Okazaki et al. (2018) have proposed a method for organizing the information contained in multiple documents into a table without limiting the information to be extracted. In this study, we propose a method for improving the accuracy of these tables. In our proposed method, information is first clustered hierarchically. Next, for the results of hierarchical clustering (with the number of clusters ranging from 1 to $n$), the degree of filling and the information density of the resulting table are calculated. The number of clusters when the balance between these two indicators is optimal is chosen as the optimal number of clusters. The results of the method using the chosen number of clusters are organized into a table. In the conventional method, the number of clusters estimated by the $X$-means method tends to be too small. As demonstrated by the results of experiments using 15 types of multiple documents, the proposed method improves this problem, with its estimated number of clusters being closer to the optimum. The average evaluation result in the tables (F-measure) when applying the conventional method was 0.43; the proposed method improves this to 0.65. We therefore confirm the effectiveness of the proposed method.

**Key Words:** Information Extraction, Table, X-means, Hierarchical Clustering

1 Introduction

In recent years, it has become significantly easier to obtain information from the vast amount of electronic documents stored on the Web. Okazaki et al. (2018) proposed a method for organizing the information contained in multiple documents into a table without limiting the information to be extracted. Here, “multiple documents” refers to a collection of documents of the same type. For example, information such as “maker” and “price” included in new product articles for smartphones of different models is organized in a table based on type. Arranging the information of multiple documents of the same type in a table in this way is useful not only for extracting information, but also for comparing information between documents. In Okazaki et al.’s study, information contained in multiple documents of the same type was extracted in sentence units,
and the results of classifying the information by the clustering method of the $X$-means method (Pelleg and Moore 2000) were organized into a table with documents as rows and clusters as columns. In this paper, we call this method the “conventional method.” The $X$-means method is a clustering method that automatically estimates the optimum number of clusters based on an index called BIC while repeating the clustering using the $K$-means method at $K = 2$. It is not necessary to specify the number of classification destinations. However, the number of clusters estimated by the $X$-means method tends to be smaller than the optimal number of clusters; in the resulting tables, too much information is organized into a single column, and the accuracy of the tables is low.

To address this problem, we propose a method for estimating the number of clusters by optimizing the balance between the degree of filling and the information density of a table. The degree of filling of a table indicates that there are fewer blanks in the table. The degree of information density of a table indicates that the information in the same column of the table is more similar. With the proposed method, it is possible to create a more appropriate table where similar contents are arranged in the same column of a table with fewer blanks by using the degree of filling and the information density of a table. Blank cells are generated when creating a table. The degree of filling of a table is useful for reducing the number of blank cells.

With the proposed method, information is first clustered hierarchically. The results of hierarchical clustering, with the number of clusters ranging from 1 to $n$, are organized in $n$ tables. Then, for each table, the degree of filling and the degree of organized information density are calculated. When the balance between these two indicators is optimal, the resulting number of clusters is estimated to be optimal. The results, with the optimal number of clusters estimated at the end, are organized in a table. In this study, we attempt to improve the accuracy of the tables of the conventional method using the above approach. This study was completed in Japanese, and Japanese text was used for the experiments.

In this paper, we propose a method for estimating the optimum number of clusters based on the degree of filling and density in a table from the results of hierarchical clustering. Several methods (a silhouette analysis, Upper Tail method, and BIC-based method) have already been proposed to estimate the optimal number of clusters.

Our proposed method estimates the number of clusters by optimizing the balance between the degree of filling and the information density of a table. In addition, the degree of filling of a table indicates that there are fewer blanks in the table. The degree of information density of a table indicates that the information in the same column of the table is more similar. With the proposed method, it is possible to create a more appropriate table in which similar contents are
arranged in the same column of the table with fewer blanks by using the degree of filling and the information density of the table.

The other methods, i.e., the X-means method, silhouette analysis, Upper Tail method, and BIC-based method use an index similar to the degree of the information density, but do not use the degree of filling of the table. However, the proposed method does have the novelty of using such degree of filling.

In the experiments described in this paper, we confirmed that the proposed method significantly outperformed other methods. The novelty of method using the degree of filling of a table seems to be the reason why the proposed method performed better than the other approaches.

This paper is related to information extraction (Finkel et al. 2005; Chang et al. 2006; Gonzalez and Turmo 2009; Fader et al. 2011; Fukuda et al. 2012; Grishman 2015; Ali et al. 2017; Stanovsky et al. 2018); a brief description of related studies and their relevance to the present work is as follows. Fukuda et al. (2012) constructed a system that extracts expressions indicating the effects of technology from research papers and visualized the extracted information. They created supervised data and used machine learning to extract the expressions. By contrast, Okazaki et al.’s study (Okazaki et al. 2018) and the present approach use clustering instead of machine learning; there is therefore no need to create learning data.

We additionally examined several studies related to tables (Swain and Cole 2016; Nishida et al. 2017). Swain et al. constructed a toolkit for an automated extraction of chemical information from tables in the scientific literature. Nishida et al. also constructed a method of understanding the semantic structures of tables. By contrast, Okazaki et al.’s study (Okazaki et al. 2018) and our own extract important information from sentences in documents and construct tables including extracted information.

Akano et al. (2017) studied tables and information extraction. In addition, Akano et al. utilized word2vec (Mikolov et al. 2013) to represent the words that appear on Wikipedia pages related to a castle as vectors and then clustered and organized them in the form of a table. In word2vec, word vectors are obtained by considering the surrounding words, and word vectors vary depending on the words that appear in the surroundings. For example, the words related to a person can be classified based on the role that they have in the castle. However, only words are organized in the table in Akano et al.’s study and there is no information about the role in the table. In Akano et al.’s method, information was extracted word-by-word, for example, “Mouri Motonari”; however, in Okazaki et al.’s study (Okazaki et al. 2018) and our own approach, information is extracted as a sentence unit, e.g., “Mouri Motonari became a castle owner.” In Okazaki et al.’s study (Okazaki et al. 2018) and our own method, we can see the role of a person.
in the sentences shown in the table.

Section 2 describes the conventional approach using the $X$-means method. Section 3 describes the proposed method using hierarchical clustering. Section 4 describes experiments confirming the effectiveness of the proposed method. Section 4 includes an explanation of comparison methods (silhouette analysis, Upper Tail method, and BIC-based method).

2 Conventional Method

In Okazaki et al.’s study (Okazaki et al. 2018), information contained in multiple documents of the same type was extracted in sentence units. The resulting information was clustered using a method called the $X$-means method (Pelleg and Moore 2000) and organized in a table.

In their study, the system creates a table with sentences that describe the same type of things in the same column. The clustering of sentences is used to collect sentences that describe the same type of things.

In the study, FastText (Bojanowski et al. 2016) is used to grasp similar contents of different words. Things with similar content may be written using different words. FastText allows us to vectorize words and make similar words into similar vectors. FastText is used such that we can grasp similar content of different words when we check the similarity of sentences containing different words.

2.1 Explanation of the Conventional Method

2.1.1 Procedure for organizing information in a table

The procedure used by Okazaki et al. for organizing sentence information in a table is shown below.

Step 1 The system divides multiple documents into sentences.

Step 2 The system calculates the sentence vector of each sentence obtained in Step 1.

Step 3 The system clusters the sentence vectors using the $X$-means method (Pelleg and Moore 2000).

Step 4 The results of clustering are organized into a table with documents as rows and clusters as columns. During clustering, which document contained the information is not taken into consideration, and thus one cell may contain multiple pieces of information.

Step 5 The system gives a name to each cluster (column) in the table.
2.1.2 Step 2: Calculation of statement vector

The vector of each sentence in Step 2 of Section 2.1.1 is obtained by the following procedure.

1. The system conducts a morphological analysis of the sentences using MeCab.
2. It extracts words whose part-of-speech is a noun and whose detailed part-of-speech classification does not contain a pronoun, number, non-independent word, or adverb among the morphological analysis results.
3. The sum of the vectors of the extracted words is treated as the sentence vector. Word vectors are obtained by learning all articles on Wikipedia by FastText (Bojanowski et al. 2016).

Things with similar contents may be written in different words. FastText is used to express the similarity of sentences and thus even such things can be grasped as similar contents.

2.1.3 Step 3: Clustering of sentences

In Step 3 of Section 2.1.1, sentences are clustered using the $X$-means method (Pelleg and Moore 2000). The $X$-means method repeats the division by the $K$-means method at $K = 2$ and determines whether to stop the division based on the Bayesian information $BIC$. In this clustering method, the number of clusters is determined automatically. If $BIC \leq BIC'$ is true for the Bayesian information $BIC$ before the division and the Bayesian information $BIC'$ after the division, the division is stopped.

Assuming that the $p$ multivariate normal distribution is given the following equation, $BIC$ is defined as follows.

$$f(\theta_i; x) = (2\pi)^{-n/2}|V_i|^{-1/2} \exp \left[ \frac{1}{2}(x - \mu_i)^t V_i^{-1}(x - \mu_i) \right]$$

Here, $\hat{\theta}_i = [\hat{\mu}_i, \hat{V}_i]$ is the maximum likelihood estimate of the $p$ multivariate normal distribution, $\mu_i$ is the $p$th mean vector, and $V_i$ is the variance-covariance matrix of $p \times p$. In addition, $q$ is the number of dimensions in the parameter space, which is $q = 2p$ if the covariance of $V_i$ is ignored, and $q = p(p + 3)/2$ if it is not. Moreover, $L$ is a likelihood function ($L = \Pi f()$).

$$BIC = -2 \log(L\hat{\theta}_i; x_i \in C_i) + q \log n_i$$

In addition, $BIC'$ (after division) is defined as follows. Here, $\hat{\theta}_i' = [\hat{\theta}_1^i, \hat{\theta}_2^i]$ is an estimated value of the maximum likelihood of the $p$ multivariate normal distribution in each of the two split clusters.

---

1 http://taku910.github.io/mecab/
If we ignore the covariance, there are two parameters, mean and variance, for each $p$, and thus $q' = 2 \times 2p = 4p$; otherwise, $q' = 2q = p(p + 3)$.

$$BIC' = -2 \log(\hat{L}_{\theta_i}; x_i \in C_i) + q' \log n_i$$

### 2.1.4 Step 4: Organizing clustering results in a table

The clustering results of Step 3 in Section 2.1.1 are organized into a table with documents as rows and clusters as columns. To improve the readability of the table, the columns are sorted in order of importance.

In the columns of the table that organize the clustering results, some columns are of high importance (such as Column 1 of Table 1, which consists only of related sentences), and some are of low importance (such as Column 2, which includes unrelated sentences). Here, the importance of a column is defined based on the density and document coverage.

The density $d_k$ in the $k$th column is defined by Equation 1. Here, $N_k$ is the total number of sentences contained in the $k$th column, and $S_{k,l}$ is the vector of the $l$th sentence contained in the $k$th column. $S_{k, mean}$ is the average of the vectors of the sentences contained in the $k$th column.

$$d_k = \frac{1}{N_k} \sum_{l=1}^{N} \frac{S_{k,l} \cdot S_{k, mean}}{|S_{k,l}| |S_{k, mean}|} \quad (1)$$

The system normalizes the column density $d_k$ obtained by Equation 1 such that the minimum value is zero and the maximum value is 1 using Equation 2. Here, $nd_k$ is the normalized column density of the $k$th column, and $K$ is the total number of columns.

$$nd_k = \frac{d_k - d_{min}}{d_{max} - d_{min}} \quad (2)$$

$$d_{min} = \min_{1 \leq k \leq K} d_k \quad (3)$$

$$d_{max} = \max_{1 \leq k \leq K} d_k \quad (4)$$

| Document 1 | The weight is about 130g. | The price is about 130g. |
|------------|--------------------------|--------------------------|
| Document 2 | The weight is 125g.      | The thinness is 8mm.     |
| Document 3 | The weight is about 100g and light. | The screen size is 5 inches. |
| Document 4 | The weight is 130g.      | The price is undecided.  |
| Document 5 | The weight is about 150g. | Scheduled to be released in January. |

**Table 1**  Example of importance in a column
The document coverage rate $c_k$ in the $k$th column is defined by Equation 5. In addition, $p_k$ is the number of documents for which sentences could be extracted in the $k$th column, and $P$ is the total number of documents.

$$c_k = \frac{p_k}{P} \quad (5)$$

The document coverage $c_k$ obtained by Equation 5 is normalized by using Equation 6 such that the minimum value is zero and the maximum value is 1. Here, $nc_k$ is the normalized document coverage of the $k$th column and $K$ is the total number of columns.

$$nc_k = \frac{c_k - c_{\text{min}}}{c_{\text{max}} - c_{\text{min}}} \quad (6)$$

$$c_{\text{min}} = \min_{1 \leq k \leq K} c_k \quad (7)$$

$$c_{\text{max}} = \max_{1 \leq k \leq K} c_k \quad (8)$$

We define the importance $i_k$ of the $k$th column using Equation 9.

$$i_k = nd_k \times nc_k \quad (9)$$

### 2.1.5 Step 5: A methodology for finding the item name of a cluster

In Step 5 of Section 2.1.1, item names are given to each column through the following procedure.

1. For each sentence included in the cluster, all nouns are extracted.
2. The document frequency is found for each word extracted in 1.
3. The word with the highest document frequency is chosen as the item name of the cluster.
4. If the highest document frequency is shared by two or more words, all are used as the item name of the cluster (separated by commas).

### 2.2 Problems with conventional method

Using the conventional method, it is often the case that either too much information is organized in one column, as in Column 1 of Table 2, or the information is classified too finely, as in Columns 2 and 3 of Table 3. A well-balanced table, such as Table 4, is desirable. In addition, the X-means method tends to underestimate the optimum number of clusters (that is, the optimum number of columns in the resulting table). This often results in a table with too much information in one column, such as Column 1 of Table 2, where memory and storage information are mixed. This was the cause of the low accuracy.
The memory is 4GB.
The internal storage is 64GB.

The release date is early October.

Table 2  Example of too much information in one column

| Column 1 | Column 2 | Column 3 | Column 4 |
|----------|----------|----------|----------|
| The memory is 4 GB. | The internal storage is 64 GB. | The release date is early October. |
| The memory is 3 GB. | The storage is 32 GB. | The release date is September. |
| The memory is 3 GB. | The internal storage is 32 GB. | The release date is January. |
| The memory is 3 GB. | The internal storage is 32 GB. | Released from August. |

Table 3  Example of information being classified too finely

| Column 1 | Column 2 | Column 3 | Column 4 |
|----------|----------|----------|----------|
| The memory is 4 GB. | The internal storage is 64 GB. | The release date is early October. |
| The memory is 3 GB. | The storage is 32 GB. | The release date is September. |
| The memory is 3 GB. | The internal storage is 32 GB. | The release date is January. |
| The memory is 3 GB. | The internal storage is 32 GB. | Released from August. |

Table 4  Example of an optimal table

3 Proposed method

To solve the problems of the conventional method, we propose a method for estimating the number of clusters by optimizing the balance between the filling degree and the density of information in the table. In the $X$-means method used in the conventional method, the table obtained in each execution is different owing to the initial value dependency. The proposed method uses a hierarchical clustering method that does not depend on the initial values.

With the proposed method, a table with the maximum score is selected after creating a corresponding individual table for each different $k$ (the number of clusters). The hierarchical clustering method is used to create individual tables.

3.1 Procedure used with our proposed method

The procedure of our proposed method is shown below.

809
Step 1 The system divides multiple documents into sentences.
Step 2 The system calculates the sentence vector of each sentence.
Step 3 The system clusters the sentence vectors using a hierarchical clustering method based on Ward’s approach.
Step 4 The results of clustering are organized into a table with documents as rows and clusters as columns. The filling degree of a table with k clusters is calculated using Equation 10, and the density of information in the table is calculated from Equation 11. In addition, \( |c_{k,i}| \) is the total number of sentences contained in the ith column of the table with k clusters, \( d_{k,i,j} \) is the vector of the jth sentence in the ith column of the table with k clusters, and \( C_k \) is the total number of columns in the table with k clusters, and \( \text{cosine}(x, y) \) is the function for finding the cosine similarity of x and y.

\[
cover_k = \frac{\text{The number of non-empty cells in the table with } k \text{ clusters}}{\text{The total number of cells in the table with } k \text{ clusters}}
\]

(10)

\[
density_k = \min_{j \neq h}(\text{cosine}(d_{k,i,j}, d_{k,i,h}))
\]

(11)

\[
i = 1, \ldots, C_k \quad j, h = 1, \ldots, |c_{k,i}|
\]

Here, \( \text{COVER} \) is the set of \( \text{cover}_k \) for all clusters, \( \max(\text{COVER}) \) is the maximum value of the set \( \text{COVER} \), and \( \min(\text{COVER}) \) is the minimum value of the set \( \text{COVER} \). The system normalizes \( \text{cover}_k \) in each cluster to the range of 0 to 1 using Equation 12.

\[
\text{norm}(\text{cover}_k) = \frac{\text{cover}_k - \min(\text{COVER})}{\max(\text{COVER}) - \min(\text{COVER})}
\]

(12)

Similarly, \( \text{DENSITY} \) is the set of \( \text{density}_k \) for all clusters, \( \max(\text{DENSITY}) \) is the maximum value of \( \text{DENSITY} \), and \( \min(\text{DENSITY}) \) is the minimum value of \( \text{DENSITY} \). The system normalizes \( \text{density}_k \) in each cluster to the range of 0 to 1 using Equation 13.

\[
\text{norm}(\text{density}_k) = \frac{\text{density}_k - \min(\text{DENSITY})}{\max(\text{DENSITY}) - \min(\text{DENSITY})}
\]

(13)

The system finds \( \text{Score}_k \) in the table with \( k \) clusters using Equation 14. The number of clusters \( k \) when \( \text{Score}_k \) is maximized is adopted as the optimum number of clusters.

\[
\text{Score}_k = \text{norm}(\text{cover}_k) \times \text{norm}(\text{density}_k)
\]

(14)

Step 5 The system gives item names to each cluster in the table.
3.2 Hierarchical clustering

Hierarchical clustering is a clustering method that repeats the integration of clusters with the shortest distance. There are several methods for hierarchical clustering depending on the definition of the distance between clusters. In this study, we used Ward’s method. In Ward’s method, the distance $D(C_1, C_2)$ between Cluster $C_1$ and Cluster $C_2$ is defined as follows.

$$D(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2)$$

$$E(C_i) = \sum_{x \in C_i} (d(x, c_i))^2$$

$$c_i = \sum_{x \in C_i} x/|C_i|$$

4 Experiment

In this section, we evaluate and compare tables generated using the conventional method and tables generated using the proposed method with the same input data. In the X-means method used in the conventional approach, a different table can be obtained from each execution, depending on the initial value. For our experiments, a table was generated 1,000 times; among the evaluation results of all tables applied, the maximum, average, and minimum value results were used as comparison targets.

In this section, we also describe comparative experiments between the proposed method and methods differing from the conventional method.

4.1 Experimental environment

4.1.1 FastText learning environment

The word vector used in the calculation of the sentence vector is obtained using FastText (Bojanowski et al. 2016). FastText is a two-layer neural network consisting of a hidden layer and an output layer; the hidden layer corresponds to the distributed representation of words. All 1,061,375 articles (22,794,659 lines) in Japanese Wikipedia (June 1, 2017)² were used as learning data for FastText. The learning data were unified in full-width for Japanese characters and numbers and half-width for Latin characters and numbers, and were divided into morpheme units with MeCab. Skip-gram was used as the learning model in FastText, with 300 vector dimensions. The default values were used for all other parameters.

² https://ja.wikipedia.org/. The data were crawled on “June 1, 2017.”
4.1.2 Experimental data

Fifteen types of multiple documents were extracted from newspaper articles, Web product article pages, and Wikipedia for use in the experiment. The details of each type are shown in Table 5.

For newspaper articles, 20 articles containing the corresponding word (e.g., “robbery”) in the headline were randomly extracted from 2016 Mainichi Shinbun newspaper articles. For Web product article pages, the latest 20 new product news articles in the corresponding category (e.g., “smartphone”) were extracted from the web pages of Kakaku.com (January 15, 2018). For Wikipedia, the summary sections of 20 randomly selected pages included in a corresponding Wikipedia category (e.g., “100 Famous Castles in Japan”) were extracted from Japanese Wikipedia (June 1, 2017). The same Wikipedia (June 1, 2017) was used for input documents and the learning data for FastText.

4.1.3 Evaluation method

The accuracy of generated tables is evaluated using the following procedure.

(1) We focus on each column of a correct table, which is manually created in advance.

| Type of Document | Number of documents | Total number of sentences | Average number of characters per sentence |
|------------------|---------------------|---------------------------|------------------------------------------|
| Newspaper article (robbery) | 20 | 128 | 39.3 |
| Newspaper article (finance) | 20 | 124 | 49.2 |
| Newspaper article (earthquake) | 20 | 91 | 37.2 |
| Newspaper article (traffic accident) | 20 | 143 | 41.7 |
| Newspaper article (recall) | 20 | 89 | 56.7 |
| New product article (smartphone) | 20 | 313 | 46.3 |
| New product article (TV) | 20 | 273 | 49.0 |
| New product article (camera) | 20 | 340 | 52.0 |
| New product article (robot vacuum cleaner) | 20 | 235 | 47.7 |
| New product article (air conditioner) | 20 | 255 | 62.3 |
| Wikipedia (castle) | 20 | 94 | 31.2 |
| Wikipedia (dinosaur) | 20 | 77 | 49.9 |
| Wikipedia (wrestler) | 20 | 103 | 28.7 |
| Wikipedia (mountain) | 20 | 76 | 31.5 |
| Wikipedia (baseball team) | 20 | 68 | 46.9 |

Table 5 Details of each type of multiple documents used in the experiment

https://kakaku.com/. The data were crawled on “January 15, 2018.”
(2) We extract those columns of an experimental table containing the most data in the column of interest. The experimental table is the output of the proposed (or other) method.

(3) We find the precision, recall, and $F$-measure using Equations 15 through 17.

(4) We perform 2 and 3 for all of the correct columns, find the average of the $F$-measures in all columns, and use this as the evaluation result of the experimental table.

Precision = \frac{\text{The number of sentences commonly included in the correct table column and outputted table column}}{\text{The number of sentences included in the outputted table column}} \tag{15}

Recall = \frac{\text{The number of sentences commonly included in the correct table column and outputted table column}}{\text{The number of sentences contained in the correct table column}} \tag{16}

F-measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}

Correct tables are created manually without using the output tables of the system. A correct table is manually created by manually checking 20 input articles, collecting sentences that have the same main content, and putting them in the same column.

Here, an example of calculating the F-measure of a column is described. We suppose the following condition: We use ten documents from Wikipedia (wrestler) and are focusing on the fifth column of the correct table. The column of the outputted table containing the most sentences in the fifth column of the correct table is the fourth column. The fifth column of the correct table is shown in Table 6. The fourth column of the output table is shown in Table 7. In this case, the precision, recall, and F-measure are calculated as follows.

Precision = \frac{5}{5} \tag{18}

Recall = \frac{5}{8} \tag{19}

F-measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times \frac{5}{5} \times \frac{5}{8}}{\frac{5}{5} + \frac{5}{8}} = 0.77 \tag{20}

With respect to the precision, there are five rows in Table 7 (output table), which is the number of sentences included in the output table column, and they all appear in Table 6 (correct table); thus, the precision is 5/5. With respect to the recall, there are eight rows in Table 6 (correct table), which is the number of sentences contained in the correct table column, and five of them appear in Table 7 (output table); thus, the recall is 5/8.
Table 6 The fifth column of a correct table (Wikipedia (wrestler))

4.2 Comparison methods

In this paper, we propose a method for estimating the optimum number of clusters based on the degree of filling and density in a table from the results of hierarchical clustering. Several indicators have already been proposed to estimate the optimal number of clusters. Therefore, experiments comparing our proposed methods with other methods using the indicators were conducted by applying the following three cluster number estimation methods in Step 4 of the proposed method.

4.2.1 Comparison method 1: Silhouette analysis

A silhouette analysis (Rousseeuw 1987) is an index based on the cohesiveness of data in clusters and the discreteness between clusters. In a silhouette analysis, the silhouette coefficients of all data are calculated for each number of clusters, and the number of clusters when the average value of the silhouette coefficients is maximized is selected.

The silhouette coefficient $s(i)$ of the datapoint $i$ is expressed by Equation 21. Here, $a(i)$ is the average distance between datapoint $i$ and each datapoint contained in the cluster to which point
### Table 7 The fourth column of an output table (Wikipedia (wrestler))

| Document | 4th column (skill) |
|----------|--------------------|
| Document 1 |                    |
| Document 2 | *tokuwaza wa migiyotsu, yori.*  
(His special skill is right-handed belt gripping and forcing out.) |
| Document 3 |                    |
| Document 4 |                    |
| Document 5 |                    |
| Document 6 | *tokuwaza wa hidariyotsu, yori.*  
(His special skill is left-handed belt gripping and forcing out.) |
| Document 7 | *tokuwaza wa migiyotsu, tsuppari, hatakikomi.*  
(His special skill is right-handed belt gripping, thrusting, and slapping down.) |
| Document 8 |                    |
| Document 9 | *tokuwaza wa oshi, hidariyotsu, yori.*  
(His special skill is pushing out, left-handed belt gripping, and forcing out.) |
| Document 10 | *tokuwaza wa tsuppari, migiyotsu, uwatenage.*  
(His special skill is thrusting, right-handed belt gripping, and overarm throw.) |

$i$ belongs. In addition, $b(i)$ represents the average distance between the datapoint $i$ and each point contained in the cluster closest to $i$ among the clusters to which $i$ does not belong. Here, the cluster closest to $i$ is the cluster for which the average distance between $i$ and each datapoint contained in the cluster is the smallest. In addition, $\max\{a(i), b(i)\}$ represents the larger value of $a(i)$ and $b(i)$.

\[ s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \]  

4.2.2 Comparison method 2: Upper tail method

The upper tail method was proposed by Mojena (1977) to determine the optimum number of clusters based on statistical stop rules for the results of hierarchical clustering. In the upper tail method, first, the reference value $\alpha_j$ is obtained for each clustering result when the number of clusters $j$ is 2 to $N - 1$. Here, $\alpha_j$ represents the minimum distance between the centroids of clusters in the clustering result with number of clusters $j$. Next, the system starts with the value
of \( j \) from 2, and increases the value of \( j \) by 1 until the following conditions are no longer satisfied.

\[
\alpha_j \leq \bar{\alpha} + ks_{\alpha}
\]

The value of \( j \) when stopped is selected as the optimum number of clusters. Here, \( \bar{\alpha} \) and \( s_{\alpha} \) represent the mean of all reference values \( \alpha_j \) and the square root of the unbiased variance of all reference values \( \alpha_j \). Regarding the value of \( k \), Mojena (1977) uses a value of 2 to 4 when the number of data is 60 to 120. In addition, Ayaka and Matsuda (2011) reported that \( k = 3 \) is a good value when the number of data in one group is approximately 30 to 50. With reference to these, in this study, assuming that the number of data in one group is less than the number of documents (20 or less), the experiment is conducted using \( k = 1 \).

4.2.3 Comparison method 3: BIC-based method

The same method as the X-means method in Section 2.1.3 is applied to the results of hierarchical clustering, and the optimum number of clusters is estimated. Specifically, every time the tree diagram obtained as a result of hierarchical clustering is divided into two parts in order from the root of the tree, the system finds \( BIC \) before division and \( BIC' \) after division in the same way as described in Section 2.1.3, and if \( BIC \leq BIC' \), the system stops splitting.

4.3 Experimental results

An example of a part of a table generated by the proposed method is shown in Table 8.

The first column of the table shows the height and weight of sumo wrestlers. The second column of the table shows the sumo ratings such as Makushita and Front. The third column of the table shows the hometown of the wrestlers. In the table, the contents related to sumo wrestlers are acquired. In the table, each column contains useful information with the same content with a few blanks. We were able to generate a table that is useful for gathering information.

We show the evaluation results using the conventional method, the proposed method, and comparison methods (silhouette analysis, upper tail method, and BIC-based method) in Table 9. We also conducted a paired two-sided \( t \) test to determine the significance of the difference in the evaluation results. The significance level was 0.05. The results of the significance test are shown in Table 10.

In an experiment using 15 types of multiple documents, the average evaluation result of the proposed method (F-measure) was 0.65, which exceeded that of the other methods. The \( t \) test confirmed that our proposed method significantly outperformed the other methods.
| Document | 1st column (height) | 2nd column (highest) | 3rd column (origin) |
|----------|---------------------|----------------------|---------------------|
| Document 1 | shincho 183 cm, taiju 191 kg | saikoi wa higashi juryo 12 maime | masakaze wa nagasaki-ken nagasaki-shi shusshin de ogurumabeya shozoku no geneki ozumo rikishi (Mototsugu Masakaze is an active sumo wrestler from Nagasaki City, Nagasaki Prefecture, who belongs to the Oguruma stable.) |
| Document 2 | shincho 175 cm, taiju 130 kg, ketsueki gata wa O gata | saikoi wa nishi makushita 2 maime | tochino yamahirushi wa, tokyo-to tatekawa-shi shusshin de, chiganoura beya no moto ozumo rikishi de, gen sewa min (Hiroshi Tochinoyama is from Tachikawa, Tokyo, and is a former grand sumo wrestler in the Chiganoura stable. He is currently Caretaker.) |
| Document 3 | geneki jidai no taikaku wa shincho 179 cm, taiju 149 kg, ketsueki gata wa AB gata | saikoi wa higashi makushita 4 maime | Takamitsu Tochinosato wa, ishikawa-ken ishikawa-gun nonoichi-cho shusshin de kasugano beya ni shozoku shiteita moto ozumo rikishi (Takamitsu Tochinosa is from Nonoichi Town, Ishikawa County, Ishikawa Prefecture, belongs to the Kasugano stable, and is a former grand sumo wrestler.) |
| Document 4 | shicho wa 180 cm, taiju wa 160 kg | saikoi wa nishi mae gashira 6 maime | homare yuki yoshiyuki wa aomori-ken nishi-tsurugarun gun ajigasawa-cho shusshin de isegahama beya shozoku no geneki ozumo rikishi (Yoshiyuki Homarefuji is from Ajigasawa-cho, Nishitsugaru-gun, Aomori Prefecture, belongs to Isegahama stable, and is an active sumo wrestler.) |
| Document 5 | shincho 192 cm, taiju 120 kg | (The height is 192 cm, and the weight is 120 kg.) | taue akira wa, nihon no jitsuyoka, moto puroresura, moto ozumo rikishi (Taue Akira is a Japanese businessman, a former professional wrestler, and a former sumo wrestler.) |

Table 8 An example of a part of a table obtained using the proposed method (Wikipedia (wrestler))

### 4.4 Discussion in comparison with conventional method

In an experiment using 15 types of multiple documents, the average evaluation result of the proposed method was 0.65, which exceeded the maximum and average values of the conventional method. In addition, based on a paired two-sided $t$ test with a significance level of 0.05, a significant difference was found between the evaluation results of the conventional method and the proposed method. In the proposed method, by optimizing the balance between the filling degree and the density of information in a table, it is possible to more accurately estimate the
number of clusters (columns in the table) suitable for organizing the table. As a result, the accuracy of the table can be improved.

A comparison of the number of clusters estimated using the conventional and proposed methods with the number of columns in the correct table is shown in Table 11. In addition, a comparison of the absolute values of the difference between the number of clusters estimated using the conventional and proposed methods and the number of columns in the correct table is shown in Table 12. From these results, the difference from the number of columns in the correct answer table was on average 6.7 for the conventional method. This difference was reduced to 3.8 in the proposed method. Therefore, the conventional method’s problem of the $X$-means method’s tendency to underestimate the optimum number of clusters was improved by the proposed method.

During the experiments, the accuracy of a table was calculated as the average of the precision rates, recall rates, and $F$-measures in the columns of the table. These results, shown in Table 13,
|                              | Conventional method | Proposed method | Correct table |
|-----------------------------|---------------------|-----------------|---------------|
|                             | Maximum  | Average  | Minimum |          |          |
| Newspaper article (robbery) | 9    | 9       | 5       | 10      | 8       |
| Newspaper article (finance) | 8    | 4       | 2       | 14      | 13      |
| Newspaper article (earthquake) | 9    | 6       | 2       | 11      | 10      |
| Newspaper article (traffic accident) | 8    | 5       | 2       | 11      | 10      |
| Newspaper article (recall)  | 8    | 5       | 2       | 14      | 11      |
| New product article (smartphone) | 34   | 26      | 7       | 40      | 24      |
| New product articles (TV)   | 25    | 12      | 5       | 24      | 25      |
| New product article (camera) | 18   | 11      | 2       | 24      | 33      |
| New product article (robot vacuum cleaner) | 20   | 11      | 3       | 23      | 27      |
| New product article (air conditioner) | 15   | 12      | 3       | 13      | 16      |
| Wikipedia (castle)          | 11    | 6       | 2       | 16      | 9       |
| Wikipedia (dinosaur)        | 4     | 3       | 2       | 12      | 8       |
| Wikipedia (wrestler)        | 13    | 6       | 5       | 13      | 10      |
| Wikipedia (mountain)        | 5     | 3       | 2       | 9       | 9       |
| Wikipedia (baseball team)   | 6     | 8       | 4       | 7       | 9       |
| Macro average               | 12.9  | 8.5     | 3.2     | 16.1    | 14.8    |

**Table 11**  Number of columns in generated tables

|                              | Conventional method and correct table | Proposed method and correct table |
|-----------------------------|---------------------------------------|----------------------------------|
|                             | Maximum  | Average  | Minimum |          |          |
| Newspaper article (robbery) | 1    | 1       | 3       | 2       |          |
| Newspaper article (finance) | 5    | 9       | 11      | 1       |          |
| Newspaper article (earthquake) | 1   | 4       | 8       | 1       |          |
| Newspaper article (traffic accident) | 2   | 5       | 8       | 1       |          |
| Newspaper article (recall)  | 3    | 6       | 9       | 3       |          |
| New product article (smartphone) | 10  | 2       | 17      | 16      |          |
| New product article (TV)    | 0    | 13      | 20      | 1       |          |
| New product article (camera) | 15   | 22      | 31      | 9       |          |
| New product article (robot vacuum cleaner) | 7   | 16      | 24      | 4       |          |
| New product article (air conditioner) | 1   | 4       | 13      | 3       |          |
| Wikipedia (castle)          | 2    | 3       | 7       | 7       |          |
| Wikipedia (dinosaur)        | 4    | 5       | 6       | 4       |          |
| Wikipedia (wrestler)        | 3    | 4       | 5       | 3       |          |
| Wikipedia (mountain)        | 4    | 6       | 7       | 0       |          |
| Wikipedia (baseball team)   | 3    | 1       | 5       | 2       |          |
| Macro average               | 4.1  | 6.7     | 11.6    | 3.8     |          |

**Table 12**  Difference between the number of columns in each table and the number of columns in an optimal table
Table 13  Precision (average) and recall (average) in a table

|                         | Conventional method (average)          | Proposed method           |
|-------------------------|----------------------------------------|---------------------------|
|                         | Average precision | Average recall | Average F-measure | Average precision | Average recall | Average F-measure |
| Newspaper article (robbery) | 0.57 | 0.71 | 0.58            | 0.67 | 0.72 | 0.66            |
| Newspaper article (finance) | 0.28 | 0.89 | 0.37            | 0.62 | 0.72 | 0.63            |
| Newspaper article (earthquake) | 0.45 | 0.81 | 0.49            | 0.76 | 0.77 | 0.72            |
| Newspaper article (traffic accident) | 0.31 | 0.76 | 0.39            | 0.53 | 0.71 | 0.51            |
| Newspaper article (recall) | 0.38 | 0.74 | 0.41            | 0.74 | 0.64 | 0.65            |
| New product article (smartphone) | 0.60 | 0.69 | 0.58            | 0.87 | 0.75 | 0.78            |
| New product article (TV) | 0.33 | 0.73 | 0.35            | 0.67 | 0.72 | 0.64            |
| New product article (camera) | 0.24 | 0.71 | 0.31            | 0.55 | 0.77 | 0.58            |
| New product article (robot vacuum cleaner) | 0.32 | 0.84 | 0.35            | 0.60 | 0.75 | 0.60            |
| New product article (air conditioner) | 0.39 | 0.68 | 0.40            | 0.48 | 0.71 | 0.51            |
| Wikipedia (castle) | 0.50 | 0.89 | 0.52            | 0.80 | 0.78 | 0.76            |
| Wikipedia (dinosaur) | 0.31 | 0.80 | 0.42            | 0.93 | 0.70 | 0.78            |
| Wikipedia (wrestler) | 0.50 | 0.89 | 0.52            | 0.90 | 0.87 | 0.84            |
| Wikipedia (mountain) | 0.26 | 0.89 | 0.35            | 0.62 | 0.82 | 0.65            |
| Wikipedia (baseball team) | 0.44 | 0.77 | 0.40            | 0.52 | 0.80 | 0.51            |
| Macro average | 0.39 | 0.79 | 0.43            | 0.68 | 0.75 | 0.65            |

indicate that the average recall rate of the conventional method is 0.79, and its average precision rate is as low as 0.39. This indicates that the characteristics of the columns in the table obtained using the conventional method tend to include irrelevant information, and the information is less likely to be missed. By contrast, with the proposed method, the average recall rate is 0.68 and the precision rate is 0.75. The gap between the two indicators is small, indicating that a well-balanced table is obtained using the proposed approach.

5  Conclusion

Okazaki et al. proposed a method for organizing sentence information contained in multiple documents into a table using the X-means method (Okazaki et al. 2018). The X-means method is a clustering approach that estimates the optimum number of clusters based on BIC. However, the number of clusters (that is, the number of columns in the table) estimated using the X-means method tends to be smaller than optimum, which can decrease the accuracy of the resulting table because too much information will be represented in a single column. To address this problem, we proposed a method for estimating the number of clusters by optimizing the balance between the degree of filling and the density of information in tables resulting from hierarchical clustering of the sentence information.
With the proposed method, information is first clustered hierarchically. For hierarchical clustering with the number of clusters ranging from 1 to \( n \), the degree of filling and the degree of density of information of the resulting table are calculated. The number of clusters when the balance between these two indicators is optimal is selected as the optimal number of clusters, and the final table is constructed accordingly. In this study, we were able to improve the accuracy of tables generated by Okazaki et al.’s method by applying our proposed approach. The average of the evaluation results (\( F \)-measure) of the tables generated using the conventional method was 0.43; in addition, our method achieved a value of 0.65, thus confirming its effectiveness.

References

Akano, H., Murata, M., and Ma, Q. (2017). “Detection of Inadequate Descriptions in Wikipedia using Information Extraction based on Word Clustering.” In IFSA-SCIS 2017, pp. 1–6.

Ali, F., Kwak, D., Khan, P., Ei-Sappagh, S. H. A., Islam, S. M. R., Park, D., and Kwak, K.-S. (2017). “Merged Ontology and SVM-Based Information Extraction and Recommendation System for Social Robots.” IEEE Access, 5, pp. 12364–12379.

Ayaka, S. and Matsuda, S. (2011). “Comparison of Automatic Cluster Number Determination Methods in Cluster Analysis.” Academia Information Sciences and Engineering, 11, pp. 17–34.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). “Enriching Word Vectors with Subword Information.” arXiv preprint arXiv:1607.04606.

Chang, C.-H., Kayed, M., Girgis, M. R., and Shaalan, K. F. (2006). “A Survey of Web Information Extraction Systems.” IEEE Transactions on Knowledge and Data Engineering, 18 (10), pp. 1411–1428.

Fader, A., Soderland, S., and Etzioni, O. (2011). “Identifying Relations for Open Information Extraction.” In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545.

Finkel, J. R., Grenager, T., and Manning, C. (2005). “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling.” In Proceedings of the 43rd Annual Meeting of the ACL, pp. 363–370.

Fukuda, S., Nanba, H., and Takezawa, T. (2012). “Extraction and Visualization of Technical Trend Information from Research Papers and Patents.” D-Lib Magazine.

Gonzalez, E. and Turmo, J. (2009). “Unsupervised Relation Extraction by Massive Clustering.” In 2009 Ninth IEEE International Conference on Data Mining, pp. 782–787.
Grishman, R. (2015). “Information Extraction.” *IEEE Intelligent Systems, 30* (5), pp. 8–15.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). “Distributed Representations of Words and Phrases and Their Compositionality.” *Advances in Neural Information Processing Systems, 26*, pp. 3111–3119.

Mojena, R. (1977). “Hierarchical Grouping Methods and Stopping Rules: An Evaluation.” *The Computer Journal, 20* (4), pp. 359–363.

Nishida, K., Sadamitsu, K., Higashinaka, R., and Matsuo, Y. (2017). “Understanding the Semantic Structures of Tables with a Hybrid Deep Neural Network Architecture.” In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17)*, pp. 168–174.

Okazaki, K., Murata, M., and Ma, Q. (2018). “Extracting Important Information from Multiple Documents and Creating Tables.” In *Proceedings of the 24th Annual Meeting of The Association for Natural Language Processing*, pp. 240–243. (in Japanese).

Pelleg, D. and Moore, A. (2000). “X-means: Extending K-means with Efficient Estimation of the Number of Clusters.” In *ICML ’00 Proceedings of the 17th International Conference on Machine Learning*, pp. 727–734.

Rousseeuw, P. (1987). “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics, 20* (1), pp. 53–65.

Stanovsky, G., Michael, J., Zettlemoyer, L., and Dagan, I. (2018). “Supervised Open Information Extraction.” In *Proceedings of NAACL-HLT 2018*, pp. 885–895.

Swain, M. C. and Cole, J. M. (2016). “ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature.” *Journal of Chemical Information and Modeling, 56*, pp. 1894–1904.

**Masaki Murata:** received his BS and MS degrees and his PhD in engineering from Kyoto University in 1993, 1995, and 1997, respectively. He worked in the Communications Research Laboratory (currently the National Institute of Information and Communications Technology (NICT)), Japan, from 1998 to 2010. In 2010, he moved to Tottori University, Japan, where he worked as a professor in the Department of Information and Electronics, Graduate School of Engineering. He is also with the Cross-informatics Research Center, Tottori University. His research interests include natural language processing, machine translation, and information retrieval.

**Kensuke Okazaki:** received his BS and MS degrees in engineering from Tottori
University, Japan, in 2018 and 2020, respectively.

Qing Ma: received his BS degree in electrical engineering from Beihang University, China, in 1983 and his MS and Dr. Eng. degrees in computer science from the University of Tsukuba, Japan, in 1987 and 1990, respectively. He worked in Ono Sokki Co., Ltd., Japan, from 1990 to 1993 and in the Communications Research Laboratory (currently, the National Institute of Information and Communications Technology (NICT)), Japan, from 1993 to 2003. In 2003, he moved to Ryukoku University, Japan, as a professor in the Applied Mathematics and Informatics Course, Faculty of Advanced Science and Technology. His research interests include machine learning and natural language processing.

(Received February 1, 2021)
(Revised April 28, 2021)
(Accepted June 9, 2021)