Statistical Analysis on Authors and Word Trend of NIPS Papers from 1987 to 2017

DUNDA, Gerry Windiarto Mohamad 20491372

Abstract

This study analyzed NIPS papers from 1987 to 2017 and identified trends in the usage of keywords, research themes, and publication patterns. The research found that the word usage of machine learning has undergone significant changes in focus and emphasis over the past three decades, with some words becoming more popular than others. Additionally, the study found a weak positive correlation between community size and publication count, with Leiden showing stronger correlation coefficients than Louvain. Lastly, clustering NIPS papers using MDS and word mover's distance metric on top 5 keywords revealed non-trivial and fairly spreading clusters. These findings demonstrate the importance of considering community size and using advanced analysis methods when studying publication patterns in scientific communities.

1 Introduction

The field of machine learning has seen significant growth over the last few decades, with the annual Neural Information Processing Systems (NIPS) conference playing a key role in disseminating new research. With the wealth of data generated by NIPS papers, statistical analysis has become an important tool for understanding trends within the field.

In this study, we analyze NIPS papers from 1987 to 2017 using statistical techniques to identify word usage trends and community detection within the research community. By analyzing the frequency of keywords and topic modeling techniques, we aim to uncover the prominent research themes over the years and track how they have evolved.

Moreover, we will use community detection techniques to identify subgroups of researchers who collaborate and publish papers together. By analyzing these communities, we aim to determine whether there is a correlation between research community and number of publications using two different community detection algorithms. Overall, this study will provide valuable insights into the evolution of the machine learning field and highlight the effect of research community on publishing more papers in NIPS. Furthermore, we want to perform unsupervised topic clustering of the papers using multidimensional scaling.

2 Methodology

Dataset. The database contains NIPS papers information such as the titles, abstracts, authors, and contents from the conference proceedings of 1987 to 2017 from the NIPS website. We downloaded the database from *https://www.kaggle.com/benhamner/nips-papers*.

Keyword extraction. Before performing the words trend, we extract some important keywords from each paper. To extract important keywords from each paper, we utilized TopicRank [2], which is an unsupervised keyword extraction algorithm. We preprocessed the text by removing stop words, stemming, and lemmatizing. Then, we applied the TopicRank algorithm to extract the top 10 keywords for each paper, which are ranked based on their importance score. Afterwards, we grouped

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

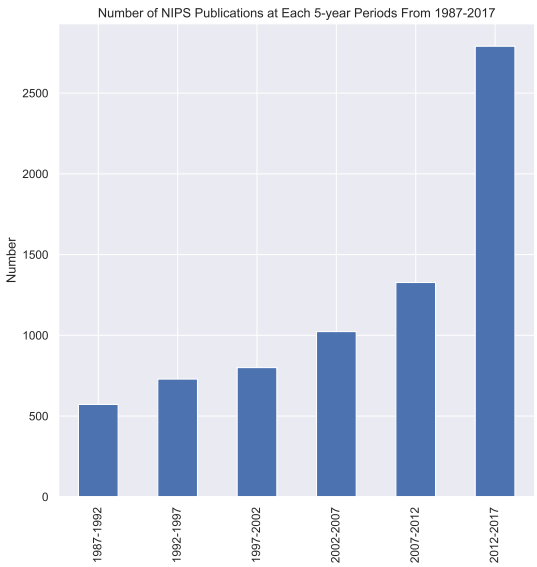


Figure 1: NIPS Publication Trends

the papers into five-year intervals. This allowed us to observe both frequent keywords and some important words in the papers.

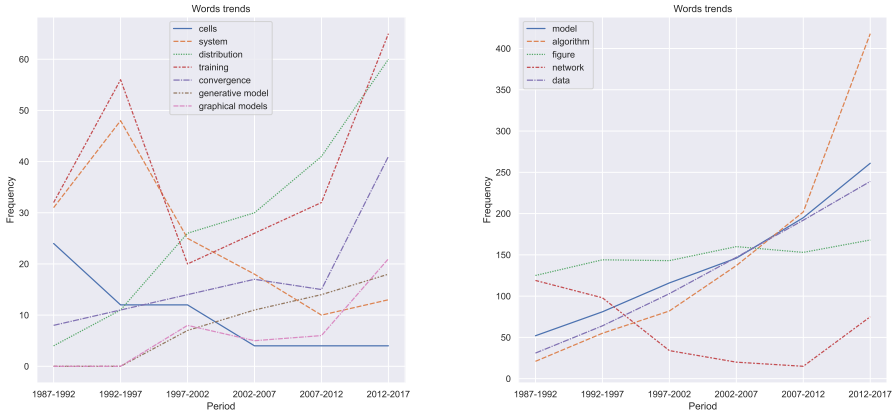
Community detection. To identify the communities of researchers who collaborate and publish papers together, we utilized two different community detection algorithms: Louvain community detector [1] and Leiden community detector [4]. These algorithms are commonly used for community detection in social network analysis and have shown promising results in detecting communities in scientific collaboration networks. For both algorithms, we constructed a weighted undirected network using the co-authorship information in the NIPS papers. We applied each algorithm to detect the communities and obtain a partition of the network into different communities. The Leiden algorithm is an improved version of the Louvain algorithm and is known to produce better results in detecting communities in large-scale networks. After constructing community, we take the average of number of publications on each community and perform linear regression analysis on that number with community size to determine correlation coefficient.

Document clustering. We only consider the first 100 papers on each 5-years period due to computational constraint. We preprocess the paper text by converting all letters to lowercase, removing special characters, stop words, and words that are less than three characters, and finally lemmatizing. Then we use topic rank to extract the top 5 keywords. After that, we calculate the pairwise distance between documents using word mover's distance [3]. Using this distance matrix, we can use multi-dimensional scaling (MDS) algorithm to generate the coordinates of each paper.

3 Results

Publications across different periods. The annual number of publications in NIPS has shown a steady increase from 1987 to 2017 followed by a remarkable increase from 2012 to 2017. The conference started with only 572 papers in 1987-1992 and gradually increased to a thousand papers by the late 1990s. The publication rate continued to grow steadily in the 2000s, and by the end of the decade, the conference was publishing more than 2700 papers. This trend highlights the rapid growth and increasing interest in NIPS fields over the past few decades. This observation is compactly depicted in Figure 1.

Word trends. We select some notable specific keywords from each paper: *cells*, *system*, *distribution*, *training*, *convergence*, *generative model*, and *graphical models*. As shown in Figure 2(a), the word *distribution* shows consistent increase. Intuitively, this word is commonly used as it is a vocabulary



(a) Specific

(b) General

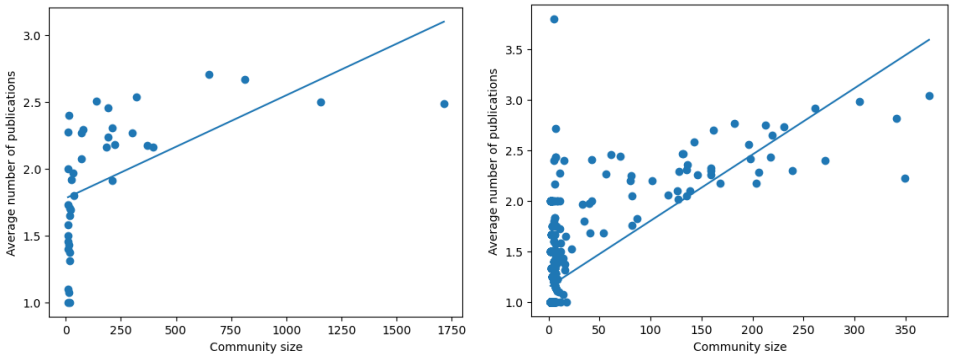
Figure 2: Word trends

used in statistics and probability to describe the machine learning model. Next, the word *training* starts to get popular in 1992-1997. However, it declines during 1997-2002 and rises again afterwards probably due to limited computation at that time. Apart from increasing trend, we can observe the opposite such as *cell* and *system*. It seems that the less frequent usage of *cell* is associated with either biology or recurrent neural network, which is not popular now. Likewise, *system* is commonly used in engineering or science fields rather than machine learning fields. Interestingly, the word *convergence* shows increasing trend, meaning the theoretical works are getting more popular. Additionally, the words *generative model* and *graphical models* are increasing at the same pace.

The general keywords that are discussed in this paper are *model*, *algorithm*, *figure*, *network*, and *data*. As expected, the word *data* is becoming more frequent due to its usage for training or evaluating models. It is also interesting to see that at early times, some researchers have tendency to describe their works as *model* rather than *algorithm* and vice versa. At 2012-2017, the usage gap between those two words is even larger. Another surprising observations is that, although the number of publications increases exponentially, the keyword *figure* is steady, indicating a less frequent usage of such word. This may suggest that latest papers may have less figures compared to the early papers. Lastly, we observe that there is decreasing trend of word *network* despite being the more popular word choice compared to *model* at 1987-1992.

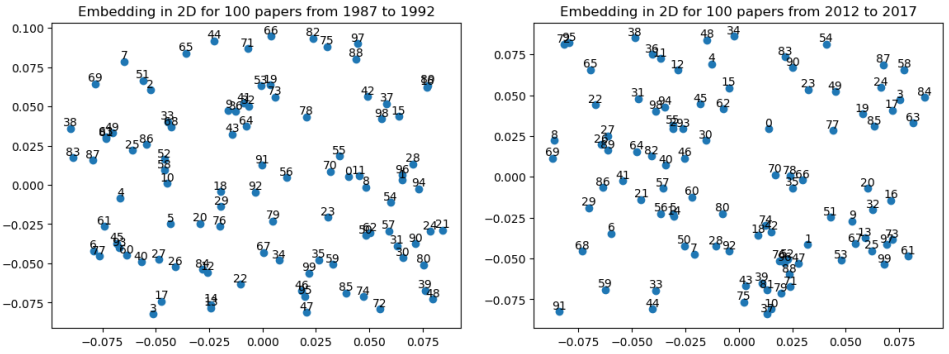
Correlation between community size and number of publications. The present study aimed to investigate the correlation between community size and the number of publications within those communities, using two popular community detection algorithms: Louvain and Leiden. As shown in Figure 3, it shows that community generated by both algorithms produced a weak positive correlation between community size and the number of publications, indicating that larger communities tended to produce more publications on average. Specifically, the correlation coefficient for Louvain was found to be 0.302, while that for Leiden was 0.464, indicating a stronger correlation between community size and publication count for the latter algorithm. These findings suggest that community size could be an important factor to consider when examining publication patterns within scientific communities, and that the choice of community detection algorithm can influence the strength of the observed correlation.

From the graphs, we see that Leiden algorithm yields smaller community compared with Louvain method. Also, Leiden algorithm produces more smaller communities as opposed to Louvain method. The underlying reason behind these is that the Leiden community detector offers enhanced resolution, meaning that it can identify smaller and more densely connected subgroups within larger communities. This allows for a more fine-grained analysis of community structure and can uncover patterns that may not be apparent with coarser resolution methods, which is Louvain algorithm. Furthermore, even though Michael I. Jordan has the most publications in the dataset (101), he does not belong to the largest community generated using Louvain method.



(a) Louvain method (b) Leiden method

Figure 3: The graphs of average number of publication in the community vs community size using



(a) Embeddings for 1987-1992 papers (b) Embeddings for 2012-2017 papers

Figure 4

Documents clustering. Figure shows the result of our clustering method for the first 100 papers in 1987-1992 and 2012-2017. In general, it can be very challenging to interpret the results even the documents that are very close to each other. For example, the neighboring documents with index 45, 93, and 60 with their corresponding titles *Associative Learning via Inhibitory Search*, *Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks*, and *What Size Net Gives Valid Generalization?* have the common keywords describing mathematical vocabularies such as function and error vector are located far away than the documents with index 37, 98, and 15 with their corresponding titles A Back-Propagation Algorithm with Optimal Use of Hidden Units, Optimization with Artificial Neural Network Systems: A Mapping Principle and a Comparison to Gradient Based Methods, and *Cricket Wind Detection*, as perhaps their keywords are related to dynamics and neurons. The important observation is that the algorithm put those three groups far away and almost opposite direction. In case of 2012-2017 papers, we found another non-trivial patterns. Papers titled *On Triangular versus Edge Representations — Towards Scalable Modeling of Networks*, *Sketch-Based Linear Value Function Approximation*, and Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses are closely located and opposite to papers titled Optimal Regularized Dual Averaging Methods for Stochastic Optimization, Learning from the Wisdom of Crowds by Minimax Entropy, and *Nonparametric Max-Margin Matrix Factorization for Collaborative Prediction*.

4 Conclusion

In conclusion, our analysis of NIPS papers from 1987 to 2017 revealed several interesting trends in the use of certain keywords and the evolution of research themes over time. Our findings suggest that the word usage of machine learning has undergone significant changes in focus and emphasis over the past three decades, with some words emerging as particularly prominent and others declining in

popularity. Moreover, our investigation into the relationship between community size and publication count using the Louvain and Leiden algorithms demonstrated a weak positive correlation between these two variables, with Leiden showing stronger correlation coefficients than Louvain. These results indicate the importance of considering community size when studying publication patterns in scientific communities and highlight the benefits of using advanced community detection methods like Leiden for more precise and efficient analysis of complex networks. Lastly, we try to cluster NIPS papers based on MDS technique with word mover's distance metric on top 5 keywords and find that the clustering are non-trivial and fairly spreading.

References

[1] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. DOI: 10.1088/1742-5468/2008/10/p10008. URL: https://doi.org/10.1088/1742-5468/2008/10/p10008.

[2] Adrien Bougouin, Florian Boudin, and Béatrice Daille. "TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction". In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan: Asian Federation of Natural Language Processing, Oct. 2013, pp. 543–551. URL: https://aclanthology.org/I13-1062.

[3] Matt Kusner et al. "From Word Embeddings To Document Distances". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 957–966. URL: https://proceedings.mlr.press/v37/kusnerb15.html.

[4] V. A. Traag, L. Waltman, and N. J. van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific Reports* 9.1 (Mar. 2019). DOI: 10.1038/s41598-019-41695-z. URL: https://doi.org/10.1038/s41598-019-41695-z.