


Which Is More Helpful in Finding Scientific Papers to Be Top-cited in the Future: Content or Citations? Case Analysis in the Field of Solar Cells 2009

Masanao Ochi¹ ^a, Masanori Shiro², Jun'ichiro Mori¹ and Ichiro Sakata¹

¹*Department of Technology Management for Innovation, Graduate School of Engineering, The University of Tokyo, Hongo 7-3-1, Bunkyo, Tokyo, Japan*

²*HIRI, National Institute of Advanced Industrial Science and Technology, Umezono 1-1-1, Tsukuba, Ibaraki, Japan*

Keywords: Citation Analysis, Scientific Impact, Graph Neural Network, BERT.

Abstract: With the increasing digital publication of scientific literature and the fragmentation of research, it is becoming more and more difficult to find promising papers. Of course, we can examine the contents of a large number of papers, but it is easier to look at the references cited. Therefore, we want to know whether a paper is promising or not based only on its content and citation information. This paper proposes a method of extracting and clustering the content and citations of papers as distributed representations and comparing them using the same criteria. This method clarifies whether the future promising papers will be biased toward content or citations. We evaluated the proposed method by comparing the distribution of the papers that would become the top-cited papers three years later among the papers published in 2009. As a result, we found that the citation information is 39.9% easier to identify the papers that will be the top-cited papers in the future than the content information. This analysis will provide a basis for developing more general models for early prediction of the impact of various scientific researches and trends in science and technology.

1 INTRODUCTION

In order to identify research worthy of investment, it is essential to identify promising research at an early stage. In addition, with the increase in the digital publication of scientific literature and the increasing fragmentation of research, there is a need to automatically develop techniques to predict future research trends. Previous research on predicting the impact of scientific research has been conducted using specially designed features for each indicator. On the other hand, recent advances in deep learning technology have facilitated integrating different individual models and constructing more general-purpose models. However, the possibility of using deep learning techniques to predict the impact indicators of scientific research has not been sufficiently explored. In this paper, we extracted the number of citations after publication, one of the typical impact indicators of scientific research, and the corresponding information in the academic literature as a distributed representation. We analyzed the possibility of identifying papers with high impact.

The analysis results show that the linguistic information of academic literature and the distributed representation using network information are different. The results of this paper may provide a fundamental analysis for the development of a more general model for early prediction of the impact of various scientific researches and the prediction of trends in science and technology.

2 RELATED WORKS

Research on the impact of science and technology has focused on developing indicators and their future projections. The development of indices mainly aims at quantifying the influence of an individual subject. For example, the number of citations for papers, the *h*-index (Hirsch, 2005) for authors, the Journal Impact Factor (JIF) (Garfield and Sher, 1963) for journals, and the Nature Index (NI) for research institutions are typical examples. Of course, various other indices have been developed, but most of them focus on papers and authors. On the other hand, some studies


^a  <https://orcid.org/0000-0002-6661-6735>

Table 1: Rank of the number of citations of the papers in the dataset (published in 2009) until 2012.

Ranking @2012	Authors	Title	Journal	Citation @2012
1	Park, S. H., <i>et al.</i>	Bulk heterojunction solar cells with internal quantum efficiency approaching 100%.	Nature Photonics	1,126
2	Chen, H. Y., <i>et al.</i>	Polymer solar cells with enhanced open-circuit voltage and efficiency.	Nature Photonics	930
3	Dennler, G., <i>et al.</i>	Polymer-fullerene bulk-heterojunction solar cells.	Advanced materials	747
4	Krebs, F. C., <i>et al.</i>	Fabrication and processing of polymer solar cells: A review of printing and coating techniques.	Solar energy materials and solar cells	495
5	Grätzel, M., <i>et al.</i>	Recent advances in sensitized mesoscopic solar cells.	Accounts of chemical research	465

have reported predicting these indices. Some studies predict the h -index of future researchers (Ayaz et al., 2018; Miró et al., 2017; Schreiber, 2013; Acuna et al., 2012), studies that predict the number of citations after publication (Bai et al., 2019; Sasaki et al., 2016; Stegehuis et al., 2015; Cao et al., 2016). Among these, the difference is that Stegehuis *et al.* and Cao *et al.* consider the number of citations one to three years after publication and predict the number of citations in the reasonably distant future. In comparison, Sasaki *et al.* predict the number of citations three years later without using citations after publication.

Recently, the application of deep learning techniques to academic literature data has been promoted. The SPECTER model (Cohan et al., 2020), trained on the SciDocs dataset, is a representative example of applying text data in academic literature. However, the SPECTER model uses the citation information of the articles, and it does not simply obtain the distributed representation of each article based on linguistic information alone. In this study, we used the learned Sentence-BERT model (Reimers and Gurevych, 2019) trained by the SNLI corpus (Bowman et al., 2015) as a method to obtain the distributed representation for each article.

On the other hand, there is an attempt to capture the citation information of academic literature data as one huge graph and use it for task evaluation such as link prediction. The SEAL model (Zhang and Chen, 2018) is the top-ranked model on #ogbl-citation2, for the citation prediction task in the academic literature dataset of the Open Graph Benchmark (OGB) (Weihua Hu, 2020), one of the benchmark datasets for graph data, as of February 2021¹. The SEAL model learns by sampling a pair of nodes in a graph and using a subgraph containing the two nodes to predict a link between the sampled nodes. The SEAL model does not use the entire graph as input but rather a large number of small subgraphs,

¹OGB:Leaderboards for Link Property Prediction: <https://ogbstanford.github.io/leaderboards/linkprop/#ogbl-citation2>

which has the advantage of being relatively easy to apply to parallelization and large graphs.

3 METHODOLOGY

The purpose of this paper is to analyze the possibility of identifying papers with high impact by extracting the number of citations after publication, which is one of the representative impact indicators of scientific research, and the corresponding information on academic literature as a distributed representation. In order to analyze the possibility of identifying papers with high impact, we use two methods to obtain the distributed representation for each paper: one is for linguistic information (title and abstract), and the other is for citation information. We compare the distribution of the papers with the highest citations after three years of the publication on the obtained variance representation. The likelihood of identifying such papers is high if the papers with the highest citations are skewed within a particular region and low otherwise. This paper compares the likelihood of identifying the papers with the highest citations by the method using linguistic information and the method using citation information for a relatively small dataset.

The method of comparison is as follows. Obtain the distributed representation of each article by two methods: one is the embedding method for linguistic information, and the other is the embedding method for citation information. After obtaining these two distributed representations, we apply a clustering method under the same number clusters k . Furthermore, we calculate the entropy of the entire dataset with the percentage of papers in the same cluster that will be the most cited papers in n years after publication. The following formula calculates the entropy.

$$H(P_t) = - \sum_{c \in C} P_t(c) \ln P_t(c) \quad (1)$$

However, the symbols in the equation are as follows:

$N(c)$: Number of papers belonging to the cluster c

$N_t(c)$: Number of papers in the cluster c that are among the top cited papers in the cluster

$P_t(c) = \frac{N_t(c)}{N(c)}$: Percentage of papers with the highest citations in the cluster c

The lower value of entropy, the more likely the papers with the highest number of citations concentrate in a particular cluster.

4 EXPERIMENT

In this section, we describe the experiment. First, we describe the scientific and technical literature data used in the experiment. Next, we explain the parameters and conditions we set for the extraction of the variance representation. Here includes how we visualized the data in two dimensions.

4.1 Scientific and Technical Literature Dataset

We received the data from Elsevier, one of the international publishers of many journals. They ran the query “(TITLE-ABS-KEY(nano AND carbon) OR TITLE-ABS-KEY(gan) OR TITLE-ABS-KEY(solar AND cell) OR TITLE-ABS-KEY(complex AND networks)) AND PUBYEAR AFT 2006” on Scopus and obtained the results of the data retrieval.

In this paper, we focus on the 57,935 papers published between 2006 and 2009 that have abstract information, and the top-cited papers are the 66 papers published in 2009 that have been cited more than 100 times by 2012($n = 3$). We show some of the top-cited papers in Table 1.

For the method based on linguistic information, we combine the title and abstract of each paper as input. For the method based on citation information, we create an undirected graph using the citation information of the period, where the nodes are the papers and the edges are the citation relations. This graph has 921,454 nodes and 1,348,424 edges.

4.2 Conditions for Distributed Representation Extraction

For the method using linguistic information, we use the Sentence-BERT(Reimers and Gurevych, 2019) trained model, “nli-bert-large”. We use the SEAL(Zhang and Chen, 2018) for the method using citation information and use the created network as input. We set the parameter $h = 1$ to represent the sam-

pling range of nodes to create the subgraph. However, 10% of the edges are used as test data to evaluate the accuracy of the trained model. The distributed representation acquisition by SEAL learns the presence or absence of an edge between two sampled nodes. For this purpose, we obtain the distributed representation of the target node from the output layer of the MLP layer. We then average with the variance representation of the target node and the neighbouring nodes.

We apply the K-means method for clustering the extracted distributed representations, and we set the number of clusters to $k = 20$. For visualization, we use the UMAP method(McInnes et al., 2018) to reduce the dimensionality to two dimensions.

5 RESULTS

In this section, we explain the results of our experiments. In the experiment, we use a pre-trained model for embedding linguistic information, while we need to train the model for embedding citation information using a dataset. For this reason, we explain the training results of the SEAL model that we selected as the method using citation information. After we confirm that both models have been sufficiently trained, we finally show the comparison results of the distributions of the top-cited papers.

5.1 Training Results of SEAL Model

We show the Precision-Recall curves of the link prediction results for the test data in Figure 1, and we show the Precision and Recall at the threshold where the F-value is the maximum in Table 2. We observe that the Precision-Recall curve has a stable shape and that the model is not sensitive to the output threshold. In addition, the F-value is 0.835 at the threshold $P_{th} = 0.960$ when the F-value is maximum, indicating that the learned model has high accuracy on the test data.

Table 2: Accuracy of Link Prediction.

Precision	Recall	F-value(Max)	P_{th}
0.916	0.768	0.835	0.960

Table 3: Distribution results of the top-cited papers by Entropy.

Model	Entropy
Sentence-Bert	2.900
SEAL	1.742

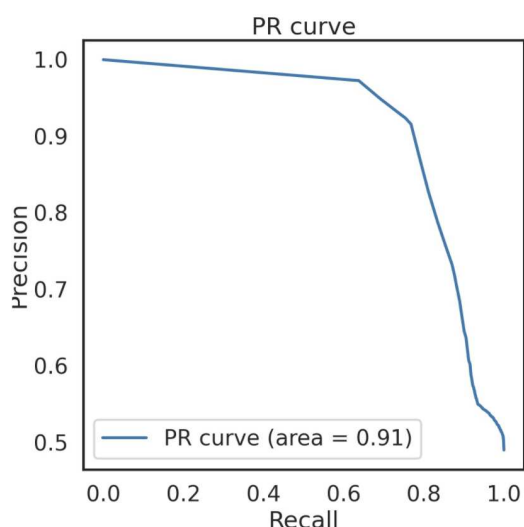


Figure 1: The Result of Precision-Recall curve for link prediction.

5.2 Results of the Distribution of the Top-cited Papers

We show the visualization results of the extracted distributed representations and the distributions of the top-cited papers by the UMAP method in Figure 2. We show the entropies of the distributions of the top-cited papers in the clusters in Table 3. In the visualization result shown in Figure 2, the colour-coding indicates the result of clustering. The red plots sparsely shown with the titles of the papers are the top-cited papers. Comparing the visualization results of Sentence-BERT and SEAL, we can observe that the top-cited papers are more concentrated in SEAL. Table 3 shows that the entropy of the top-cited papers is 2.900 for the Sentence-BERT model, while it is 1.742 for SEAL. In other words, the SEAL model is more biased than the Sentence-BERT model by more than 1.1 points in terms of the number of papers with the highest citations.

6 DISCUSSION

In this section, we discuss the results presented in the 5 section.

First, the SEAL model shows an MRR (Mean Reciprocal Rank) of 0.8767 for the OGB (#ogbl-citation2) leaderboard². The result indicates that the target node is the 1.1th candidate on average. Although the learning results of the link prediction are

²https://ogb.stanford.edu/docs/leader_linkprop/#ogbl-citation2

not as accurate as this, the learning results are comparable to those of the network in this experiment with a smaller size than the #ogbl-citation2 network, which indicates that the learning result is sufficient.

Next, the bias of the papers with the highest citations is more skewed in SEAL than in Sentence-BERT, indicating that the papers are concentrated in specific clusters. This result indicates that the citation relationship is more likely to concentrate the papers whose citations are more likely to increase than the content of the title or abstract. The effect of citations on SEAL learning is limited since the present analysis only covers the papers published in 2009 and marks the top-cited papers after three years.

7 CONCLUSION

In this paper, we conducted an identifiability analysis using distributed representation extracted from academic literature information for predicting the impact of scientific research. Specifically, we used the trained Sentence-BERT model, a method for obtaining distributed representation for linguistic information, and the SEAL model, which is a method for obtaining distributed representation for citation information. We apply these models to identify the top-cited papers three years after publication using only linguistic information and citation information at the time of publication. We evaluate the results by applying the entropy index.

The results show that the SEAL model is more likely than the Sentence-BERT model to bias the top-cited papers to a specific cluster by about 1.1 points. This result indicates that the citation information is more likely to identify the top-cited papers three years after publication than the linguistic information.

On the other hand, there are some limitations to our results. The trained Sentence-BERT model used in this study does not use the academic literature data as training data. It may show different results if the model is trained only on academic literature corpus. In addition, the analysis is a case in a technological field related to solar cells. In addition, we have used the technological field related to solar cells as a case study for 2009. In this study, we analyzed the data as of 2009, using the technology field related to solar cells as a case study. This result may be because solar cells are a highly specialized field or a phenomenon specific to a particular year. In the future, we could obtain different results if the analysis is carried out for different periods in different fields, especially in fields that develop in an interdisciplinary manner.

We will need to discuss further the possibility of

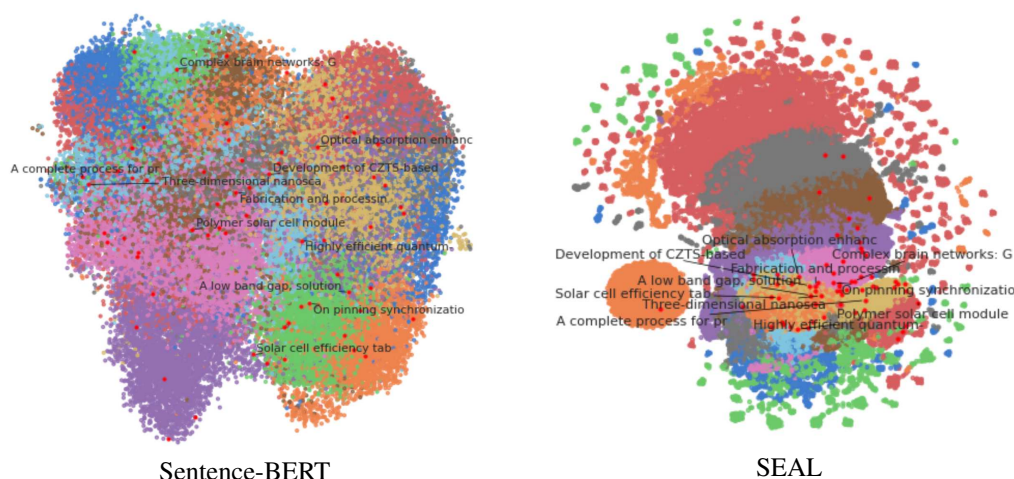


Figure 2: Visualization results of the acquired distributed representation. Color coding is the result of the K-means method.

identifying studies that will be heavily cited in the future by analyzing more models and examples.

ACKNOWLEDGEMENT

This article is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

- Acuna, D. E., Allesina, S., and Kording, K. P. (2012). Predicting scientific success. *Nature*, 489(7415):201–202.
- Ayaz, S., Masood, N., and Islam, M. A. (2018). Predicting scientific impact based on h-index. *Scientometrics*, 114(3):993–1010.
- Bai, X., Zhang, F., and Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics*, 13(1):407 – 418.
- Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Cao, X., Chen, Y., and Liu, K. R. (2016). A data analytic approach to quantifying scientific impact. *Journal of Informetrics*, 10(2):471 – 484.
- Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. In *ACL*.
- Garfield, E. and Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 14(3):195–201.
- Hirsch, J. E. (2005). An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.
- McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.
- Miró, Ò., Burbano, P., Graham, C. A., Cone, D. C., Ducharme, J., Brown, A. F. T., and Martín-Sánchez, F. J. (2017). Analysis of h-index and other bibliometric markers of productivity and repercussion of a selected sample of worldwide emergency medicine researchers. *Emergency Medicine Journal*, 34(3):175–181.
- Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Sasaki, H., Hara, T., and Sakata, I. (2016). Identifying emerging research related to solar cells field using a machine learning approach. *Journal of Sustainable Development of Energy, Water and Environment Systems*, 4:418–429.
- Schreiber, M. (2013). How relevant is the predictive power of the h-index? a case study of the time-dependent hirsch index. *Journal of Informetrics*, 7(2):325 – 329.
- Steghuis, C., Litvak, N., and Waltman, L. (2015). Predicting the long-term citation impact of recent publications. *Journal of Informetrics*, 9.
- Weihua Hu, Matthias Fey, M. Z. Y. D. H. R. B. L. M. C. J. L. (2020). Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*.
- Zhang, M. and Chen, Y. (2018). Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, pages 5165–5175.