POSTER NUMBER: 383

Algorithm Entities Usage in Chinese Academic Articles from The Domain of Information Science

Yuzhuo Wang, Heng Zhang, Chengzhi Zhang *
Department of Information Management, Nanjing University of Science and Technology, Nanjing, China, 210094

Introduction

With the emergence of the fourth paradigm for science, the demand for big data-related algorithms is increasing. As one of the three key elements of artificial intelligence, algorithms have made great contributions to both social science and natural science fields.

Academic papers are excellent resources for scholars to learn algorithms. However, manually finding algorithms from massive articles is time-consuming and high-cost. Scholars need automatic methods to extract algorithms.

In this article, a deep learning model with a manual filtering method is proposed to find algorithm entities and carry out further exploration. Taking Chinese academic papers in the information science (IS) domain as an example, we plan to explore:

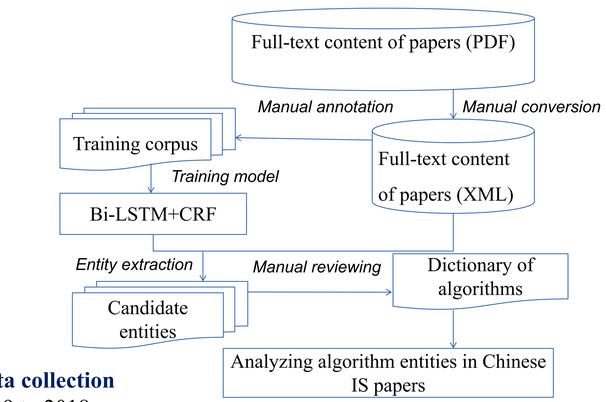
1.How to extract algorithm entities with an automatic method?

2.What is the distribution of algorithms in Chinese IS academic papers?

We define algorithm entities as nouns or noun phrases representing the name of algorithms or models that are algorithms in nature.

Method

We collected the full-text content of papers and then built corpus to train extraction models. Algorithm entities were extracted and reviewed to explore the distribution of algorithm entities in IS papers.



Data collection

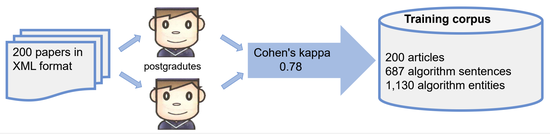
2009 to 2018

China Society for Scientific and Technical Information (JCSSTI)

The full-text content of 1,367 articles published in the JCSSTI, which is the top journal in the field of IS in China.

Corpus construction

The papers in PDF format were converted into XML format manually. Two postgraduates annotated algorithm entities in 200 random selected papers independently and corrected the inconsistent results.



Candidate entities extraction

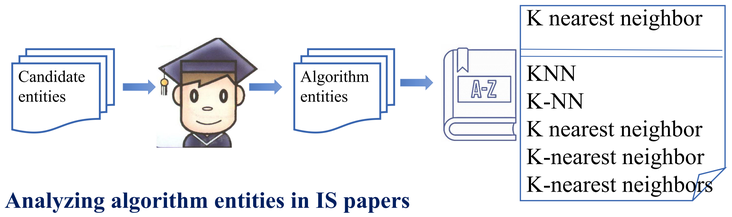
Model: Bi-LSTM+CRF

training set : testing set = 8:2

After using **FastText** (https://fasttext.cc/) to train word-vector, training data was inputted to the Bi-LSTM+CRF model. The trained model was then utilized to extract candidate entities.

Reviewing algorithm entities

A Ph.D. candidate reviewed all the candidate entities and picked out algorithm entities. A dictionary was compiled where full name, abbreviation, and aliases representing the same algorithm were organized into a set.



Analyzing algorithm entities in IS papers

(1) Classifying algorithms

Algorithms entities were classified into three types:

A *single algorithm* is an explicit algorithm not showing the category and function in the name (Support vector machine).

A *functional algorithm* indicates a specific function or type in the name (Classification algorithm).

A *composite algorithm* contains the actual task solved by the algorithm in the name (Literature recommendation algorithm based on user interest topics).

(2) Influence of algorithms

We use the same method in Wang's work (Wang & Zhang, 2020) to calculate the influence of the algorithm entity in IS field. For an algorithm j,

$$Influence(j) = (\sum_{i=2009}^{2018} \left(\frac{N_{ij}}{N_j}\right))/T_j$$

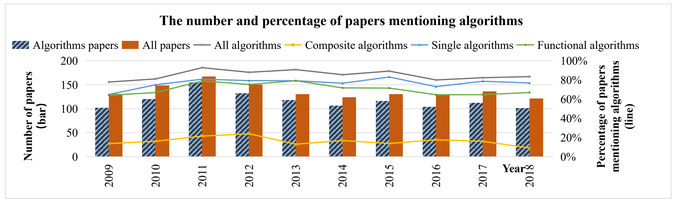
Where i represents the year, ranging between 2009 and 2018. Ni is the number of articles published in year i, Nij is the number of articles that mentioned algorithm j in year i. Tj is the duration from the year when algorithm j first appeared in papers to 2018.

Result

Recognition result of algorithm entities

Model preference	Precision	Recall	F1-value
	67.26%	72.82%	69.93%
Results	23,352 candidate entities	2,122 algorithm entities	962 Single algorithms
			775 Functional algorithms
			385 Composite algorithms

Algorithm entities distributed in different year



Single algorithms with high influence in IS papers

Rank	Algorithm	Rank	Algorithm
1	TF*IDF	6	Neural network
2	Vector space model	7	LSTM
3	SVM	8	K-means
4	Cosine similarity	9	Latent dirichlet allocation
5	Mutual information	10	Decision trees

Conclusion

For the algorithms mentioned in Chinese academic articles in IS field:

- Algorithms play an essential role in IS papers in China.
- The influence of the single algorithm is the highest and the most stable.
- Classical machine learning algorithms and emerging deep learning algorithms are more influential.

Reference

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). Introduction to Algorithms, Third Edition. The MIT Press.

Hou, L., Zhang, J., Wu, O., Yu, T., Wang, Z., Li, Z., Gao, J., Ye, Y., & Yao, R. (2020). Method and Dataset Entity Mining in Scientific Literature: A CNN + Bi-LSTM Model with Self-attention. arXiv:2010.13583. http://arxiv.org/abs/2010.13583

Wang, Y., & Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. Journal of Informetrics, 14(4), 101091.

Contact

Yuzhuo Wang : wangyz@njust.edu.cn
Heng Zhang : zh_heng@njust.edu.cn
Chengzhi Zhang : zhangcz@njust.edu.cn
* Corresponding author