Engineering World
Open Access Journal
ISSN: 2692-5079

Web Search Engine Based Naming Procedure for Independent Topic

Takahiro Nishigaki, Takashi Onoda

Abstract—In recent years, the number of document data has been increasing since the spread of the Internet. Many methods have been studied for extracting topics from large document data. We proposed Independent Topic Analysis (ITA) to extract topics independent of each other from large document data such as newspaper data. ITA is a method for extracting the independent topics from the document data by using the Independent Component Analysis. The topic represented by ITA is represented by a set of words. However, the set of words is quite different from the topics the user imagines. For example, the top five words with high independence of a topic are as follows. Topic1 = {"scor", "game", "lead", "quarter", "rebound"}. This Topic 1 is considered to represent the topic of "SPORTS". This topic name "SPORTS" has to be attached by the user. ITA cannot name topics. Therefore, in this research, we propose a method to obtain topics easy for people to understand by using the web search engine, topics given by the set of words given by independent topic analysis. In particular, we search a set of topical words, and the title of the homepage of the search result is taken as the topic name. And we also use the proposed method for some data and verify its effectiveness.

Keywords—Independent topic analysis, topic extraction, topic naming, web search engine.

I. Introduction

THERE are a lot of studies of topic extraction from a large amount of document data. In this paper, we focus on topic extraction which is one of the challenges of text mining. The topic is the information represented by a co-occurrence of words between the large number of documents. As a method of topic extraction, there are a lot of studies of topic model such as PLSA (Probabilistic Latent Semantic Analysis) [4] proposed by Hofmann et al. and LDA (Latent Dirichlet Allocation) [1] proposed by Blei et al.. The topic model is a method of extracting the topic focusing on probabilistic generative model[2]. The topic model generates the model between the document, word and topic. The topic is latent variable in the topic model. Moreover, the topic model represents the bias of the topic with the document by defining the distribution of the topic for each document. In many of the topic model, a document is represented as the bag of its words (bag-of-words), disregarding grammar and even word order but keeping multiplicity. In the bag-of-words, it commonly used the frequency of the words or tf-idf[8]. It is possible to represent the probability model of the relationship between the document, the word and the topic by using the topic model. However, the PLSA and the LDA is not focus on the relation between topics as correlation relationship or independence.

T. Nishigaki is Department of Industrial and Systems Engineering, Aoyama Gakuin University, Kanagawa, Japan (e-mail: nishigaki@ntt.dis.titech.ac.jp).
T. Onoda is Department of Industrial and Systems Engineering, Aoyama Gakuin University, Kanagawa, Japan (e-mail: onoda@ise.aoyama.ac.jp).

On the other hand, LSI (Latent Semantic Indexing)[3] and ITA (Independent Topic Analysis)[9] are focus on the relation between topics. LSI is able to extract the topic as large variance of the word or the document. It is possible to extract the topic without the correlation relationship between each topic by using LSI. The topic without the correlation relationship between each topic shows that there is no linear relationship such as a topic increase as well as another topic increase. ITA is able to extract the highly independent topic by using ICA (Independent Component Analysis)[6]. ICA is a computational method for separating a multivariate signal into additive subcomponents in signal processing. The highly independent topic show the mutual information between each topic is small topic. It is easy to make document summarization with a large of information by extracting highly independent topic. The independent topic contains the topic without the correlation relationship. In other words, the independent topic is not equal to the topic without the correlation relationship. Note that when the topic in the document is normal distribution, the independent topic is equal to the topic without the correlation relationship. However, the topic in actual document data is not necessarily the normal distribution. Thus, it is necessary to extract the independent topic for extracting the topic which is small mutual information.

In this paper, we describe the ITA which is topic extraction method focusing on the independence of the topics. However, the extracted independent topic are expressed the collection of important words. So it is difficult to understand the independent topics. In this study, we propose a method that help to naming independent topic. We propose method that help to naming independent topic with Web Search Engine.

In the following sections, we introduce the ITA (Independent Topic Analysis) in Section II. In section III, describe the proposed method's concept and algorithm. In Section IV and Section V, we show the experimental setup and experimental result. Finally in Section VI, we describe the conclusion and future works.

II. ITA: Independent Topic Analysis

In this section, we introduce the Independent Topic Analysis (ITA) [10]. This method extract to topics from the document data by Independent Component Analysis (ICA)[6]. In the followings that a small letter expresses scalar, a bold small letter expresses vector, and a bold capital letter expresses matrices. As common variables, $t \in \{1, \ldots, k\}$ express topic variables, $d \in \{1, \ldots, n\}$ express document variables, and $w \in \{1, \ldots, m\}$ express word variables.

ISSN: 2692-5079

Firstly, we describe the concepts of ITA. Matrices $\mathbf{V}$ have m rows and k columns, which called "importance of the word w in the topic t". And vector $\mathbf{v}_t$ represent the t-th columns of vector of matrices $\mathbf{V}$. The vector $\mathbf{v}_t$ is $(v_{1,t}, \ldots, v_{m,t})^{\mathrm{T}}$. Vector $\mathbf{v}_w^{\mathrm{T}}$ represent the transposition of the w-th rows of vector of matrices $\mathbf{V}$. The vector $\mathbf{v}_w$ is $(v_{w,1}, \ldots, v_{w,k})$. Matrices $\mathbf{U}$ have n rows and k columns, which called "importance of the document d in the topic t". And vector $\mathbf{u}_t$ represent the t-th columns of vector of matrices $\mathbf{U}$. The vector $\mathbf{u}_t$ is $(u_{1,t}, \ldots, u_{n,t})^{\mathrm{T}}$. Vector $\mathbf{u}_d^{\mathrm{T}}$ represent the transposition of the d-th rows of vector of matrices $\mathbf{U}$. The vector $\mathbf{u}_d^{\mathrm{T}}$ is $(u_{d,1}, \ldots, u_{d,k})$. In the same way, matrices $\mathbf{A}$ have n rows and m columns, which called "frequency of word w in a document d". And $\mathbf{a}_w$ represent the w-th columns of vector of matrices $\mathbf{A}$. The vector $\mathbf{a}_w$ is $(a_{1,w}, \ldots, a_{n,w})^{\mathrm{T}}$. $\mathbf{a}_d^{\mathrm{T}}$ represent the transposition of the d-th rows of vector of matrices $\mathbf{A}$. The vecotr $\mathbf{a}_d^{\mathrm{T}}$ is $(a_{d,1}, \ldots, a_{d,m})$.

We use the kurtosis of fourth moment of the standard score as the measure of topic of independence. We define the measure of topic of independence as follows.

$$\sum_{w}^{m} \left(v_{w,t}^4 P(w) \right) - 3 \left(\sum_{w}^{m} v_{w,t}^2 P(w) \right)^2$$

Where $v_{w,t}$ is component of w-th rows and t-th columns of the matrix $\mathbf{V}$. And $P(w)$ is defined as follows.

$$P(w) \equiv \frac{\sum_d^n a_{d,w}}{\sum_{d,w}^{n,m} a_{d,w}}$$

Where $a_{d,w}$ is component of d-th rows and w-th columns of the matrix $\mathbf{A}$. When this measure is large, many components of matrix $\mathbf{V}$ and $\mathbf{U}$ become 0 value. So it is possible to express the topic only in a small number of words and documents.

Secondly, we describe the algorithm of ITA. ITA is formulated as an optimization problem as follows.

$$\underset{\mathbf{R}}{\text{maximize}} \quad \left| \sum_{t}^{k} \left\{ \sum_{w}^{m} \left((\mathbf{VR}).^4 P(w) \right) - 3 \left(\sum_{w}^{m} (\mathbf{VR}).^2 P(w) \right)^2 \right\} \right|$$

$$\text{subject to} \quad \mathbf{R}^{\mathrm{T}}\mathbf{R} = 1, \quad \|\mathbf{R}\| = 1$$

Where $(\mathbf{VR}).^4$ is fourth power of each component of the matrix $\mathbf{VR}$. In follow, we show algorithm of ITA. In this method, the number of topics k is random variable.

1) Get the matrix $\mathbf{A}$. And using by [5], we normalize the $\mathbf{A}$ to make $\tilde{\mathbf{A}}$.
2) Performs a singular value decomposition of the matrix $\tilde{\mathbf{A}}$, such that $\hat{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{A}}\hat{\mathbf{V}} = \hat{\mathbf{S}}$. Where $\hat{\mathbf{S}}$ is a diagonal matrix of singular values.
3) Extrac the matrix $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{V}$ from the matrix $\hat{\mathbf{U}}$, $\hat{\mathbf{S}}$ and $\hat{\mathbf{V}}$. Extracted by k components in descending order of the value of the matrix $\hat{\mathbf{S}}$.
4) The matrix $\mathbf{X}$ of the topic in the k-dimensional space is defined as follows.

$$\mathbf{X} = \mathbf{S}^{-1/2}\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{A}}$$

5) Independence maximization between each topic: calculate the rotation matrix $\mathbf{R}$ of the maximum independence based on FPIC as follows.

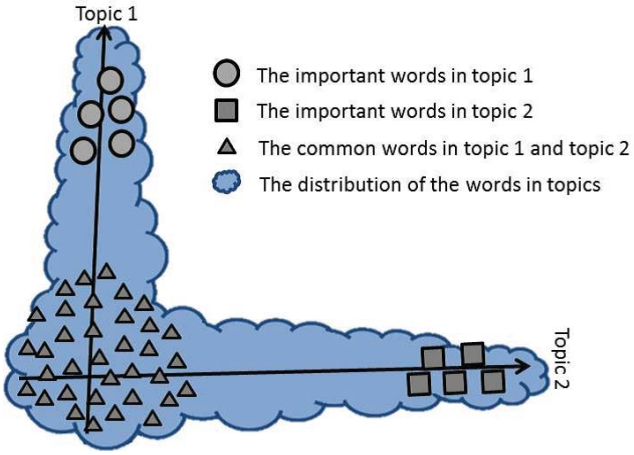

Fig. 1 The image of ITA

a) Initialize the $\mathbf{R}$ to $k \times k$ zero matrix.

$$\mathbf{R} = \mathbf{0}$$

b) Substitute t-th column vector of the $\mathbf{R}$ to t-th column vector $\mathbf{e}_t$ of identity matrix $\mathbf{I} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k)$

$$\mathbf{r}_t = \mathbf{e}_t$$

c) Initialize the $\mathbf{r}^{(old)}$ to $k \times 1$ zero vector.

$$\mathbf{r}^{(old)} = (0, 0, \ldots, 0)^{\mathrm{T}}$$

d) Update the $\mathbf{r}_t$ as follows.

$$\mathbf{r}^{(old)} = \mathbf{r}_t, \quad \mathbf{r}_t = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{r}_t).^3 - 3\mathbf{r}_t$$
$$\mathbf{r}_t = \mathbf{r}_t - \mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{r}_t, \quad \mathbf{r}_t = \mathbf{r}_t / \|\mathbf{r}_t\|$$

Where $(\mathbf{X}^{\mathrm{T}}\mathbf{r}_t).^3$ is cube of components of the $\mathbf{X}^{\mathrm{T}}\mathbf{r}_t$.

e) If $\mathbf{r}$ is convergence under the same conditions as FPICA[5], go to Step (5f). Otherwise go to Step (5d).
f) If $t < k$, increasing one t and go to Step (5b). If $t = k$, go to Step (6).

6) Calculate the ${}^*\mathbf{V}$ and the ${}^*\mathbf{U}$ as follows.

$${}^*\mathbf{V} = \mathbf{VR}, \quad {}^*\mathbf{U} = \mathbf{UR}$$

Thus, it is possible to extract highly independent topics, such as shown in Fig. 1. We apply ITA to Los Angeles Times (LA Times) data as benchmark data. In the Table I, it is shows that the important words of extracted 6 topics by applying ITA to LA Times. The important words which are the word with the large value of $v_{w,t}$.

However, it is difficult to understand the independent topics. Because we can only look the collection of important words of each independent topic. So, we want to name of each independent topics. Therefore, we propose a method of naming the independent topic with Web search engine. In Section III, we describe the proposed method.

ISSN: 2692-5079

TABLE I
THE IMPORTANT WORDS OF THE EXTRACTED 6 TOPICS BY APPLYING ITA TO LA TIMES

No. topics	important words of each topic				
	$w = 1$	$w = 2$	$w = 3$	$w = 4$	$w = 5$
1	scor	game	lead	quarter	rebound
2	bush	polic	budget	reagan	presid
3	million	earn	bank	quarter	billion
4	soviet	israel	afghanistan	foreign	afghan
5	aleen	macmin	art	entertain	report
6	polic	counti	offic	orang	game
	$w = 6$	$w = 7$	$w = 8$	$w = 9$	$w = 10$
1	league	goal	fullerton	half	victori
2	senat	tower	earn	congress	million
3	compani	rose	revenu	corpor	stock
4	israe	militari	datelin	counti	palestinian
5	morn	nation	intern	new	film
6	arrest	citi	car	team	bowl

TABLE II
THE IMPORTANT WORDS OF THE EXTRACTED 9 TOPICS BY APPLYING ITA TO KOS BLOG

No. topics	important words of each topic				
	$w = 1$	$w = 2$	$w = 3$	$w = 4$	$w = 5$
1	bush	president	bushs	administration	edwards
2	november	account	electoral	governor	house
3	dean	clark	edwards	primary	gephardt
4	kerry	edwards	john	general	kerrys
5	iraq	war	iraqi	military	troops
6	percept	poil	voters	polls	vote
7	campaign	million	money	edwards	media
8	party	democratic	democrats	votes	republican
9	race	senate	democrats	state	republican
topics	$w = 6$	$w = 7$	$w = 8$	$w = 9$	$w = 10$
1	george	states	kerry	delay	people
2	polls	poll	repubicans	senate	vote
3	democratic	iowa	lieberman	campaign	republican
4	administration	percent	voters	polls	debate
5	american	saddam	soldiers	threat	forces
6	democrats	number	race	lead	results
7	race	poll	election	democrats	senate
8	house	republicans	states	election	vote
9	seat	house	gop	district	democtrat

TABLE III
THE IMPORTANT WORDS OF THE EXTRACTED 4 TOPICS BY APPLYING ITA TO NIPS

No. topics	important words of each topic				
	$w = 1$	$w = 2$	$w = 3$	$w = 4$	$w = 5$
1	netwark	unit	neural	input	training
2	model	function	data	system	weight
3	learning	algorithm	action	data	control
4	neuron	cell	input	function	training
	$w = 6$	$w = 7$	$w = 8$	$w = 9$	$w = 10$
1	output	hidden	weight	layer	algorithm
2	control	object	parameter	recognition	algorithm
3	task	reinforcement	policy	function	system
4	data	visual	circuit	synaptic	response

III. NAMING TOPICS WITH WEB SEARCH ENGINE

In this section, we describe proposed method that help to naming topics with Web search engine.

We exlpain the concepts of proposed method. First step, it extract the important words of independent topics. Second step, it perform an AND search using the extracted words. Next step, it extract the Title from the Web page of the search result.

We describe the algorithm of proposed method as follows.

1) Get the k independent topics by ITA.
2) Extract the p important words of each independent topic with high $\mathbf{v}_{w,t}, t \in 1\ldots,k$.
3) Using of googlesearch[15] library in Python, perform an AND search with the extracted words by Step 2.
4) Extract the Title from HTML of the search result first Web page by Step 3.
5) Let the title be the name of that topic.
6) Perform Step 2, 3, 4 with all independent topics.

IV. EXPERIMENTAL SET UP

In this section, we apply proposed method to three benchmark dataset as follows.

- Los Angeles Times (LA Times)[13], [14]: the number of document × words is 6279×31472, the number of topic is 6
- Neural Information Processing Systems (NIPS)[7]: the number of document × words is 1500×12419, the number of topic if 4
- DAILY KOS blog (KOS blog)[7]: the number of document × words is 3430×6906, the number of topic is 9

In the experiments, the number of words to be extracted and searched is $p = 10$.

V. EXPERIMENTAL RESULTS & DISCUSSION

In this section, we show the experimental results of applying proposed method to the benchmark data, and we discuss it. In the experiment, we perform an AND search with 10 important words of independent topics. We show some of the words used in the AND search in the Table. I and III. And we use googlesearch[15] in Python for AND search.

Firstly, we explain the result of proposed method apply to LA Times. We show that the extracted Title from HTML of the search result applying proposed method to LA Times.

Topic1 Philadelphia 76ers 2019 Statistics - Team and Player Stats - ESPN
Topic2 THE PRESIDENT'S BUDGET; Transcript of President Bush's Message to Congress on His Budget Proposal - The New York Times
Topic3 Citigroup earnings Q3 2018 beat expectations
Topic4 The Soviet War in Afghanistan 1979 - 1989 - The Atlantic
Topic5 AceRec/newdocs.dat at master - mickeystroller/AceRec - GitHub
Topic6 Crime - The Buffalo News

Looking at the titles, we find that Topic1 is a topic of NBA. Similarly, we find that Topic2 is a economic policies of the former U.S. president Tonald Reagan, Topic3 is the earn and finance, Topic4 is a foregin topic for the U.S., Topic5 is GitHub, Topic6 is incident near Los Angeles. Topic5 is very difficult to understand. So, we access this web page, we are able to learn that the details of this LA Times dataset. From this results, the extracte title is easier to understand than a collection of words.

Secondly, we explain the result of proposed method apply

ISSN: 2692-5079

to NIPS. We show that the extracted Title from HTML of the search result applying proposed method to NIPS.

Topic1 Mind: How to Build a Neural Network (Part One)
Topic2 Build your own object classification model in SageMaker and import it to DeepLens | AWS Machine Learning Blog
Topic3 Reinforcement learning - Wikipedia
Topic4 Synapses and Memory Storage

Looking at the titles, we find that Topic1 is a topic of neural network. Similarly, we find that Topic2 is classification model, Topic3 is reinforcement learning, Topic4 is Brain Synapse.

The result of NIPS, as well as the results of LA Times, the extracte title is easier to understand than a collection of words.

Thirdly, we explain the result of proposed method apply to KOS Blog. We show that the extracted Title from HTML of the search result applying proposed method to KOS Blog.

Topic1 Presidency of George W. Bush - Wikipedia
Topic2 Ben Shapiro: 10 Reasons To Vote Republican In November | Video | RealClearPolitics
Topic3 CNN.com - Edwards campaign book details attacks - Jan. 21, 2004
Topic4 Poll: Bush And Kerry Tied - CBS News
Topic5 U.S. SENDS FORCE AS IRAQI SOLDIERS THREATEN KUWAIT - The New York Times
Topic6 Democrats Have Numbers on Their Side in Battle for the House. Republicans Have the Map. - The New York Times
Topic7 Election Countdown: Latest on the 2018 Senate money race | Red-state Dems feeling the heat over Kavanaugh | Dem doubts about Warren | Ocasio-Cortez to visit Capitol Hill | Why Puerto Ricans in Florida could swing Senate race | TheHill
Topic8 Democrats secure 218 seats in midterms to win control of House - as it happened | US news | The Guardian
Topic9 Democrats Capture Control of House; G.O.P. Holds Senate - The New York Times

Looking at the titles, we find that Topic1 is a topic of George W. Bush. Similarly, we extracted the title from Web page.

The result of KOS Blog, as well as the results of LA Times, the extracte title is easier to understand than a collection of words.

In these experiments, we tried searching for the number of words by $p = 10$. However, we don't know the appropriate number of words easy to understand.

VI. Conclusion

In this paper, we proposed method that help to name the independent topic using the Search Engine. This methods was to solve one of the problems of the Independent Topic Analysis (ITA). The problem is that it is difficult to understand the independent topic from collection of words. To evaluate the proposed method, we implemented and tested in Python. We applied proposed method to several benchmark datasets. This experimental results show the extracted topics are easier to understand than a collection of words. However, we don't know the appropriate number of words easy to understand. And when HTML style are different, we don't understand of title tag.

For our future works, we would like to improve these problem. And it is necessary to apply other benchmark datasets.

References

[1] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation, *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.
[2] Blei, D. M. 2012. Probabilistic topic models, *Commun. ACM*, Vol. 55, No. 4, pp. 77–84.
[3] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. Indexing by latent semantic analysis, *Journal of the American Society of Information Science*, Vol. 41, No. 6, pp. 391–407.
[4] Hofmann, T. 1999. Probabilistic latent semantic analysis, *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99)*, pp. 289–29, Morgan Kaufmann Publishers Inc..
[5] Hyvärinen A. 1999. Fast and robust fixed-point algorithms for independent component analysis, *IEEE Trans. on Neural Networks*, Vol. 10, No. 3.
[6] Hyvärinen, A., Karhunen, J. and Oja, E. 2001. Independent component analysis, John Wiley & Sons.
[7] Lichman, M. 2013. UCI machine learning repository, http://archive.ics.uci.edu/ml , Accessed on 11/11/2016.
[8] Salton, G., Fox, E. A., Wu, H. 1983. Extended boolean information retrieval, *Commun. ACM*, Vol. 26, No. 11, pp. 1022–1036.
[9] Shinohara, Y. 1999. Independent Topic Analysis : Extraction of Characteristic Topics by maximization of Independence, Technical report of IEICE.
[10] Shinohara, Y. 2000. Development of Browsing Assistance System for finding Primary Topics and Tracking their Changes in a Document Database, CRIEPI Research Report.
[11] Sirovich, I., and Kirby, M., 1987. Low-Dimensional procedure for the caracterization of human faces, *Journal of Optical Society of America A*, Vol.4, No.3, pp.519–524.
[12] Tanaka, M, Shinohara, Y. 2003. Topic-Based Dynamic Document Management System for discovering Important and New Topics, CRIEPI Research Report.
[13] Zhao, Y. and Karypis, G. 2002. Evaluation of hierarchical clustering algorithms for document datasets, *Conference of Information and Knowledge Management (CIKM)*, pp. 515–524, ACM.
[14] Zhong, S., and Ghosh, J. 2003. A comparative study of generative models for document clustering, *Data Mining Workshop on Clustering High Dimensional Data and Its Applications*.
[15] google-search 1.0.2, https://pypi.org/project/google-search/, 2018/11/15

Takahiro Nishigaki Takahiro Nishigaki was born in Japan 1987. He recieved the B. E. Degree from Kyoto Institute of Technology in 2011. He recieved the M. E. Degree from Tokyo Institute of Technology in 2013. He recieved the Dr. Eng Degree from Tokyo Institute of Technology in 2017. He is a member of the Japanese Society of Artficial Intelligence (JSAI). He has been a Assistant Professor with Aoyama Gakuin University since 2017.

Takashi Onoda Takashi Onoda was born in Japan in 1962. He received the B. S. Degree from International Christian University in 1986. He received the M. S. Degree from Tokyo Institute of Technology in 1988. He received the Dr. Eng. Degree from University of Tokyo in 2000. He was a Visiting Researcher in GMD FIRST (Fraunfoher FIRST) from 1997 to 1998. Since 2007, he has also been an Visiting Professor at Department of Computational Intelligence and System Science, Tokyo Institute of Technology, Japan. He is a member of JSAI (Japanese Society of Artficial Intelligence). He researched in Central Research Institute of Electric Power Industry from 1988 to 2015. He has been a Professor with Aoyama Gakuin University since 2016.