

## BIBLIOGRAPHIC INFORMATION SYSTEM

**Journal Full Title:** [Journal of Biomedical Research & Environmental Sciences](#)

**Journal NLM Abbreviation:** J Biomed Res Environ Sci

**Journal Website Link:** <https://www.jelsciences.com>

**Journal ISSN:** 2766-2276

**Category:** Multidisciplinary

**Subject Areas:** [Medicine Group](#), [Biology Group](#), [General](#), [Environmental Sciences](#)

**Topics Summation:** 133

**Issue Regularity:** [Monthly](#)

**Review Process:** [Double Blind](#)

**Time to Publication:** 21 Days

**Indexing catalog:** [IndexCopernicus ICV 2022: 88.03](#) | [GoogleScholar](#) | [View more](#)

**Publication fee catalog:** [Visit here](#)

**DOI:** 10.37871 ([CrossRef](#))

**Plagiarism detection software:** [iThenticate](#)

**Managing entity:** USA

**Language:** English

**Research work collecting capability:** Worldwide

**Organized by:** [SciRes Literature LLC](#)

**License:** Open Access by Journal of Biomedical Research & Environmental Sciences is licensed under a Creative Commons Attribution 4.0 International License. Based on a work at SciRes Literature LLC.

Manuscript should be submitted in Word Document (.doc or .docx) through

**Online Submission**

form or can be mailed to [support@jelsciences.com](mailto:support@jelsciences.com)

**IndexCopernicus  
ICV 2022:  
83.03**

 **Vision:** Journal of Biomedical Research & Environmental Sciences main aim is to enhance the importance of science and technology to the scientific community and also to provide an equal opportunity to seek and share ideas to all our researchers and scientists without any barriers to develop their career and helping in their development of discovering the world.

RESEARCH ARTICLE

# A Distributed Representation for Domain Names: An Initial Report

Akihiro Satoh<sup>1\*</sup>, Gen Kitagata<sup>2</sup>, Yutaka Fukuda<sup>1</sup> and Yutaka Nakamura<sup>1</sup>

<sup>1</sup>Kyushu Institute of Technology, Japan

<sup>2</sup>Morioka University, Japan

## Abstract

We propose a distributed representation approach for domain names based on DNS queries. This distributed representation enables domains to be embedded into vector spaces with reflecting data exchange in networks. Since the ground truth of the distributed representation is unknown, we indirectly evaluate our distributed representation based on the premise that the accuracy of the distributed representation is strongly related to the validity of similarity between domains in the distributed representation. The results suggest the feasibility of the concise and versatile representation for numerous domain names with accurately capturing their interrelations.

## Abbreviations

DNS: Domain Name System

## Introduction

The DNS (Domain Name System) is closely related to the network activities of devices because the devices directly interact with the DNS before data exchange via networks [1]. These activities must be monitored by administrators to ensure network security. Recent work has been performed to assess the effectiveness of graph-theoretical techniques in gaining insight into these activities [2,3]. Graph-based approaches represent the relations of domains using vertices and edges to numerically analyze complex structures. However, the main problem is the strict limitation on the acceptable number of domains. This is because the graph scale rapidly increases with the number of unique domains and the graph structure is ultimately represented by an adjacency matrix in which the values become exceedingly sparse and uneven. Our challenge is to establish a concise and versatile representation for numerous domain names with accurately capturing their interrelations, and the results contribute to advances in network security.

The remainder of this paper is organized as follows: In Section 2, we propose a distributed representation approach for domain names based on DNS queries. We describe experiments conducted to analyze the

### \*Corresponding author(s)

**Akihiro Satoh**, Kyushu Institute of Technology, 1-1 Sensuicho, Tobata, Kitakyushu, Fukuoka, 804-8550, Japan


**Email:** [satoh@isc.kyutech.ac.jp](mailto:satoh@isc.kyutech.ac.jp)

**DOI:** [10.37871/jbres2117](https://doi.org/10.37871/jbres2117)

**Submitted:** 19 May 2025

**Accepted:** 08 June 2025

**Published:** 09 June 2025

**Copyright:** © 2025 Satoh A, et al., Distributed under Creative Commons CC-BY 4.0 

**OPEN ACCESS**

### Keywords

- Domain name
- Distributed representation
- Machine learning
- Network security

VOLUME: 6 ISSUE: 6 - JUNE, 2025



Scan Me

**How to cite this article:** Satoh A, Kitagata G, Fukuda Y, Nakamura Y. A Distributed Representation for Domain Names: An Initial Report. J Biomed Res Environ Sci. 2025 Jun 09; 6(6): 642-645. doi: [10.37871/jbres2117](https://doi.org/10.37871/jbres2117), Article ID: JBRES2117, Available at: <https://www.jelsciences.com/articles/jbres2117.pdf>

effectiveness of our approach in Section 3. Finally, we summarize our conclusions and future work in Section 4.

## Methodologies

In this paper, we propose a distributed representation approach for domain names based on DNS queries. This distributed representation enables domains to be embedded into low-dimensional, dense, and continuous vector spaces. Our approach is motivated by the following observation: since several DNS queries occur behind data exchange in a network, the queried domains have strong dependencies; thus, individual domains are indirectly characterized by those relations. Notably, the concept of distributed representations for domain names is not novel [4]. However, the main focus has traditionally been topical classification, which maps “kyutech.ac.jp” to the “education” category and “bbc.com” to the “news and media” category. In contrast, our concern lies in vector representations for domain names reflecting data exchange in networks, and this representation is highly suitable as an input for machine learning and deep learning in network security [5].

Figure 1 shows an overview of our approach. This approach mainly consists of a data preprocessing function and a distributed representation training

function. The following sections detail the two functions.

### Data preprocessing function

A query log for the input of our approach is a record of queries to recursive DNS servers from devices on a network. In the query log, each query has attributes such as a timestamp, source address, queried domain, and record type.

This function divides a query log into query sub-logs. A query sub-log is a set of consecutive queries that have the same source address within a time-interval of  $T_a$  seconds or less. In this division, any query in a query sub-log must satisfy the following conditions:

$$\forall x_i, x_{i+1} \in X_n : f(x_i, x_{i+1}) \leq T_a \text{ and } g(x_i, x_{i+1}).$$

Here,  $x_i$  and  $x_{i+1}$  indicate the  $i$ -th and  $(i+1)$ -th consecutive queries in query sub-log  $X_n$ ;  $f(x_i, x_{i+1})$  is the time-interval between  $x_i$  and  $x_{i+1}$ ; and  $g(x_i, x_{i+1})$  is boolean corresponding to the same or different source address in  $x_i$  and  $x_{i+1}$ . The total  $N$  query sub-logs  $X_n$ , where  $n \in \{1 \dots N\}$ , resulting from this process are passed to subsequent functional units to train the distributed representation.

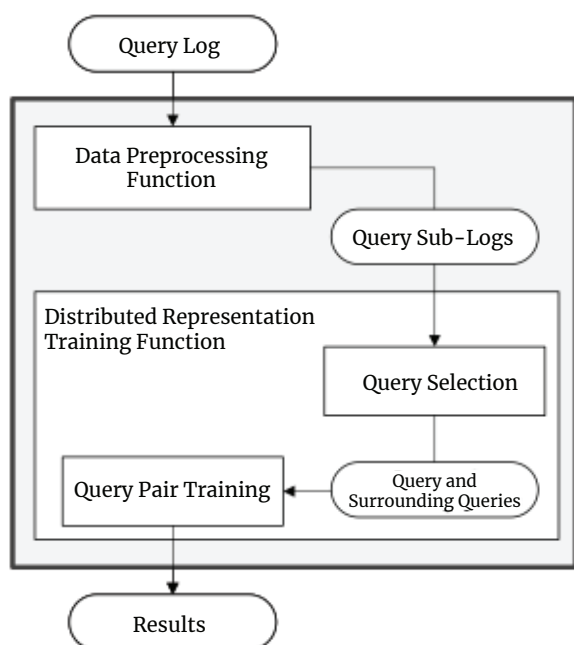
### Distributed representation training function

A distributed representation is initially devised for semantic analysis in natural language processing. Word2Vec, a representative model, trains the relations between words and their surrounding words through a neural network [6]. We modify Word2Vec to shift its focus from words in a sentence to queries in a query sub-log as follows: (1) we replace words with queried domains; and (2) to measure co-occurrences, we adopt the time-interval between queries instead of the distance between words.

First, this function selects the pairs of queries and their surrounding queries, where “surrounding” means occurrences within  $T_b$  seconds before and after a query. A set of surrounding queries,  $x_j \in S_i$ , paired with query  $x_i$  in query sub-log  $X_n$  must satisfy the following conditions:

$$\forall x_j \in S_i : x_i, x_j \in X_n \text{ and } f(x_i, x_j) \leq T_b \text{ and } h(x_i, x_j).$$

Here,  $f(x_i, x_j)$  is the time-interval between  $x_i$  and  $x_j$ ; and  $h(x_i, x_j)$  is true if both queries  $x_i$  and  $x_j$  have A-records. Only such queries are considered because they arise from data exchange.



**Figure 1** Overview of our distributed representation approach for domain names based on DNS queries.

Next, the pairwise relations between query  $x_i$  and surrounding queries  $S_i$  are trained in accordance with the general Word2Vec model. Specifically, the weights in the neural network are optimized to infer the domains for surrounding queries  $S_i$  from the domain for query  $x_i$ . By iteratively training the pairwise relations, our approach finally yields a distributed representation of domains.

## Evaluation

In our approach, we set the parameters to the following values:  $T_\alpha=15.0$  and  $T_\beta=1.0$ . For other parameters in the general Word2Vec model, the number of vector dimensions, number of iterations, initial learning rate, down-sampling rate, batch size, and negative sampling size are set to 200, 500, 0.01,  $1e-5$ , 1024, and 5, respectively. Refer to the literature [6] for details of these parameters.

We collected a dataset from a recursive DNS server in our campus wireless network during a one-month period beginning on 1 March 2020. The dataset comprised a total of 26266740 queries, with a total size of approximately 8.5GB. For the queries with A-records, we aggregated domains with a frequency of less than 10 as “Other”. The resulting number of unique domains was 36261. Note that the number of domains was extensive, and the true distributed representation was unknown. Accordingly, based on the premise that the accuracy of the distributed representation was strongly related to the validity of similarity between domains in the distributed representation, we indirectly evaluated the distributed representation by validating each domain and its similar domains in the dataset.

Figure 2 shows the number of domains similar to each domain in the dataset, where the horizontal and

vertical axes indicate the number of similar domains and the cumulative rate. The similarity criterion between domains involved a cosine similarity value of more than 0.7 for the distributed representation. The results indicate that (A) 90% of domains in the dataset had less than 9 similar domains, (B) 1% of them had more than 109 similar domains, and (C) the maximum number of similar domains reached 255.

To indirectly evaluate our distributed representation, we assessed the relations between domains with more than 0.7 cosine similarity. We found that they could be categorized into the following cases. In the first case, domain  $d_i$  directly co-occurs with domain  $d_j$ , and they are strongly dependent on data exchange. In the second case, co-occurring domains with domain  $d_i$  are similar to those with domain  $d_j$ , and they are potential alternates in data exchange. The third case involves indirect similarities. Specifically, since the relation between domains  $d_i$  and  $d_k$  could fall within either of the above two cases and the same is true for domains  $d_j$  and  $d_k$ , domain  $d_i$  is indirectly similar to domain  $d_j$ . The relations between domains in the first, second, and third cases account for 39%, 29%, and 25% of the total, respectively. Thus, our approach realizes to accurately embed domains into vector space while maintaining their relations. The remaining 7% experience embedding errors caused by the low frequency of co-occurrence between domains. In conclusion, the results suggest the feasibility of the concise and versatile representation for numerous domain names with accurately capturing their interrelations.

## Conclusion

We proposed a distributed representation approach for domain names and indirectly confirmed the accuracy of our distributed representation. The evaluation results indicated the feasibility of the concise and versatile representation for numerous domain names with accurately capturing their interrelations. Our distributed representation is highly suitable as input for machine learning and deep learning models in the field of network security. Consequently, it can be applied to novel security systems based on such models, including the visualization of domain interrelations, detection of malware infections, inference of unknown domains, and enhancement of threat intelligence. Moreover, in security-specialized LLMs (Large Language Models), this representation is expected to facilitate the automation of tasks traditionally performed by

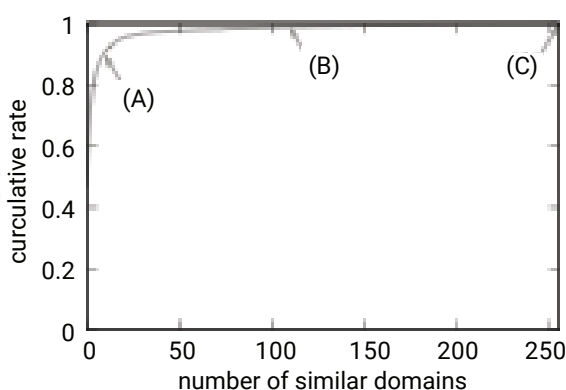


Figure 2 Number of domains similar to each domain in the dataset.



network operators, such as the analysis of security logs and security reports, as it constitutes a fundamental technology for the semantic understanding of domain names. In the future, we plan to deeply analyze the results and the influential parameters.

## Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP24K14932.

## References

1. Lawson C, Watts J. Quick answer: How can organizations use dns to improve their security posture? Gartner Research. 2021.
2. Ma Z, Li Q, Meng X. Discovering suspicious APT families through a large-scale domain graph in information-centric IoT. IEEE. 2019;7:13917-13926.
3. Xu H, Zhang Z, Yan J, Ma X. Evaluating the impact of name resolution dependence on the DNS. Security and Communication Networks. 2019;1-12. doi: 10.1155/2019/8565397.
4. López W, Merlino J, Rodríguez-Bocca P. Learning semantic information from internet domain names using word embeddings. Engineering Applications of Artificial Intelligence. 2020;94:1-13. doi: 10.1016/j.engappai.2020.103823.
5. Pang G, Shen C, Cao L, Hengel AVD. Deep learning for anomaly detection: A review. ACM Computing Surveys. 2022;54(2):1-38. doi: 10.1145/3439950.
6. Di Gennaro G, Buonanno A, Palmieri FAN. Considerations about learning Word2Vec. The Journal of Supercomputing. 2021;77(11):12320-12335. doi: 10.1007/s11227-021-03743-2.