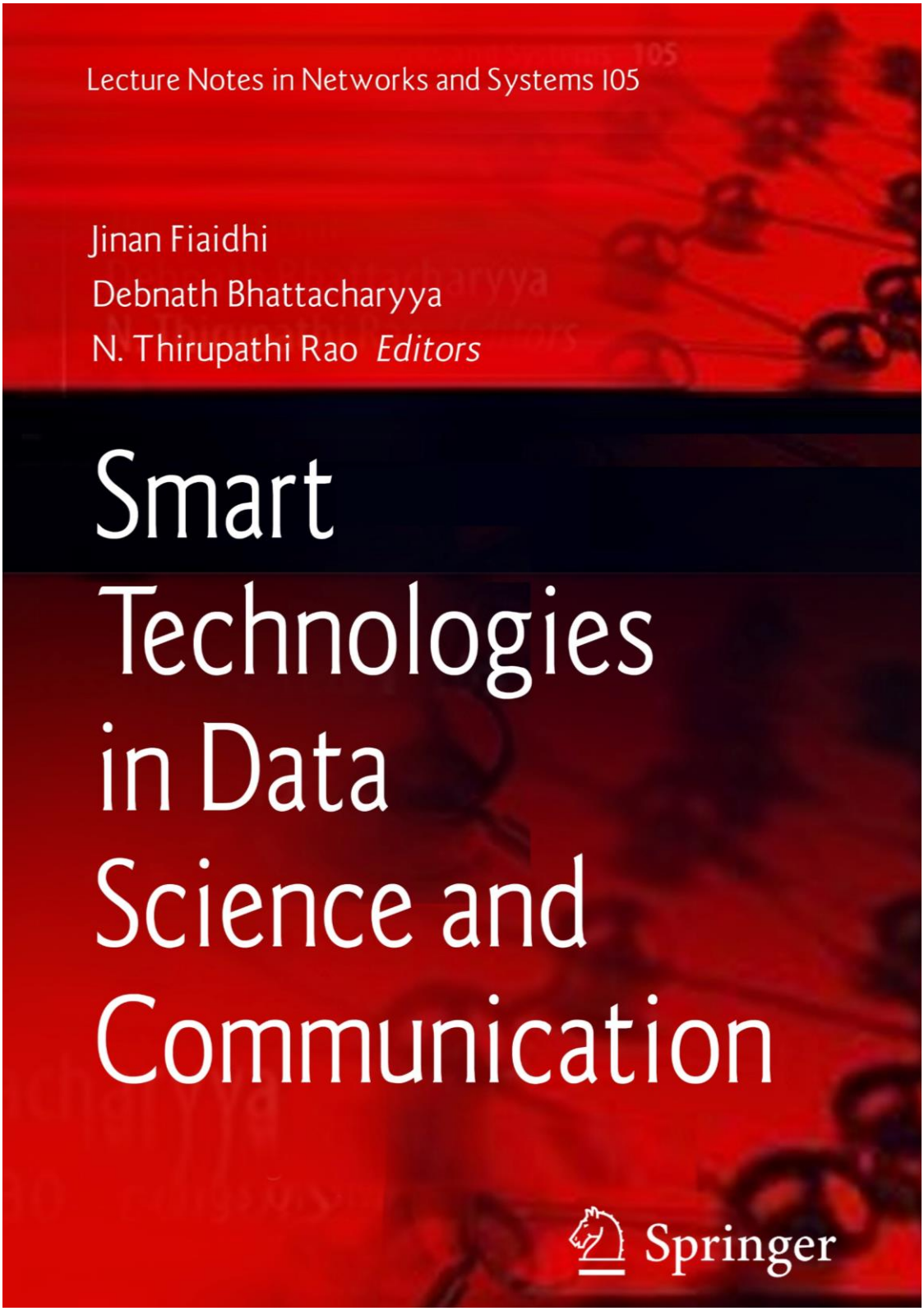
Lecture Notes in Networks and Systems 105
Jinan Fiaidhi
Debnath Bhattacharyya
N. Thirupathi Rao Editors
Smart Technologies in Data Science and Communication
Springer

About the Book:

This book features high-quality, peer-reviewed research papers presented at the International Conference on Smart Technologies in Data Science and Communication (Smart-DSC 2019), held at Vignan's Institute of Information Technology (Autonomous), Visakhapatnam, Andhra Pradesh, India on 13–14 December 2019. It includes innovative and novel contributions in the areas of data analytics, communication and soft computing.

Springer

ISBN: 978-981-15-2406-6

Lecture Notes in Networks and Systems

Volume 105

Editors
Jinan Fiaidhi
Department of Computer Science
Lakehead University
Thunder Bay, ON, Canada

N. Thirupathi Rao
Department of Computer Science
and Engineering
Vignan's Institute of Information
Technology
Visakhapatnam, Andhra Pradesh, India

Debnath Bhattacharyya
Department of Computer Science
and Engineering
Vignan's Institute of Information
Technology
Visakhapatnam, Andhra Pradesh, India

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-15-2406-6 ISBN 978-981-15-2407-3 (eBook)
https://doi.org/10.1007/978-981-15-2407-3

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

LDA Topic Generalization on Museum Collections

Zeinab Shahbazi and Yung-Cheol Byun

1 Introduction

Text segmentation is one of the important sections in natural language processing. The meaning of segmentation is dividing the document into coherent parts that can be based on words/sentences or paragraphs and finds the similarity between these materials. It prepares the document structure to be useful in text summarization and information extraction [1, 2]. Totally, segmentation is used for removing the repetitive or same meaning sections to get more information from the topics, and it is famous for topic generalization. In this paper, we proposed topic modeling method by using linear regression machine learning algorithms to solve the problem of similarity and find the correspondence between museums in different countries. The dataset which we used through this process is the total information about museums and visitor's comments.

2 Related Works

Try to find information through the document is the first way that comes as an idea to human being which is mostly coming from searching between sentences, but it can make it more complex to detect topics out of that material. To come out of this challenge, proposing machine reader system which is known as SECTOR is defined that it segments the text into relevant parts [3] (SECTOR: a neural model for coherent topic segmentation and classification).

Z. Shahbazi · Y.-C. Byun (✉)
Jeju National University, Jeju City 63243, South Korea
e-mail: yungcheolbyun@gmail.com

Z. Shahbazi
e-mail: z.shahbazi72@gmail.com

J. Fiaidhi et al. (eds.), *Smart Technologies in Data Science and Communication*, Lecture Notes in Networks and Systems 105, https://doi.org/10.1007/978-981-15-2407-3_12

Recently, technologies of speech become one of the famous areas in monitoring and it contains high output for user consent. This output is because of analyzing dialogues to catch customer problems and the ways to overcome them (multiple topic identification in telephone conversations).

It is to analyze the structure of document topics or relationship between speeches to show the goal of the speech such as people discussion or customer service dialogues is a significant part to figure out the dialogue, generalize it, and also summarize it (a weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning) [4].

In late studies in natural language processing, supervised learning has a problem of searching in large amount and labeled data compared with unsupervised learning. To do this, adjusting text segmentation as a document label for every sentence until it ends with segment and contains rather than 727,000 texts from Wikipedia (text segmentation as a supervised learning task) [5].

3 Methodology

In this part, we are trying to present LDA model (latent Dirichlet allocation) to find the museums' activities per year and also find the similarity between different museums. To do this, the proposed method is divided into four main parts: preprocessing, text categorization, topic modeling, and finally similarity measurement between topics. Figure 1 shows the proposed methodology in detail.

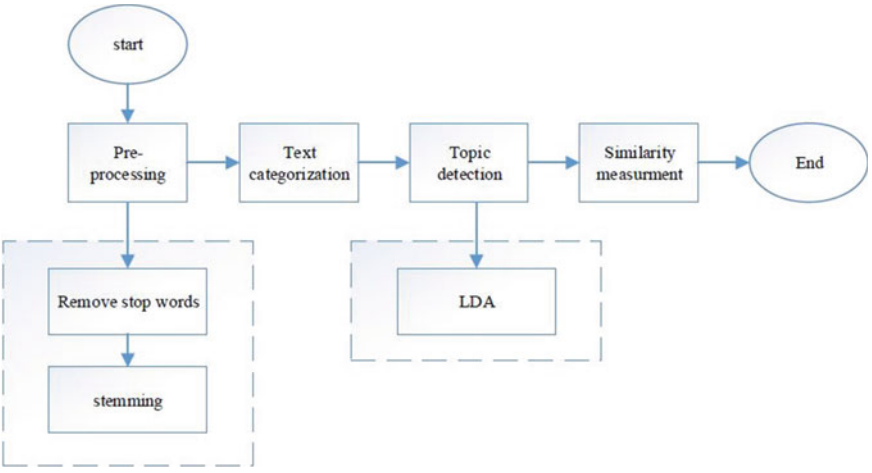

Fig. 1 Proposed system architecture

3.1 Preprocessing

Preprocessing is one of the important steps in the machine learning algorithms. To start the preprocessing section, stop word removal and stemming are used for normalizing the data. Analyzing data can cause problem if the process not done carefully. Preprocessing steps start with extracting terms, transforming, cleaning the data, and end with loading the data.

3.2 Text Categorization

Text categorization is part of natural language processing tasks which is also known as text classification. In this section, unique words extracted from unstructured data types which in this procedure are text documents. Text categorization has two ways that process for extracting information: First step is manually and second is automatically. It usually shows the result based on good quality, but it can be time-consuming. In this paper, we used automatically categorization by applying machine learning ("K-means" and "linear regression") algorithms. The first step toward training the algorithms is feature extraction which converts the text into numerical representation based on vectors. Figure 2 shows the categorization process in detail.

3.3 Topic Modeling

Topic modeling and topic classification are the most popular approaches in the machine learning system and natural language processing. Topic modeling is used to extract the generic information from the document. The proposed system is to recognize the most repeated topics and generalize them into more limited categories based on number of visitors.

3.4 Similarity Measurement

To continue the process of topic detection by using similarity measurement, extracted topics, generalized, and limited them into less topic numbers by categorizing the same meaning topics in one general sub-topic. In case of museum, a comparison of different museums is detected and as a last step recognition between similar topics is processed. Figure 3 represents the similarities between different kinds of museums.

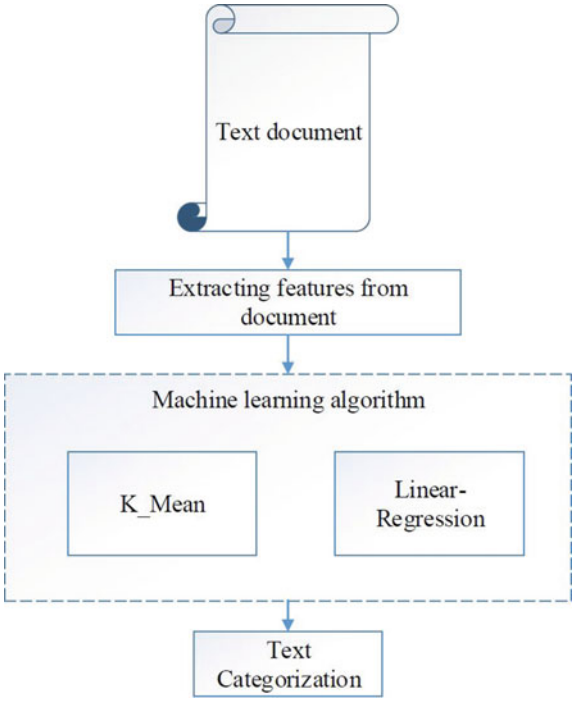

Fig. 2 Text categorization process

4 Experiments

To test the proposed methodology, trip advisor museum reviews used as input dataset which contains 1013 museum information such as museum description, address, entrance fee, and maximum time of visit. This section is to show the detailed results and information about topic generalization between museums.

4.1 Dataset

This dataset is created based on trip advisor information and visitor's reviews. It contains different types of museums, e.g., art, war, history, and car; based on this information, museums activities per year and topic similarity are calculated. Figure 4 shows the comparison between museums that rely on their activities per year and number of visitors.

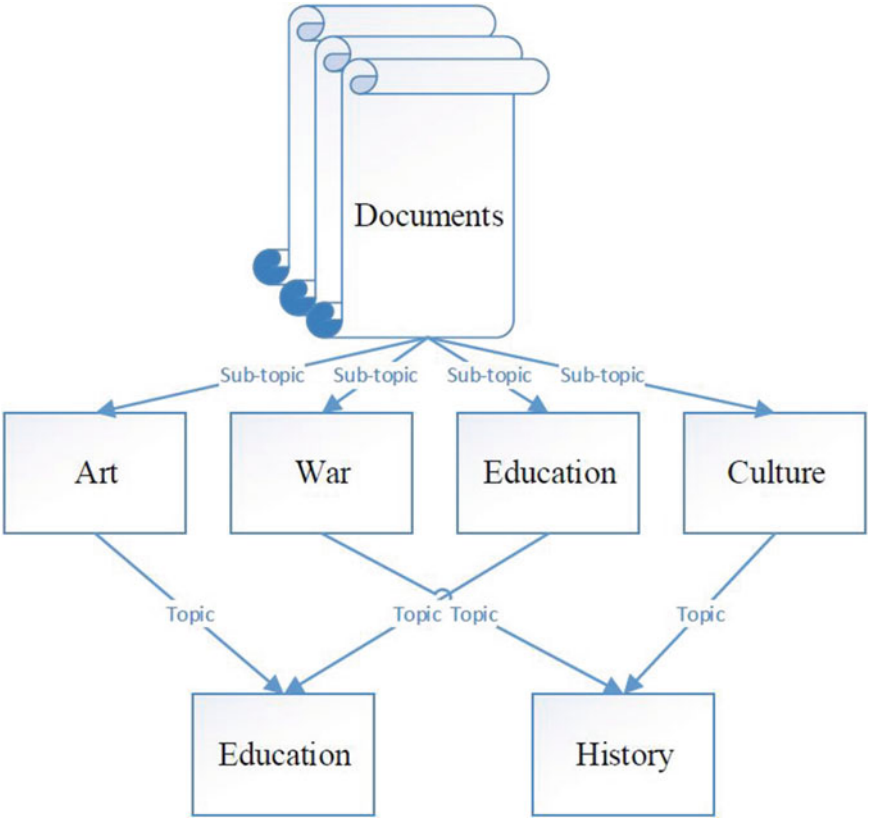

Fig. 3 Preview of topic generalization

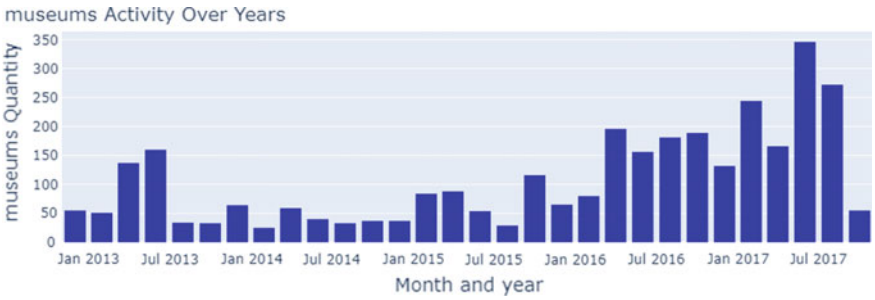

Fig. 4 Yearly preview of museums

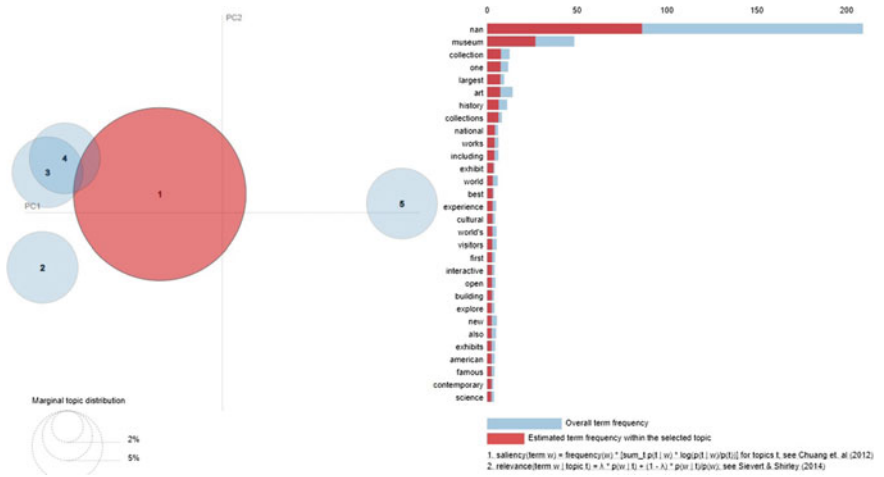

Fig. 5 Representation of topic similarity

4.2 LDA Topic Modeling

Topic modeling is a part of unsupervised natural language processing which display document based on related topics. To automatically organize, understand, search, and summarize large type of texts, topic modeling system has proposed some steps:

1. Finding the hidden part of document.
2. Divide the document based on that information.
3. Apply achieved information to organize, summarize, search, and form predictions.

In this part, document is divided into different topics and after doing the similarity process, topics are generalized into more limited parts, and finally, similarity between these topics is shown in Fig. 5. Table 1 describes the number of topics and percentage of the similarity between topics, running online (single-pass) LDA training, 5 topics, 1 pass over the supplied corpus of 3218 documents, updating model once every 2000 documents, evaluating perplexity every 3218 documents, iterating 50x with a convergence threshold of 0.001000.

4.3 Linear Regression Algorithm

Linear regression machine learning algorithm is used to find the relationship between independent and dependent variables and analyses between topics in document.

Table 2 shows the comparison between latent Dirichlet allocation, similarity feature, and linear regression machine learning algorithm which LDA has 4% higher result than linear regression.

Table 1 Merging changes from 2000 documents into a model of 3218 documents

Topic number	Topic words	LDA similarity PA
1	Museum, history, collection, art, one	0.6
2	Nan, museum, art, one, collection	0.37
3	Museum, history, collection, one	0.2
4	Nan, museum, collection, one, largest	0.4
5	Museum, art, new, collection, one	0.16
Topic number	Topic words	LDA similarity PA

Table 2 Comparison between LDA and linear regression

Algorithms	Topic differences
LDA	0.74
Linear regression	0.70

5 Conclusion

Text segmentation is one of the important aspects of natural language processing which is used for topic modeling. In this paper, we used the combination of LDA similarity feature and linear regression algorithm to find the similarity between museums in different countries. The output represents that proposed system which has 0.4% higher accuracy rate than linear regression one.

Acknowledgements This research was financially supported by the Ministry of SMEs and Startups (MSS), Korea, under the "Regional Specialized Industry Development Program" supervised by the Korea Institute for Advancement of Technology (KIAT).

This work was supported by "Jeju Industry-University Convergence Foundation" funded by the Ministry of Trade, Industry, and Energy (MOTIE, Korea). [Project Name: "Jeju Industry-University convergence Foundation/Project Number: N0002327"].

References

1. F. Author, Article title. Journal **2**(5), 99–110 (2016)
2. F. Author, S. Author, *Title of a Proceedings Paper*, ed. by F. Editor, S. Editor. Conference 2016, vol. 9999 (LNCS, Springer, Heidelberg, 2016), pp. 1–13
3. F. Author, S. Author, T. Author, *Book Title*, 2nd edn. (Publisher, Location, 1999)
4. F. Author, Contribution title. 9th international proceedings on proceedings, Publisher, Location, 1–2 (2010)
5. LNCS Homepage, http://www.springer.com/lncs. Last Accessed 2016/11/21