SUMMARY Nowadays, there are a great deal of e-documents being accessed on the Internet. It would be helpful if those documents and significant extract contents could be automatically analyzed. Similarity analysis and topic extraction are widely used as document relation analysis techniques. Most of the methods being proposed need some processes such as stemming, stop words removal, and etc. In those methods, natural language processing (NLP) technology is necessary and hence they are dependent on the language feature and the dataset. In this study, we propose novel document relation analysis and topic extraction methods based on text compression. Our proposed approaches do not require NLP, and can also automatically evaluate documents. We challenge our proposal with model documents, URCS and Reuters-21578 dataset, for relation analysis and topic extraction. The effectiveness of the proposed methods is shown by the simulations.

key words: topic extraction, document analysis, PRDC, relation analysis, clustering, data compression

1. Introduction

Digital data are increasingly used in daily life and work. When handling enormous number of digital documents, it is convenient to appropriately know the topics or keywords in advance. There are often a large number of overlapped topics among documents which makes the analysis of documents difficult. The traditional method adopted by portal site, directory type services, demands the manual classification and arrangement of the web documents including different topics. However, as the number of web pages is getting increased, the loads imposed by manual classification have become impractically heavy. Hence, searching engine technology for information retrieval becomes popular recently. Document analysis and classification is one of the important technologies to implement information retrieval. Therefore, methods for automatic document analysis with multiple topics are required [1].

Topic analysis helps to understand data, classify information, and retrieve information. Recently, there are some proposals about topic extraction and its applications. Yokoi, T. et. al. applied non-negative matrix factorization (NMF) [2] to document sets, and carried out an evaluation with the compatibility of topics between the integrated topics and the topics from the large document set by the NMF. An efficient topic-based document retrieval model (TDRM) was proposed by Jia Xi-ping et al [3]. Instead of documents, a common topic space was used to represent them in TDRM. Compared to the traditional document retrieval methods, the authors showed that their method provided higher average precision and recall when retrieving documents. Bingjun Sun et al. introduced an unsupervised method for shared topic detection and topic segmentation of multiple similar documents using mutual information (MI) and weighted mutual information (WMI) [4], based on the consideration of the optimal segmentation maximizes MI. As a result, their method worked well for single-document segmentation, shared topic detection, and multi-document segmentation. There is also a method summarizing topic-focused multi-document with integrating the relevance of the sentences to the specified topic into the graph-ranking based method proposed by Xiaojun Wan et al. [5]. The importance of the cross-document relationships between sentences was showed in their experiment results, using real life documents, for topic-focused multi-document summarization. Also Xi-Ping Jia et al. proposed a model based on a bipartite graph in which document correlation is represented as the optimal matching of the graph [6]. They performed experiments on correlated document search that showed their method outperformed the vector space model. On the other hand, Yap, I. et al. used maximally frequent word sequences (MFSs) to extract topics [7]. The authors introduced word sequences to a document clustering method and showed the effectiveness of their method. Recently, Zhou Jie et al. proposed an opinion analysis method for real network corpus analysis based on building opinion topics in network reviews [8]. Their method could extract opinion topics under word-segmentation error and phrase-topic. As a useful approach to matrix factorization, singular value decomposition (SVD) is a well-known data analysis method and widely used in document analysis. For topic extraction of documents, SVD is applied to latent semantic indexing (LSI) to identify the eigensystem for large scale data. In comparison with the classic data analysis model based on index terms which is vague and noisy, LSI and SVD are concept-based methods. Rodrigues, R. and Asnani, K. presented some experimental results to show the effect of LSI and SVD on document analysis [9]. Recently, independent component analysis (ICA) is applied to data analysis. Wang Xiao-bin et al. proposed an ICA-based algorithm for classifying hidden Web databases with the relation to the topic domains [10]. Their proposal showed the effect of ICA applying to document analysis and provided high average precision.
Natural language processing (NLP) is widely used for analyzing documents. It is a fundamental technique of the above mentioned methods as well. NLP helps to provide high analytic performance for a special language, whereas it also restricts its application areas. The computation cost of the above methods is also high. In this paper, we propose a simple topic extraction approach providing high performance. We have developed a method to uniformly analyze sound, image and text data without natural language processing. We named this method as PRDC (Pattern Representation Scheme Using Data Compression) [11], which represents the feature of data as compressibility vector. Then analysis is made among the vectors. We have applied it to analyze multi-media data and some effective results were obtained.

In this paper, we will analyze document relation and extract topics with adapted PRDC [11]. We will propose a compressibility-based document relation analysis method for analyzing small scale corpus manually, and a compressibility-based topic extraction method for analyzing large scale corpus automatically. Our proposals focus on the compressibility of documents without carrying out natural language processing. Model documents will be created to verify the proposed methods. Real documents are also used in the simulations. Simulation results will show that the proposed methods can help us to understand complex relationship among documents without using natural language processing.

The rest of this paper is organized as follows. Our proposed approaches are explained in Sect. 2. The experimental results of the proposed methods are showed in Sect. 3. Section 4 concludes this paper.

2. Data Compression for Relation Analysis and Topic Extraction

2.1 Relation between Documents

The overlap of topics among a collection of documents from the Internet is complex. The essential relation among documents is represented in Fig. 1. As it is shown, documents A, B, and C have common topics and the relation between them can be considered as partial inclusion relation. The oblique lines show the relation between them. Document D is isolated because that it has no common topic with documents A, B or C. And complex relation among documents can be regarded as the extension of Fig. 1. The document analysis method proposed in this paper is to extract the relation (overlap topics) showed in Fig. 1. We first present the proposed compressibility-based relation analysis method. Based on the results and analysis of the compressibility-based relation analysis method, we extend it to a more robust and automatic compressibility-based topic extraction method.

2.2 Representation of Document Features Using PRDC

To extract features of documents, text compression is used for the representation of the documents in this paper. The analysis of document relation and the extraction of overlapped topics are both derived from this representation.

A model of input information source is used for encoding the input string in text compression. Moreover, a compression dictionary is used as the model in general. The compression dictionary is automatically produced when compressing an input document with Lempel-Ziv (LZ) compression [12], [13]. This implies that the dictionary derived from a document can describe those features frequently appeared in the document. The greedy compression of LZW means the more frequent a word appeared, the longer part of that word will be registered in the dictionary tree. Different tenses of a regular verb shares a same prefix are considered identical and are registered in the same child tree in a dictionary. Such that it improves the representation ability of our proposed approach. Then the dictionaries are used to represent other documents. According to this idea, PRDC constructs a compressibility space, also called feature space, using compression dictionaries. For analyzing object documents, PRDC projects them into the feature space to investigate. Therefore, we can get the feature of documents represented by a compressibility vector. Finally, PRDC classifies documents by analyzing these compressibility vectors.

PRDC is implemented as follows for relation analysis of similar documents with overlapped topics. In PRDC, a few compression dictionaries are obtained by compressing randomly selected input documents with Lempel-Ziv (LZ) compression. These dictionaries are used to constitute a compressibility space. Compressibility vector table (Table 1) is made by projecting the documents into the compressibility space and computing the compressibility vector for each document. Let \( N_i \) be an input document. By compressing the input document, a compression dictionary is

```
| \( D_{N_1} \) | \( C_{N_1 D_{N_2}} \) | \( C_{N_1 D_{N_3}} \) | \( \ldots \) | \( C_{N_1 D_{N_n}} \) |
| \( D_{N_2} \) | \( C_{N_2 D_{N_3}} \) | \( C_{N_2 D_{N_4}} \) | \( \ldots \) | \( C_{N_2 D_{N_n}} \) |
| \( D_{N_3} \) | \( C_{N_3 D_{N_4}} \) | \( C_{N_3 D_{N_5}} \) | \( \ldots \) | \( C_{N_3 D_{N_n}} \) |
| \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) | \( \vdots \) |
| \( D_{N_n} \) | \( C_{N_n D_{N_1}} \) | \( C_{N_n D_{N_2}} \) | \( C_{N_n D_{N_3}} \) | \( \vdots \) |
```

Table 1 Compressibility vector table.
obtained, which is expressed as $D_{N_j}$. Compressing document $N_j$ by $D_{N_j}$, we get compression ratio $C_{N_j D_{N_j}} = \frac{K_{N_j}}{L_{N_j}}$. Where, $L_{N_j}$ is the length of the input stream $N_j$, $K_{N_j}$ is the length of the output stream. Compressing with all of the dictionaries, we obtain a compression ratio vector for each input document. In the compressibility vector table (Table 1), the columns show the document data $N_j$, the rows show the compression dictionary $D_{N_j}$ formed by the same document, and the elements show the compression ratio $C_{N_j D_{N_j}}[\%]$. The similar text in different document can be extracted by clustering compression ratio vector.

In this paper, we apply PRDC to document analysis. In our proposal, a novel construction method of the compressibility space for PRDC is investigated. And removal of stopwords under PRDC framework is introduced. Hence, natural language processing, such as stemming, stop words removal, is not necessary.

2.3 Approach of Document Relation Analysis and Topic Extraction Based on PRDC

The proposed document relation analysis and topic extraction are described in this section. The flow chart of the proposed method is shown in Fig. 2. As in the preprocessing procedure of most document analysis methods, we unite the code of characters, remove the white spaces, newlines, tabs, and all non-alphabetical characters.

2.3.1 Approach of Compressibility-Based Relation Analysis

For manually analyzing small scale documents, the proposed compressibility-based relation analysis method is described in this section. First, the given dataset is separated to several clusters. During this process, stop words are removed. There is a kind of noise called stop words in documents and also in data collections which decreases the performance of document analysis. Stop words are literally meaningless words which cannot be used to category documents in English documents. Stop list has been utilized to remove stop words for English documents so far. A word in the stop list almost always appears in an English document no matter which field the document is in. However, when analyzing documents in a special field, more common words may be considered as stop words which are not in the stop list. Based on this consideration, PRDC is adapted in this proposed method to remove those stop words when processing a dataset by taking account of compressibility of each word. We mapped each word or segment into the compressibility vector space. The words or segments near the origin, which are compressed by all of the dictionaries, are thought to be stop words. After the processing of removing the words, the performance of analyzing documents will be more efficient and accurate.

Then, in each separated cluster, there are common topics between the documents. Here, we extract the common topics in each cluster. The topics in a certain cluster are composed of the repetition of some specific phrases or words appeared more than once in different documents. If these phrases or words can be found, it is possible to understand the content of the common topics in the cluster. PRDC is again used to discover the specific phrases and words. A compressibility space consisted of the dictionaries generated from all documents belong to the same cluster, is constructed. Next, all of the documents in the cluster are divided into fragments of length $L$ characters or words, and all fragments are mapped into the compressibility space (Fig. 3).

From the study of the compressibility vector of each fragment, specific phrases or words can be discovered. Because those fragments that are compressed into the same level by all compression dictionaries appear in common in a cluster. In other words, the fragments plotted in the diagonal of the compressibility space can be considered as
The problem of random selection of dictionaries in PRDC, may be considered as selection of a number of dictionaries to feature a large cluster in the data space. In that case, it can be thought there are only a few large clusters in the data space. A large cluster occupies a wide area in the data space. The selected dictionaries in which may locate far away from each other. The cluster cannot be represented properly, if some dictionaries belong to the margin area of the cluster are selected. Furthermore, the larger the clusters become, the larger the probability of selecting bias is.

Small clusters in a data space are considered as ‘pure’ ones. In which the selection of dictionaries which are close to the centroid of each cluster, is possible to make a representative feature space. This kind of selection may let us know the number of dictionaries to select on the whole. Through this selection method, one also is able to know where to select dictionaries in a data space. For example, one can select one or more dictionaries in each small pure cluster to make a dictionary space. We first randomly select some dictionaries to build a dictionary space. We classify the input documents into a large number of clusters to obtain small and pure ones. Then, we select one centroid document from each cluster to construct dictionaries and a new dictionary space is constructed. Better performance can be obtained by using the new compressibility vector space to represent documents.

To extend compressibility-based relation analysis method and automatically extract topics, we compose a long document by concatenating all incoming documents first. Then the long document is separated to many fragments with length of \( L \) words or characters. The adapted PRDC is employed to classify the fragments to many clusters. Each centroid fragment is considered as a representation of the cluster which it belongs to. Moreover, the words in the centroid fragment are extracted topics. In comparison with our compressibility-based relation analysis, this proposal is more flexible and simple but with less specification.

3. Simulation Results and Analysis

In this section, we will show four simulations with using the proposed methods. In Simulation 1 and 2, we use artificially constructed documents (model documents) to verify our proposed methods at the principle level. This is because that the topics can be different according to readers’ understanding and purpose of a real document, and it is difficult to capture an objective topic extraction from the documents. Later in Simulation 3 and 4, we will use real documents. The quantitative experiments will be carried out to compare our proposed method with SVD (singular value decomposition) and ICA (independent component analysis). We will start with explaining how to make the model documents.

3.1 Generation of Model Document

A topic of a document consists of the repetition of special phrases or words (basic phrase). We use a few words as the basic phrase and change the number of the repetition arbitrarily. After which, we insert some words randomly as noise character, between basic phrases. These words (noise
character) are not included in the basic phrase vocabulary. In order to have a large number and non-repeatability for characters, we use multibyte character set. In Simulation 1, the noise character is extracted from the collection of 7000 characters. Each character is selected randomly, and is buried between basic phrases. By changing the topic and the number of repetition, we can make different model documents with different relation among them (Fig. 4). The proposed relation analysis method will be applied to the model documents to verify the properties of the proposed method.

3.2 Simulation 1 (Principle Examination)

In this simulation, three model documents with common basic phrases are generated. We will show how the fragments with or without the basic phrases included are distributed in the compressibility space. The basic principle of the proposed method is also examined.

3.2.1 Model Document

According to the generation method, model documents A, B and C with a common topic are made. The relation among these documents is shown in Fig. 5. The number of noise characters is fixed to 1000 for each document. To insert the topic, two types of basic phrase are made and each of which appears ten times respectively in each document.

3.2.2 Distribution of Common Fragments

The distribution of fragments in model documents A, B and C is studied. First, a model document is divided to fragments of L characters length respectively. To prevent punctuating basic phrase in fragments, we assume that \( L = 10 \) (> maximum length of both basic phrases). All fragments are mapped into the compressibility space which is composed with the dictionaries by compressing documents A, B and C. We focus on model document A (the other cases of focusing on B or C are similar). To study the relation that the fragments in document A have with documents B and C, the fragments in A are mapped onto the plane of B-C (Fig. 6). The points in the figure indicate the fragments. Each point in Fig. 6 may stand for more than one fragment, for the reason of that there are a few different patterns in this simple example, so that many fragments have the same compressibility.

As shown in Fig. 6, the fragments including the basic phrases are plotted on the diagonal, since these fragments in document A exist both in documents B and C. This result shows the proposed method can correctly represent the relation among documents A, B and C in principle level. The distance to the origin is shorter when the basic phrase becomes longer. On the other hand, \((C_{AD}, C_{CD}) = (100, 100)\) indicates fragments with only noise characters that are neither compressed by compression dictionaries B nor C. These fragments can be considered as unknown fragments. They are plotted as distances from the origin.

3.3 Simulation 2 (Compressibility-Based Relation Analysis on Model Documents)

3.3.1 Data for Simulation

We made model documents A, B and C with the relation shown in Fig. 7. The variety of the noise character is decreased from 7000 types (Simulation 1) to 20 types, to show the influence by stop words between documents. From the 20 types of noise characters, we randomly extracted 1000 characters repeatedly, to make 3 model documents A, B and C, so the documents can share the same stop words. Seven of basic phrase are used 10 times in each document.

- monkey : basic phrase in A, B and C
- eagle : basic phrase in A and B
Fig. 7 Model document topics of Simulation 2.

Table 3 Compressibility vector table [%].

|   | A   | B   | C   | D   | E   | F   | G   | H   | I   | J   |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| D_A| 55.3| 55.3| 55.3| 55.3| 55.3| 55.3| 55.3| 55.3| 55.3| 55.3|
| D_B| 52.3| 52.3| 52.3| 52.3| 52.3| 52.3| 52.3| 52.3| 52.3| 52.3|
| D_C| 50.3| 50.3| 50.3| 50.3| 50.3| 50.3| 50.3| 50.3| 50.3| 50.3|
| D_D| 48.3| 48.3| 48.3| 48.3| 48.3| 48.3| 48.3| 48.3| 48.3| 48.3|
| D_E| 46.3| 46.3| 46.3| 46.3| 46.3| 46.3| 46.3| 46.3| 46.3| 46.3|
| D_F| 44.3| 44.3| 44.3| 44.3| 44.3| 44.3| 44.3| 44.3| 44.3| 44.3|
| D_G| 42.3| 42.3| 42.3| 42.3| 42.3| 42.3| 42.3| 42.3| 42.3| 42.3|
| D_H| 40.3| 40.3| 40.3| 40.3| 40.3| 40.3| 40.3| 40.3| 40.3| 40.3|
| D_I| 38.3| 38.3| 38.3| 38.3| 38.3| 38.3| 38.3| 38.3| 38.3| 38.3|
| D_J| 36.3| 36.3| 36.3| 36.3| 36.3| 36.3| 36.3| 36.3| 36.3| 36.3|

- cat : basic phrase in A and C
- horse : basic phrase in B and C
- dog : basic phrase in A
- owl : basic phrase in B
- hawk : basic phrase in C

Moreover, besides documents A, B and C, we extracted 1000 noise characters from the 20 types repeatedly to generate another 7 model documents D, E, ···, J without basic phrase.

3.3.2 Document Relation Analysis

The generated model documents are compressed respectively, and the compressibility space is composed of the generated compression dictionaries. Table 3 shows the result of mapping all fragments of model documents (A, ···, J) to the compressibility space.

In documents A, B and C of simulation 1, we used many different noise characters with each one of them appearing only once in each document. So that the fragments do not contain basic phrases that can neither be compressed by compression dictionary $D_B$ nor $D_C$ where $(C_A|D_B, C_A|D_C) = (100, 100)$.

If a fragment contains stop word(s) that are registered in compression dictionary $D_B$ or $D_C$, it moves close to the origin and shows closer relation between each other. Clearly stop words cause poor performance in analyzing documents relation.

Generally, stop words exist in any documents. Therefore, it is more accurate to remove stop words before document analysis and the essential common topic can be obtained. We implement removal of stop words by using the method described in Sect. 2.3.1. The results are shown in Tables 4 and 5.

Table 4 shows that fragments which do not include “monkey”, a common topic among model documents A, B and C, are extracted. A similar phenomenon is also found with model documents B and C. This means that the basic phrases of common topics and fragments with only noise characters are both extracted as the topics. That is due to the existence of stop words. Table 5 shows the extracted topics after removing the fragments including stop words.

After removal of stop words, the relationship of documents A, B, ···, J is shown in Fig. 8. We can see that model documents A, B and C that have the related topics were extracted as the same group, and separated from others. That shows that basic phrases in A, B and C have similar compressibility by sharing common topics.

Next, the relation of extracted group (composed of model documents A, B and C) is analyzed. First, the model documents are divided into fragments of $L$ characters ($L = 10$) respectively as that in Simulation 1. And, they are mapped into the compressibility space which is composed of compression dictionaries $D_A$, $D_B$ and $D_C$.

Here, we focus on document A and study its relation between model documents B and C. For that, the fragments of document A are mapped onto the B-C plane. Then, the relation of model documents A, B and C is shown in Table 6. As the common topic among A, B and C, fragments including basic phrase “monkey” are extracted. Fragments with “eagle” in A and B, and “cat” in A and C are extracted. Moreover, each peculiar fragment “dog”, “owl”, and “hawk” in model documents A, B and C respectively, can also be extracted. Therefore, the relation analysis be-
between documents can be well achieved in model documents that even contain stop words, by using compressibility-based relation analysis.

3.4 Simulation 3 (Dictionary Selection)

In this simulation, we show how to choose the dictionaries to effectively construct a compressibility space while processing large scale document set. A popular clustering method called k-means and purity will be used to evaluate the performances.

Purity is a simple and transparent evaluation measure [14]. The overall purity of a clustering solution is defined as the weighted sum of individual cluster purities:

$$\text{Purity} = \sum_{r=1}^{k} \left( \frac{n_d}{n} \right) P(S_r)$$

where \( P(S_r) \) is the purity for a particular cluster of size \( n_r \), \( k \) is the number of clusters and \( d \) is the total number of data items in the dataset. Purity of a single cluster is defined by

$$P(S_r) = \frac{n_d}{n_r}$$

where \( n_d \) is the number of documents in cluster \( r \) that belong to the dominant (majority) class in \( r \), i.e., the class with the most documents in \( r \). Obviously, the higher the purity value is, the purer the cluster in terms of the class labels of its members is, and the better the clustering results becomes.

We performed this simulation using the proposed method with URCS dataset to demonstrate the validity of our proposed method on optimized compressibility vector space construction. URCS (University of Rochester Computer Science Technical Reports) consists of 609 abstracts from 4 categories. They are, Artificial Intelligence (119 items), Robotics (97 items), Systems (218 items), and Theory (175 items). They are all derived from computer science items), Robotics (97 items), Systems (218 items), and Computer Science Technical Reports) consists of 609 abstracts

3.5 Simulation 4 (Comparison of Compressibility-Based Topic Extraction, ICA, SVD and n-Gram)

Singular Value Decomposition (SVD) is a classical linear algebra method. It is widely used in signal processing and information retrieval. Independent Component Analysis (ICA) is a signal processing method and applied to information retrieval recently. N-gram is a well-known text processing method. In this study, we compared n-gram, SVD and ICA with our proposed approach on topic extraction.

In the preprocessing procedure, the white spaces, newlines, and tabs are replaced by a single space. Non-alphabetical characters are also replaced by a single space. Upper case characters are all converted to lower case. And all stop words are removed based on the standard van Rijssbergen stop word list [15]. After that, each word is stemmed by using Porter’s Stemmer [16]. After the preprocessing, a long text stream is made by concatenating the words in all the input documents. Then the long text stream is separated to a number of fragments with length of \( L \) words or trigrams. Each fragment is represented by a vector with dimension \( n \), where \( n \) is the number of the words appeared in the text stream. A feature matrix is then obtained to represent of all
fragments. SVD or ICA is applied to the feature matrix to extract topic vectors.

For PRDC, the preprocessing procedure for processing the white spaces, newlines, tabs, non-alphabetical characters and upper case characters is carried in the same way with SVD and ICA. And all stop words are removed by using the standard van Rijjsbergen stop word list, in consideration of computation complexity and different aim from the proposed compressibility-based relation analysis. Also, a long text stream is made by concatenating all input documents. Then the long text stream is separated to a number of fragments with length of \( L \) words. Then, the proposed compressibility-based topic extraction is used to classify the fragments to a number of clusters. Each centroid fragment is considered as a representation of the cluster which it belongs to. Moreover, the words in the centroid fragment are extracted topics.

A classical measure called Jaccard (Eq. (2)) is utilized to evaluate the topic extraction results. In this equation, \( \alpha \) and \( \beta \) are word sets, \( \cup \), and \( \cap \) are the number of union and intersection between \( \alpha \) and \( \beta \). In this simulation, the top 20 topics are extracted from each set by using TF-IDF. Because a subset of each topic is used, we run simulation 10 times and compute the average for evaluation.

\[
Jaccard(\alpha, \beta) = \frac{\cap(\alpha, \beta)}{\cup(\alpha, \beta)} \tag{2}
\]

A subset of Reuters-21578 is utilized in this simulation. This dataset consists of 21578 news appeared on the Reuters newswire in 1987 [17]. The documents were assembled and indexed with categories by personnel from Reuters Ltd. and Carnegie Group, Inc. Reuters-21578 is currently one of the most widely used test collection in information retrieval, machine learning, and other corpus-based research. Since the dataset contains some noise, such as repeated documents, unlabeled documents, and empty documents, we choose a subset of 10 relatively large groups (aqc, coffee, crude, earn, interest, money-fx, money-supply, ship, sugar, and trade) in our experiments. For each article in the 10 categories that will be used, only the text bodies are extracted.

3.5.1 Simulation with Large-Scale Dataset

In this simulation, we used 100 documents from the set of each topic, and the total number of used documents is 1000. The results are shown in Fig. 10. \( L \) is set to be 8, 16, 32, 64, 128, and 256 to test the performance in different situation. ICA almost shows better performance than that of SVD. The proposed method shows better performance than that of SVD and ICA methods, except \( L = 32 \). ICA and SVD show the best results when \( L = 32 \) or \( L = 64 \). However, the performance of ICA and SVD is not stable as \( L \) changes. In the contrast, the performance of the proposed method is stable when the value of \( L \) changes. N-gram method showed better results when \( L \) is smaller than 32. The proposed method showed the best results than all other three methods when \( L \) is larger than 32.

The proposed method shows better performance comparing with SVD and ICA methods. N-gram method shows better performance when \( L \) is small. The proposed method shows good and stable performance with a set of chose dictionaries, and it can avoid complicated computation of dimension reduce.

3.5.2 Simulation with Small-Scale Dataset

In this simulation, we used 10 documents from each set of topic, and the total number of used documents is 100 to test the performance in different situation. The results are shown in Fig. 11. \( L \) is set to be 8, 16, 32, 64, 128, and 256. The performance of ICA is similar to SVD, and it is better than that of the proposed method when \( L < 32 \). When \( L \) becomes larger, the proposed method shows better performance than that of SVD and ICA methods. SVD and ICA show the best results when \( L = 32 \) and \( L = 64 \). As \( L \) becomes larger, the performance of ICA and SVD do not become better. N-gram method showed better results when \( L \) is smaller than 32. The performance of the proposed method becomes better when \( L \) grows larger.

The proposed method shows better performance comparing with n-gram, SVD and ICA methods when \( L > 32 \). The proposed method shows good performance based on the selection of dictionaries, and it can avoid complicated computation of dimension reduce.
4. Conclusion

In this paper, we proposed two methods for relation analysis and topic extraction of documents by using the compressibility of data. The methods can handle documents with overlapped topics. The results from using model documents and actual documents showed the effectiveness of the proposed methods. The proposed methods do not need any natural language processing technique. In the simulations, we achieved the goal of generating the model documents and removing the stop word as well.

5. Future Works

For topic extraction, n-gram method has better results than our proposed approach when L is small. The proposed approach showed its advantage in terms of extracting more topics than ICA and SVD for most cases of large-scale data and some cases of small-scale data in our simulations, it still can be improved by constructing independent dictionaries. Construction of independent dictionaries will improve the representation and computation performance of the proposal when a compressibility space becomes very large. This will be studied in our future work. Knowledge discovery and topic extraction are attractive areas and have been developed very fast. We will carry on more comparative studies in our future studies.

References

[1] K. Saito, “Multiple topic detection by parametric mixture models (PMM) automatic web page categorization for browsing,” NTT Technical Review, vol.3, no.3, pp.15–18, 2005.
[2] T. Yokoi and H. Yanagimoto, “Topic extraction for a large document set with the topic integration,” 2010 3rd International Conference on Knowledge Discovery and Data Mining (WKDD 2010), pp.46–49, 2010.
[3] X.-P. Jia, H. Peng, Q.-L. Zheng, S.-X. Shi, and Z.-L. Jiang, “Topic-based document retrieval model,” Journal of South China University of Technology, vol.36, no.9, pp.37–42, 2008.
[4] B.J. Sun, P. Mitra, C.L. Giles, J. Yen, and H.Y. Zha, “Topic segmentation with shared topic detection and alignment of multiple documents,” 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.199–206, 2007.
[5] X. Wan, J. Yang, and J. Xiao, “Using cross-document random walks for topic-focused multi-document summarization,” 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp.1012–1018, 2006.
[6] X.-P. Jia, H. Peng, Q.-L. Zheng, Z.-L. Jiang, and Z. Li, “A topic-based document correlation model,” 2008 International Conference on Machine Learning and Cybernetics (ICMLC), vol.5, pp.2487–2491, 2008.
[7] I. Yap, T.L. Han, S. Lixiang, and L. Ying, “Topic detection using MFS,” 19th International Conference on Industrial, Engineering and other Applications of Applied Intelligent Systems, IEA/AIE 2006, vol.4031, pp.342–352, 2006.
[8] J. Zhou, C. Lin, and B. Li, “Research on the identification of opinion topic expressed in web comments,” J. China Society for Scientific and Technical Information, vol.29, no.5, pp.858–863, 2010.

Nuo Zhang received the M.S. degree in foundations of information and computer science in 2003 and the Doctorate in information science in 2006 both from Chiba University, Japan. Since 2006, he has been with the Graduate School of Information Systems, the University of Electro-Communications in Japan, where he is an assistant professor. His current research interests include pattern recognition, information retrieval, image processing and machine learning. He is a member of IEEE.

Toshinori Watanabe received the B.E. degree in aeronautical engineering in 1971 and the D.E. degree in 1985 both from Tokyo University. Since 1971, he worked at Hitachi as a researcher in the field of information systems design. His experience includes production management systems design, VLSI design automation, knowledge-based systems, and evolutionary learning machines including TAMPOPO. He also engaged in FGCS (Fifth Generation Computer System) project of Japan and developed a concurrent VLSI layout problem solver on PIM (Parallel Inference Machine) in 1991. Since 1992, he has been a professor at the graduate school of information systems, university of Electro-Communications at Tokyo. His areas of interest include media analysis, learning intelligence, and the semantics of information systems. He is a member of the IEEE.