Implementation of National R&D Report Linking Service with Topic Modeling

Wonjun Choi, Seokjong Lim and Hyekyong Hwang

Korea Institute of Science and Technology Information, Korea

ABSTRACT

National R & D research achievements are nationally funded research outputs. The management and distribution of these research outputs are also enacted by law and have been built and serviced since 2008. Among the research achievements, reports are better managed and distributed as national R & D achievements. However, due to the lack of the concept of publications and private reports, the report remains a challenge to be published and spread in accordance with OA policy. The Korea Institute of Science and Technology Information(KISTI) is a management and distribution service (NDSL, NTIS) that is dedicated to papers and reports among national R & D research achievements. KISTI constructs and services about 200,000 reports (160,000), private reports (40,000) and non-text content (1.4 million tables, 3.4 million figures). Even now, even if it is a private report, after 3 years, it has a system that can be converted into a public service. This study introduces the development of utilization service by applying topic modeling to the national R & D report. Topic modeling methodology uses the popular LDA methodology and supports the analysis of important keywords and clustering services for each author in the report. We created a data dictionary by extracting key keywords from the report document, and developed a service scenario that shows that linking service is possible based on the contents of the report meta and the original text.

INTRODUCTION

The Korean government nominates research performance management and distribution agencies in accordance with the 2008 National R & D Program Management Rules to enhance the management and use of research performance. National R & D research achievements are nationally funded research outputs. The management and distribution of these research outputs are also enacted by law and have been built and serviced since 2008.

List	Total	Open	Closed
R&D Report (text)	**209,533**	166,516	43,017
R&D Report (non-text)	**72,620**	49,163	23,457
Non-text contents	**4,880,683**	3,479,199	1,401,484
- Table	1,473,267	1,146,143	327,124
- Figure	3,407,416	2,333,056	1,074,360

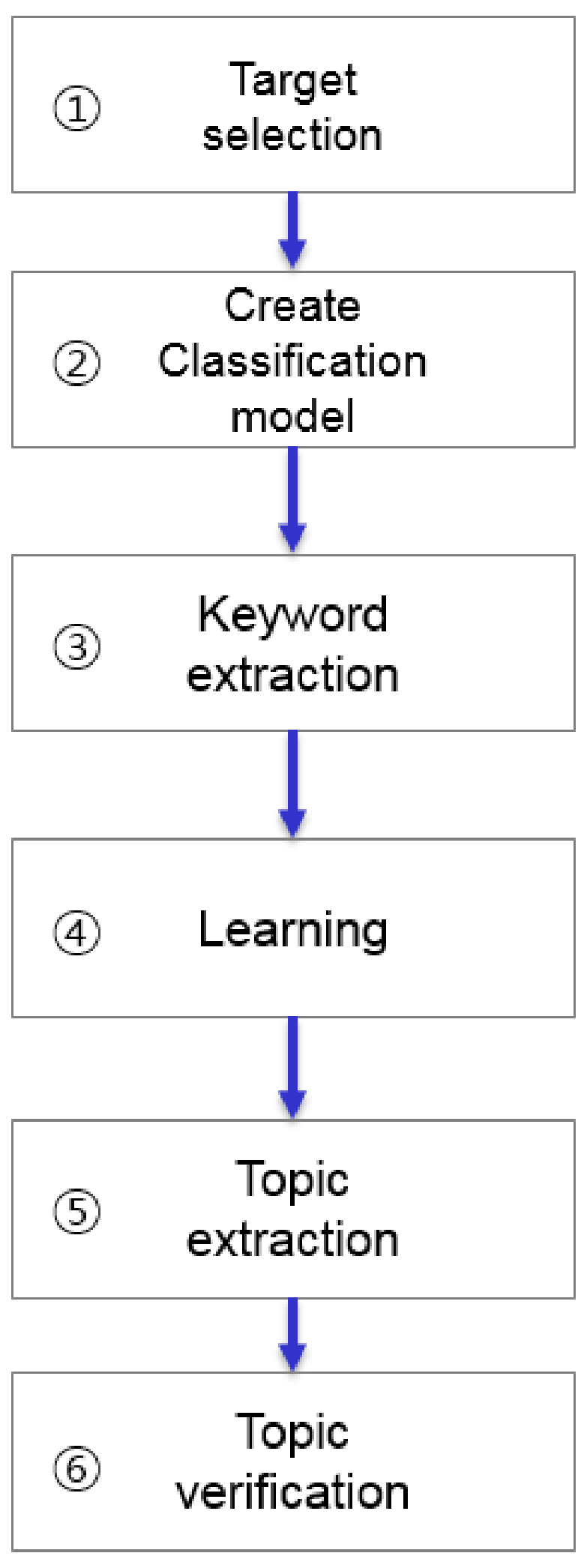
Table 1. Statistical data for R&D report(included non-text, 2019.10)

KISTI constructs and services about 200,000 R&D reports (open: 160,000, closed: 40,000), and non-text content (1.4 million tables, 3.4 million figures) as above Table 1.

As a way of summarizing and explaining the accumulated research reports, it is persuasive to extract the topics in the research reports and show them through linking related information. The purpose of this study is to implement the author-topic service of the research report and to verify its performance by using the LDA topic modeling method. The reason for choosing Latent Dirichlet Allocation(LDA) topic modeling method is that there are many topic extraction methodologies such as LSI and HDP, but Latent Semantic Indexing(LSI) can produce intuitive results, Hierarchical Dirichlet Precess(HDP) is suitable for subdivision, and LDA method is reflected in the model when new topic is derived. The LDA method was chosen because it is easy and flexible to do. Since LDA is an unsupervised learning methodology, a set of correct answers is required for verification. In order to generate the correct answer set, the key words were extracted from the summary of the study report by frequency and ranked. This set of answers is compared with the result of using LDA. Finally, the performance was verified.

ANALYSIS PROCEDURE

Extract Keywords and Topics



① Target : 2017 Research Report (Total 3,413 cases)

② Model: 35 classification models are created
- selected 35 of the total 38 science and technology standard classification codes, where the report

③ Criteria : 35 classification models generated from ②
- Extract target : 2,041 reports (excluding 1,372 without classification code out of 3,413 total)
- Experimental group : full-text report content
- Control group: Summary_Research goal, Summary_Study contents, Korean keyword
- Extraction method : LDA

④ Target : 50 learning progresses with keywords extracted from ③
- Cluster progress by 35 classification models created in ②

⑤ Criteria: Keywords by 35 Classification Models Learned from ④
- Extract target: 2,041 reports (excluding 1,372 without classification code out of 3,413 total)
- Extraction method: LDA

⑥ Topic Verification
- Criteria : Answer Set (based on Korean keywords in each report)

The meaning of topic is the main representative keyword of research report in this paper. Keyword extraction was conducted based on 2,041 (2017) research reports with science and technology standard classification, and 35 classification models were selected to generate models and select the correct answer set. The LDA model was applied by using the summary of the research report and the Korean keyword field. The learning was performed 50 times with extracted keywords and clustered by 35 classification models. In order to extract the topics (major keywords) for each of the 35 classification models, a maximum of 100 rankings were selected from the summary and Hangul keyword fields. In order to verify the topic, Top 20 by field was selected and the answer set and the result data were verified using the similarity comparison method. Keyword extraction was conducted based on 2,041 (2017) research reports with science and technology standard classification, and 35 models were selected to generate models and select the correct answer set.

COMPARISON OF THE RESULTS

	# of answer set T	# of full-text T	Matched # of full-text T	# of abstract T	Matched # of abstract T
1	1	17	0	4	1
2	3	25	1	8	3
3	4	36	0	17	4
4	2	24	0	7	2
5	3	45	1	14	3
7	3	27	0	13	3
8	4	65	0	10	4
10	5	55	0	13	5
11	7	73	0	13	7
12	4	127	0	9	4
13	5	36	0	20	5
14	2	24	0	19	2
15	5	37	0	20	5
16	2	24	0	20	2
17	3	41	0	12	3
18	3	63	0	20	3
19	5	23	0	18	5
20	5	30	0	15	5
...	...	...	...	...	...
2041	5	86	0	20	0
SUM	8962	94099	161	30845	5613

Table 2. Comparison of the results

Recall rate of abstract T	Correct rate of abstract T	Recall rate of full-text T	Correct rate of full-text T
62.63%	18.20%	1.80%	0.17%

: The number
T : Topic

Table 3. Results of recall rate and correct rate

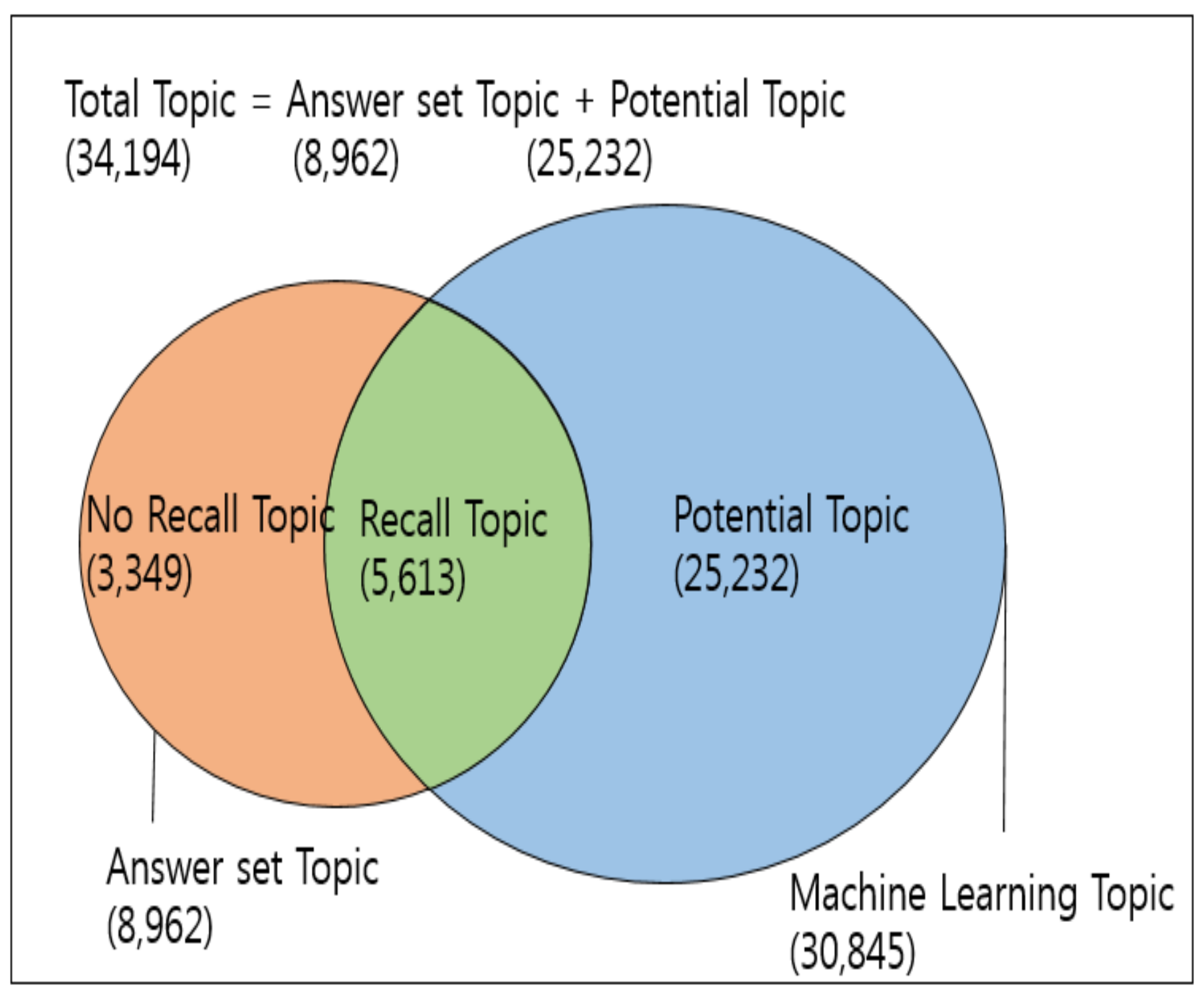
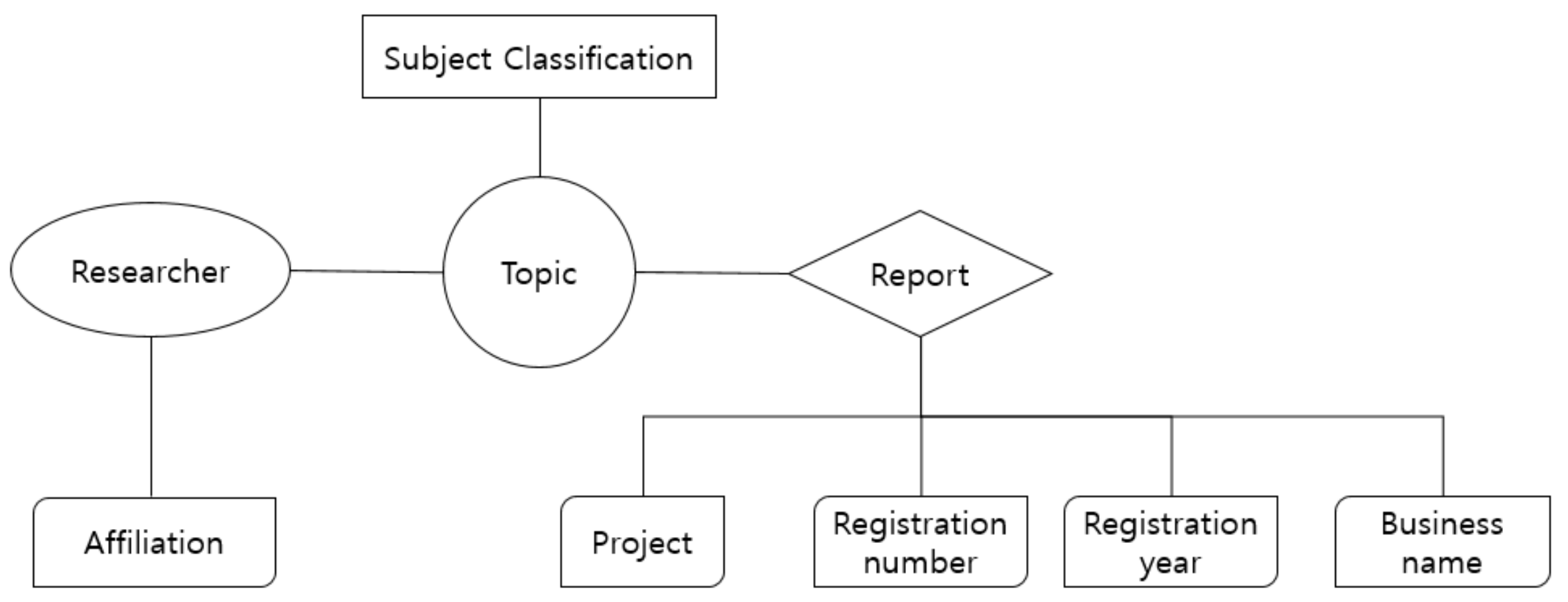


Figure 1. Comparison of abstract topic and answer set topic

The formula used for comparison is: Recall rate = (Number of matched topic / Total number of answer set's topic) * (1 / Total number of report), Correct rate = (Number of matched topic / Total number of extracted topic) * (1 / Total number of report). As shown in the results, the Abstract results are superior to the full-text results. Table. 2 shows the comparison result between topics of abstract and topics of full-text. Recall rate of abstract is higher than full-text result as shown Table. 3. Figure 3. display the comparison of abstract topic and answer set topic



Main topics extracted from national R & D research reports can be linked with thematic classification, can be linked with the researcher, and can be linked to the institution to which the researcher belongs. In addition, by linking reports according to topics, it is possible to link information such as project, registration number, registration year, and business name related to the report, thereby enabling research report linking service.

DISCUSSION

In this paper, the report topic linking service is implemented to facilitate understanding of national R & D research reports. Data dictionary was created by extracting important keywords in the report document, and data was verified by extracting data showing linking service based on the contents of the report meta and the original text. The accuracy of the result data was confirmed to be high and the linking service was implemented to verify that the contents of the report can be checked without reading the contents directly.

REFERENCES

[1] Y. Papanikolaou and G. Tsoumakas, Subset Labeled LDA for Large-Scale Multi-Label Classification (2017, September 16), arXiv.org.

[2] M. Pavlinek and V. Podgorelec, "Text classification method based on self-training and LDA topic models," Expert Systems with Applications, Vol.80, pp.83-93, 2017. http://doi.org/10.1016/j.eswa.2017.03.020

[3] M. Rani, A. K. Dhar, and O. P. Vyas, "Semi-automatic terminology ontology learning based on topic modeling," Engineering Applications of Artificial Intelligence, Vol.63, pp.108-125, 2017. http://doi.org/10.1016/j.engappai.2017.05.006