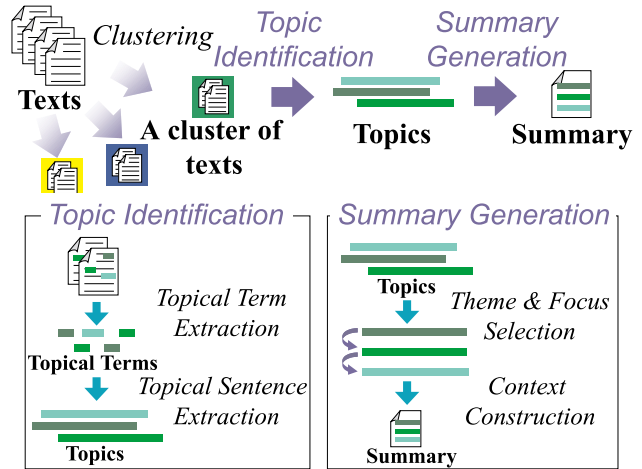


Figure 1: Overview of multiple-text summarization.

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 a collective knowledge formation, but don’t organize information directly. Organizing information such as classifying, extracting or summarizing information is needed for accelerating a process of collective knowledge formation.

We propose a text summarization method. Our aim in text summarization is to assist members of getting an overview of information resource in a community. Community’s information resource in early stage of the community is regarded as a disordered information pool. By summarizing disordered information pool, community members can find out what kind of topics are discussed in a community.

2.2 Multiple-text Summarization

We propose a multiple-text summarization method for accelerating a collective knowledge formation process. We regard information resource in a community as multiple-text which include various kinds of topics. By Summarizing information resource, a process of a 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 refer-

ence of citation among papers. Identifying a purpose of citation, proposed method generates a summary according to the purpose of reference. These methods can produce a detailed summary which considers the contents and relationships among texts. However, these methods have limitation on texts which the methods summarize.

To summarize information resource in a community, domain independent summarization method is needed. There are various kinds of texts in a early stage of community. These texts include various types such as news articles, scientific papers, WWW pages, e-mails and so on. To summarize these texts, domain independent summarization method is needed. We propose a multiple-text summarization method being domain independent. We use statistical information of words and identifies topics.

Figure 1 shows our approach. Proposed method consists of two 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.

We describe each sub-method in the following sections.

3 Topic Identification based on Statistical Information

We identify topics based on classification and statistical information of texts. We firstly describe that skewness and kurtosis which indicate the partiality of topics over the entire texts. Then, we show our proposed method.

3.1 Skewness and Kurtosis

We use skewness and kurtosis for identifying topics. **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 definition of each 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, \dots, x_n are word frequency in i -th text, \bar{x} is an average and s is a standard deviation.

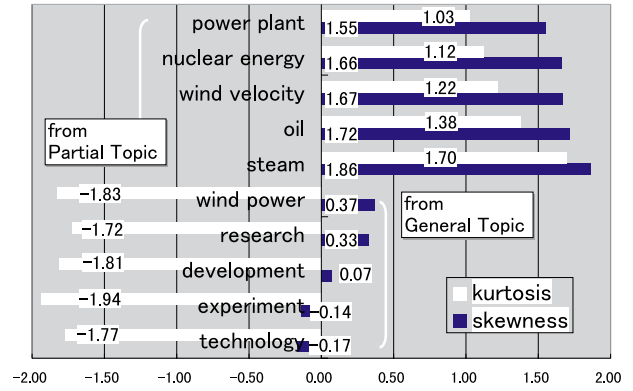


Figure 2: Comparison of skewness and kurtosis among words within a general or partial topic.

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 similar 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 i -th term in a text, w_1, w_2 ($w_1 > 0, w_2 > 0$) are weights for skewness(α_3) and kurtosis(α_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 (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 (τ_g, τ_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 facilitating 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 relevant 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 focus⁴.

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 relevant to the theme of the current topic.

Summarizing process consists of following process.

Summary

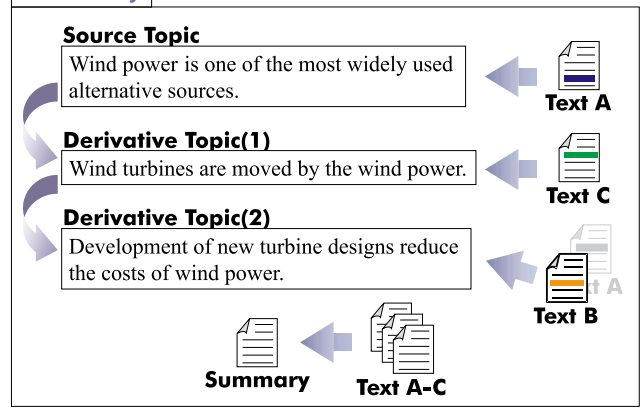


Figure 3: Context formation by linking theme and focus.

Table 1: Algorithm for the context formation.

| | |
|----------------|---|
| Given: | $i = 1, c_i$ (source topic) |
| Return: | $C = \{c_1, c_2, \dots, c_N\}$ (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 $\mathcal{F}(c_i) = \{f_1^i, f_2^i, \dots, f_m^i\}$ 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 $\mathcal{S}(f_p) = \{s_1^i, s_2^i, \dots, s_n^i\}$ 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. |

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 ($C = \{c_1, c_2, \dots, c_N\}$). 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 .

⁴We specify cases using case analysis tool: KNP which analyze 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.

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.

System retrieves texts from news article⁵. 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 any time.

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.

⁵Data of Topic Showcase is Japanese news articles.



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

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. 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 comparison 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, analysis 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 developed 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

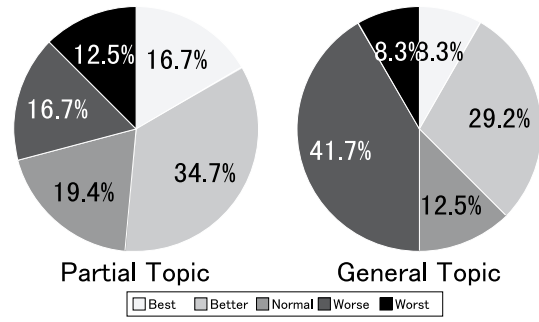


Figure 6: General evaluation for each topic.

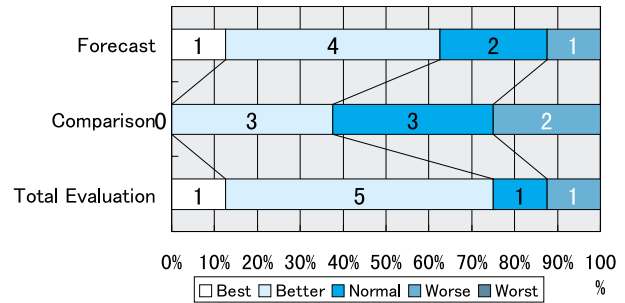


Figure 7: Evaluation for the summary.

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 supporting 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 facilitated.

Multiple-text summarization system is needed for facilitating the process of the collective knowledge formation. In this paper, we use news article as community’s information resource and evaluated the system’s 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 a process of a collective knowledge formation. In early stage of the community, information resource in a community is disordered. Summarizing informa-

tion resource in a community is needed for facilitating a collective knowledge formation process. We proposed a topic identification method and context-based summarization method. We found an availability of the proposed method for knowing topics of each categories and facilitating to imagine the contents of texts from a summary. In future work, evaluation in a real community is needed.

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).