Thematic Transitions in Data Center Research: A Longitudinal Topic Modeling Study

Kwang-Hoon Kim

kh.kim@kisti.re.kr

Korea Institute of Science and Technology Information (KISTI)

데이터센터 연구의 주제적 전환: 종단적 토픽 모델링 분석

김광훈

Abstract

Data center is an infrastructure facility that collects and manages IT equipment such as servers, storage, and networks in one building, and its importance is emphasized to the extent that the number of data centers is used as an indicator of the country's IT competitiveness. The data center industry plays a pivotal role in the digital economy with advances in advanced technologies such as Chat GPT, cloud computing, and autonomous driving, and is now considered a major infrastructure to support AI technology through high-performance computing resources beyond just communication infrastructure. Despite this importance, systematic time series analysis of what kind of thematic transformation process domestic data center research has gone through is limited. The purpose of this study is to analyze the time series evolution of research topics in various ways, targeting academic papers related to domestic data centers from 2005 to 2025. This study is expected to provide basic data for future data center-related research planning and policy establishment by empirically identifying long-term changes in domestic data center research.

Ⅰ. Introduction

Data centers are a key national infrastructure for revitalizing new industries and is in charge of data storage, processing, and distribution for the realization of services such as AI, next-generation mobile communication, cloud, and IoT. Data traffic is rapidly increasing due to the expansion of the digital economy, and as a result, demand for data centers, which are power-intensive facilities, is also increasing.

Analysis of existing data center research trends has been mainly focused on dictionary classification by researchers or specific technology areas, and due to this, there are limitations in systematically capturing structural changes and potential changes in research topics over a long period of time. In order to compensate for these limitations, topic modeling techniques that can automatically derive latent subject structures from large literature sets have recently been used for research trend analysis. In particular, LDA (Latent Dirichlet Allocation) topic modeling techniques are drawing attention as a way to analyze the multidimensional characteristics and time-series changes of research topics simultaneously by assuming literature as a probabilistic mixture of several topics [1]. This study aims to empirically investigate the long-term evolution of research interests by deriving the latent theme structure inherent in domestic data center research using LDA topic modeling and combining it with time series analysis.

Ⅱ. Data and Methodology

In this paper, LDA topic modeling was performed on 1533 KCI papers related to the data center using only the title, author keyword, and english keyword. If abstracts are included in the process of performing LDA, there is a possibility that the degree of topic aggregation will be slightly widened, so this study derives a research topic structure that directly reflects the author's intention by performing LDA topic modeling using only

title and keyword information except for the thesis abstract. As a result of LDA topic modeling using only titles and author keywords, data center research was divided into six major research topics: cloud service platforms, energy efficiency and cooling infrastructure, network performance, AI/data-intensive workloads, facility design and operation, and location and policy issues. In this paper, based on the results of LDA topic modeling(removal of stop-words), the annual change in the proportion of topics from 2005 to 2025 was analyzed. As a result of time series pattern analysis, Topic 3 tends to have a relatively high proportion in the early stages (2005-2015). This suggests that data center research started from the initial infrastructure problem centered on "network and performance optimization." Since 2015, it has been confirmed that interest has shifted to the "cloud service and platform architecture." It is estimated that this is the time when "cloud platform/service architecture" has become the mainstream frame of data center research. What is interesting here is that in the case of topic 6(Siting, Policy & Sustainability), it showed a high value in the early 2000s, but has been on the decline since 2020. In the early 2000s, it can be speculated that many keywords related to location, regulation, and environmental impact appeared to discuss the introductory period and early norms of the system. However, the decline in the proportion of the topic after 2020 can be interpreted as a result of policy and social issues gradually internalizing into the technology-oriented research context such as energy efficiency and facility operation rather than weakening policy importance. In other words, policy and social issues were initially discussed as independent research topics, but over time, they became internalized into the context of technology, operation, and energy research. Since the mid-2010s, policy or location issues have changed from a full-scale topic to a prerequisite due to the 'backgrounding of the policy'. After 2020, Topic 6 tends to be absorbed into "Energy Efficiency & Cooling (Topic 2)" or "Facility Design & Operation (Topic 5)" due to the "segmentation and internalization" of policy issues.

Ⅲ. Conclusion

As a result of the analysis, it was confirmed that domestic data center research started from the initial network and performance-oriented discussion, went through cloud service and platform architecture, and recently shifted toward energy efficiency and cooling infrastructure.

ACKNOWLEDGMENT

This research was supported by the Korea Institute of Science Technology Information (No. K26L4M2C3-01)

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

Table 1. LDA-based Research Theme Identification

Topics	Theme Identification Results	Top Keywords
Topic 1	Cloud Services & Platform Architecture	Cloud, Service, Virtualization, Resource, Platform, Management
Topic 2	Energy Efficiency & Cooling Infrastructure	Energy, Power, Cooling, Efficiency, PUE(Power Usage Effectiveness), HVAC(Heating, Ventilation, and Air Conditioning)
Topic 3	Network Performance & Scalability	Network, Latency, Traffic, Performance, SDN(Software-Defined Networking)
Topic 4	AI & Data-Intensive Computing	AI, Big Data, Workload, GPU, HPC
Topic 5	Facility Design, Construction & Operation	Facility, Design, Construction, Operation, Management
Topic 6	Siting, Policy & Sustainability	Siting, Location, Policy, Regulation, Carbon, Environment