Finding Division Points for Time-Series Corpus Based on Topic Changes

Hiroshi Kobayashi and Ryosuke Saga

Osaka Prefecture University, 1-1, Gakuen-cho, Naka-ku, Sakai-shi, Osaka, Japan
kobayahi@mis.cs.osakafu-u.ac.jp, saga@cs.osakafu-u.ac.jp

Abstract. This paper describes the discovery method of finding proper points for dividing a corpus with time series information for extracting local and frequent keywords. Local and frequent keywords express a corpus with time series information and are useful for comprehending it. To extract keywords from the corpus, the previous works proposed corpus separating method. However, this method divides the corpus at equal intervals so that it cannot take into account the change of topic. To consider the change of topics and divide the corpus based on it, we utilize the idea of topic model and the topic extracted by Latent Dirichlet Allocation (LDA). In the experiment using newspaper articles during five years topics, we confirm that the topics of each document change as time passed by using the output from LDA and the point which is available on dividing the corpus by the change of topics notably is observable.

Keywords: keyword extraction, time series information, LDA.

1 Introduction

In recent years, many documents have been converted to digital form by the development of the Internet. This transition to "paperless" documentation reduces the time and effort required to manage documents and makes necessary information instantly available. The amount of digital document data is increasing enormously. Such a large amount of document data may be collectively saved as a corpus. A corpus aggregates a considerable amount of digitalized document data and is used for research such as natural-language processing [1]. To efficiently acquire information from a corpus, the features of the corpus must be known.

Keyword extraction is one way to effectively learn the features of a corpus [2]. Keywords and keyword extraction make searching and summarizing a corpus easy [3]. For example, when searching a document, the keyword expressing the document's content is useful to filter your results [4]. The keywords are first ranked by importance, and the top *N* keywords are extracted. The frequency of a word in the whole corpus is conventionally used as a measure of importance. Based on the frequency, there are several proposed ranking methods such as term frequency and term frequency–inverse document frequency (TF–IDF).

In contrast, several types of corpuses have time series, such as newspaper articles, magazines, papers, and blogs. Corpuses also include locally characteristic keywords,

S. Yamamoto (Ed.): HIMI 2014, Part I, LNCS 8521, pp. 364–372, 2014.
© Springer International Publishing Switzerland 2014

which appear locally and frequently. These keywords are important to comprehend trends and situations. For example, these keywords appear on Twitter as Trends. For extracting keywords from a corpus with time-series information, conventional methods based on word frequency are not very effective because the keywords are compared with the whole corpus. Therefore, the words which appear frequently and locally are not extracted because the weights from the methods tend to become low. But, when such a word also expresses a corpus, the word which appears frequently locally isn't extracted because the weight becomes low.

To solve this problem, Saga et al. proposed a method of dividing a corpus [5]. The words that appear locally can be extracted as keywords from the divided corpus because division narrows the period to observe documents. However, although normally the change of topic should be considered, because these keywords express the trend of topics, a corpus is divided at equal intervals in this method. Therefore, we should divide the corpus while considering changes of topics.

This paper proposes a method to find division points of a corpus with time-series information while also accounting for changes of topics. To comprehend a topic, we utilize the idea of a topic model and identify the topics using Latent Dirichlet Allocation (LDA). The paper is organized as follows. In Section 2, we introduce a conventional keyword extraction method for when a corpus includes time-series information and identify this method's problem. In Section 3, we describe a topic model and LDA. In Section 4 we perform an experiment for news articles to divide corpus by presuming the distribution of a topic in each document using LDA. Finally, Section 5 contains our conclusions and plans for future work.

2 Keyword Extraction Methods for Corpus with Time-Series Information

The keyword extraction method calculates the weight of all the words appearing in each document of a corpus. Then, words with large weights from each document are extracted as keywords for the document, and words with large weights across many documents are extracted as keywords for the corpus. The flow is shown in Fig.1. Words are generally assigned weights using TF–IDF and residual IDF [6].

However, for some corpuses, time information is important. For example, newspaper articles and blogs are published and updated daily or weekly, and conference proceedings and annual reports are published annually. These documents are organized in time series and change continuously during a certain period. This paper treats such a time-series corpus.

In a time-series corpus, some words occur frequently only during a specific period of time on account of certain occurrences and incidents. As shown in Fig. 2, when a region receives record amounts of snowfalls out of season, the phrase “heavy snow” appears in newspapers frequently and locally. In this situation, the phrase can be an important keyword (phrase). However, the frequency of the phrase is low from the viewpoint of the whole corpus because many people lose interest after the event.

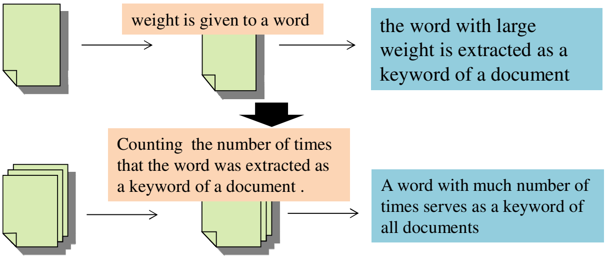

weight is given to a word
the word with large weight is extracted as a keyword of a document
Counting the number of times that the word was extracted as a keyword of a document .
A word with much number of times serves as a keyword of all documents

Fig. 1. The flow of keyword extraction

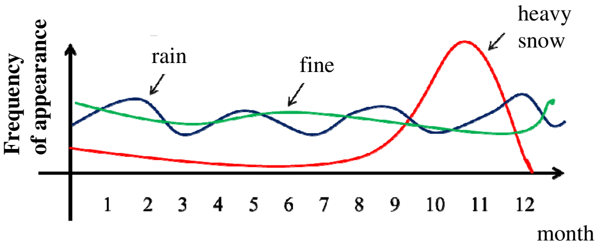


Fig. 2. The word which occurs frequently locally

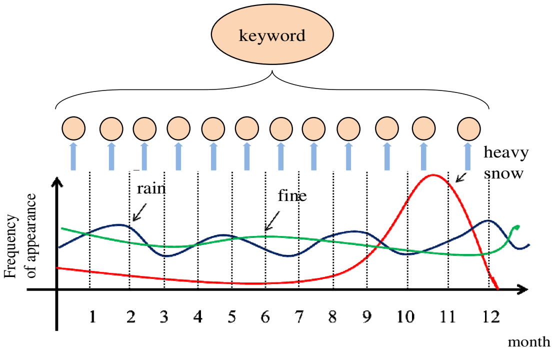


Fig. 3. The aspect of division of a corpus

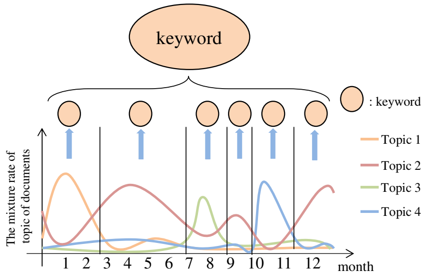


Fig. 4. The aspect of division of a corpus based on the change of topics

So the weight of the word may be lower than those of the words appearing in only a few documents during that period, but constantly, such as "fine" and "rain". As a result, words that express the features of a corpus like "heavy snow" are often not extracted as keywords.

To solve the problem of local keywords not being correctly extracted, Saga et al. proposed a method of separating a corpus according to time series. If a corpus is separated by this method, even conventional methods can be used to extract words that appear frequently but locally. This method is based on the idea that extraction of words appearing frequently and locally is attained by narrowing the period of observation for documents within a time series. Fig. 3 shows how to apply this approach to the corpus from Fig. 2. As shown in the figure, when the corpus is divided, the phrase "heavy snow," which frequently appears between October and November, can be extracted as a keyword without the interruption of other words.

This approach is useful for real data. Previous studies show how the approach can extract local keywords from newspaper articles. However, the approach divides the corpus at equal intervals. In fact, local keywords express the trends of the occurrence during certain periods, but previous studies ignore these trends.

Therefore, we must find suitable compartmental points to extract characteristic words based on changes of topics. Fig. 4 shows how to divide a corpus of documents based on changes of topics. To observe identify the changes of topics, we can utilize use the idea of topic model [7] and the topics of documents are presumed by use Latent Dirichlet Allocation (LDA) to infer the topics of documents. Using a topic model and LDA, we can try to find suitable points for compartmentalization.

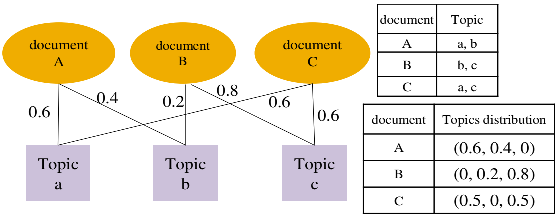
3 Corpus Separation Approach Based on Topic Change.

3.1 Topic Model and LDA (Latent Dirichlet Allocation)

A topic model is the probable generation model of a document. The probability that a word will appear varies based on document topics, and then, the probabilistic of appearance of a word depends on topic and each document has multiple-topics. This A topic model is also a statistical model for discovering a certain topic from a document. Each document of a document set is classified into topics. Topics show the contents briefly and make searching and classifying documents easy. One document may have multiple topics. This image is shown in Fig. 5. In this figure, document A is categorized into Topic a and Topic b, and the mixture ratios of two topics are 0.6 and 0.4 for each.

The LDA model is a kind type of the language probabilistic model which assumes that a document consists of multiple-topics. A word can be classified into topics and a document can be classified for every topic. Prior distribution is assumed to be Dirichlet distribution. It is because multinomial distribution can be assumed to how to choose words and topics. The graphical model of LDA is shown in Fig.6. The box expresses the number of times of execution. The nodes within box D correspond to the number of times of execution in the document. N is the number of words and T is number of topics. Each topic is defined based on the distribution of words by node φ. Then, it is probability distributions. The distribution of topics in each document is generated by node θ. The topic is randomly selected from the distribution of topics by node Z. The word is randomly selected from the distribution of words by node W. When W is observed, the posterior probability of {φ, θ, Z} is calculated based on probability density approximated by sampling. Then the posterior probability is calculated based on prior probability and the likelihood of W calculated from a generation model. α, β is the parameters to assume that the prior distribution is the Dirichlet distribution and is called the hyper-parameter.

The change of topic is observable because of the distribution of topics obtained by LDA.



document	Topic
A	a, b
B	b, c
C	a, c

document	Topics distribution
A	(0.6, 0.4, 0)
B	(0, 0.2, 0.8)
C	(0.5, 0, 0.5)

Fig. 5. The image of a multiplex topic

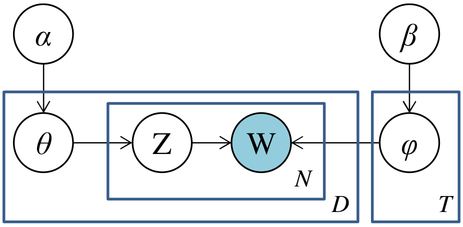


Fig. 6. Graphical model of LDA

3.2 Process of Separating a Corpus Based on Topic Change by LDA

The number of topics and hyper-parameters must be set before applying LDA to a corpus. Then, LDA learns by inputting document and outputs the topic distribution of each document. The distribution shows the mixture ratio of the topic which the document contains. Then, by illustrating the mixture ratio using time series, the rate changes of each topic in each document can be identified. These changes are easier to identify in graphs. A corpus is divided at the points where large changes of topics occur.

4 Experiments

4.1 Goal and Dataset

We performed an experiment to find the points where the topics of documents significantly changed. We used 1827 news articles from CNN.com between January 1, 2000 and December 31, 2004 as a time-series document set and the MALLET [9], a Java toolkit for machine-learning applied to natural language, to create the LDA model.

We set the number of topics (T) to 20, and optimized the hyper-parameters based on the number of topics.

4.2 Experimental Result

The word distribution outputted by LDA is shown in Table 1. We could guess some contents of topics from the LDA result. For example, Topic 6 in Table 1 is presumed to be the Iraq War because we can see words such as “iraq,” “war,” and “baghdad.” The mixture ratio of topics in each document is shown in Table 2. We found the changes of topics of a document set, including time-series information, by observing the mixture ratio of a document for every topic. Based on Table 2, we expressed the changes of topics as the line graph shown in Fig.7.

Table 1. The word distributions of 20 topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
john	gore	police	president	male	iraq	plane	afghanistan	today	class
president	today	area	house	ph	iraqi	space	taliban	percent	width
kerry	news	case	white	great	war	flight	bin	news	style
campaign	percent	county	clinton	big	baghdad	china	laden	tax	john
senator	mccain	question	senator	thing	saddam	air	united	president	stewart
election	online	shooting	blitzer	break	forces	crew	war	press	src
vote	government	morning	george	news	military	question	president	question	table
voters	correspondent	sniper	day	commercial	hussein	chinese	states	economy	href
george	elian	investigation	united	novak	troops	today	al	market	height
state	anchor	washington	country	lin	city	aircraft	military	cut	option
debate	located	point	john	female	coalition	day	today	online	brien
tonight	fdch	scene	american	story	air	miles	terrorism	anchor	color
presidential	secure	person	republican	love	today	united	qaeda	house	today
gore	clinton	man	bill	camera	iraqis	states	york	fdch	dean
democratic	york	law	state	three	general	morning	osama	case	martha
states	states	shot	party	coming	soldiers	officials	country	kagan	border
candidate	case	enforcement	states	day	marines	point	security	bill	type
night	bill	chief	great	brien	question	course	pakistan	located	input
al	day	virginia	senate	long	army	airport	government	correspondent	align

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
palestinian	storm	anthrax	iraq	material	cnnbodytext	case	court	condit	iraq
israeli	hurricane	enron	war	user	class	today	florida	levy	today
israel	florida	question	weapons	news	iraq	question	gore	police	kerry
arafat	hour	fbi	united	copyright	case	court	supreme	news	morning
palestinians	area	government	saddam	network	media	morning	election	congressman	president
peace	winds	president	cnnbodytext	license	fdch	death	state	chandra	brien
israelis	water	news	class	cable	today	family	votes	today	costello
east	miles	today	hussein	purposes	california	trial	ballots	gary	security
middle	morning	house	president	redistribute	schwarzenegger	jury	county	question	american
yasser	power	security	inspectors	rights	brien	police	vote	case	hemmer
sharon	damage	war	nations	house	news	judge	campaign	washington	break
minister	coming	states	council	media	dean	attorney	today	online	al
jerusalem	coast	september	north	long	court	neville	recount	located	iraqi
violence	city	united	security	high	question	child	case	correspondent	live
president	beach	point	states	specific	defense	evidence	count	secure	intelligence
prime	hit	pakistan	korea	federal	cooper	man	judge	fdch	war
state	hours	fact	resolution	provide	state	defense	president	anchor	government
secretary	center	intelligence	iraqi	interests	morning	harris	law	point	day
united	live	mail	today	prepared	governor	fire	vice	family	case

From Fig. 7, we confirmed that periods existed in which the topics of documents changed notably. Topics changed significantly when events related to the topics occurred, and there are related keywords about the topics in each period. Thus, we found clues about suitable separation points for changes of topics.

However, some topics, such as topic 2 and topic 10, could not be guessed from words, whereas some topics, such as Topic 5, remained at a steady state throughout the test period, i.e., we could not find any significant changes of topic. We can guess several reasons for the appearance of such a topic. For example, the number of topics may be wrong for LDA, or LDA may not be able to extract proper topics naturally from a corpus with a time series. Regardless of the reason, some topics extracted from LDA change significantly and some slightly, so we must identify and filter useful topics to discover separation points.

Table 2. Part of the topic distribution of documents

document	topic	The mixture ratio ofthe topic	topic	The mixture ratio ofthe topic	topic	The mixture ratio ofthe topic
2000/1/1	4	0.428	6	0.207	1	0.178
2000/1/2	4	0.339	1	0.337	6	0.113
2000/1/3	4	0.377	1	0.367	3	0.108
2000/1/4	1	0.467	4	0.226	3	0.131
⋮						
2004/12/28	11	0.581	4	0.144	19	0.132
2004/12/29	11	0.598	4	0.110	19	0.095
2004/12/30	11	0.542	19	0.132	4	0.126
2004/12/31	11	0.365	4	0.246	19	0.137

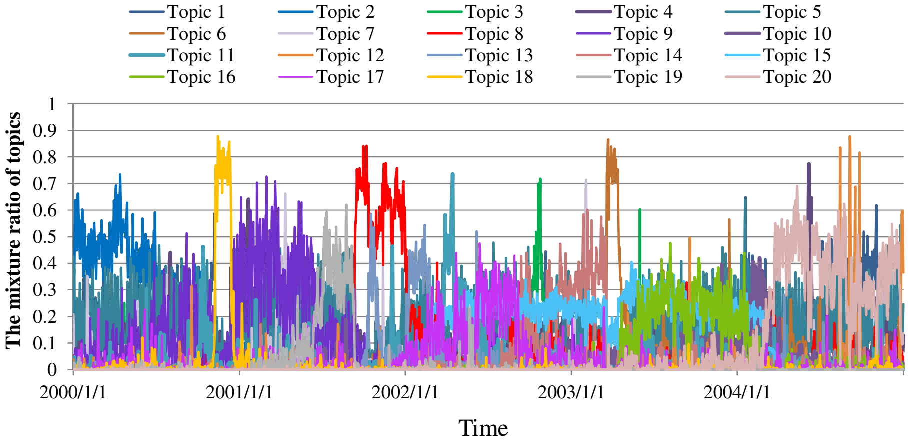


Fig. 7. The change of the mixture ratio of a document

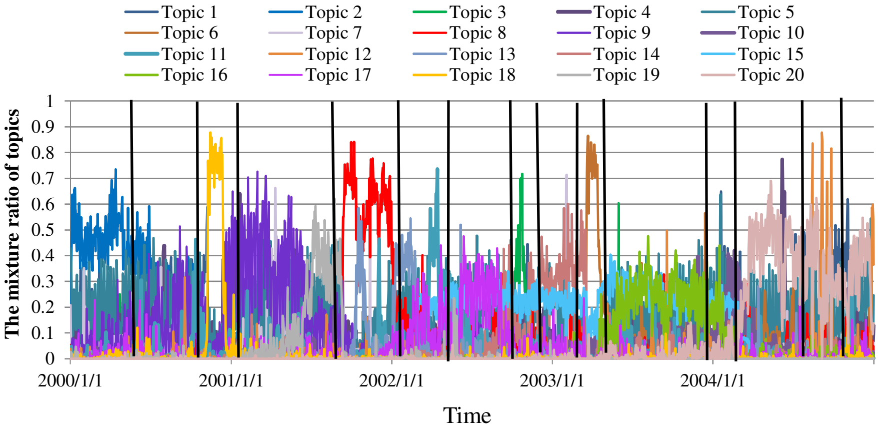


Fig. 8. The image of dividing a time series document set

5 Conclusion

In this paper, we proposed a method for finding suitable points of a corpus that includes time-series information to extract—as keywords—characteristic words appearing locally. In this case, dividing a corpus into equal time intervals is insufficient. Therefore, we needed to identify suitable compartments into which the corpus could be divided. The characteristic words appearing in a period, which is related to the topic of the document, change significantly over time. Therefore, we used LDA to examine changes of topics. As a result, we noticed that the main topics of documents changed as time passed. We can expect to extract characteristic words as keywords by dividing a document group at those points where the topic changes significantly. Fig. 8 illustrates the division of a time-series document set.

In this paper, when we performed LDA, we manually set the number of topics to 20. However, the optimal number of topics is unknown. Therefore, in future work we aim to automate the process of choosing the number of topics and evaluate the method of keyword extraction for a time-series document set

Acknowledgement. This research was supported by The Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science(JSPS), Grant-in-Aid for Young Scientists (B), 23760358 and Grant-in-Aid for Scientific Research (C), 25420448.

References

1. Liu, V., Curran, R.: Web Text Corpus for Natural Language Processing. In: Proceedings of the 11th Conference of The European Chapter of The Association for Computational Linguistics, Trento, Italy, pp. 233–240 (2006)

2. Liu, F., Liu, F., Liu, Y.: Automatic Keyword Extraction for TheMeeting Corpus Using Supervised Approach and Bigram Expansion. In: IEEE Workshop on Spoken Language Technology, pp. 181–184 (2008)
3. Dredze, M., Wallach, H., Puller, D., Pereira, F.: Generating Summary Keywords for Emails UsingTopics. In: Proceedings of The 2008 International Conference on Intelligent User Interfaces, pp. 199–206 (2008)
4. Litvak, M., Last, M.: Graph-based Keyword Extraction for Single-Document Summarization. In: Proceeding of The Workshop on Multi-source Multilingual Information Extraction and Summarization, pp. 17–24. Association for Computational Linguistics (2008)
5. Saga, R., Tsuji, H.: Improved Keyword Extraction by Separation into Multiple Document Sets According to Time Series. In: HCII, CCIS 374, pp. 450–453 (2013)
6. Church, K., Gale, W.: Inverse Document Frequency (IDF): A Measure of Deviations from Poisson. In: Proceedings of the Third Workshop on Very Large Corpora, pp. 121–130 (1995)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
8. McCallum, K.A.: MALLET, http://mallet.cs.umass.edu