Predicting the future influential researchers in big scholarly network

GU QINGJIAN

CONTENTS

ABSTRACT

The purpose of this paper is to predict academic influential researchers in the future by using the weighted model of factors such as author, publication venue, title, abstract, and publication year, etc. There are two main challenging: 1) some factors do not have exact numeric values to be used in regression models. 2) there is not direct relationship between citation count and these factors. Word2vec is chosen among several methods to process words. The main research method is the combination of theoretical model and code simulation.

1 INTRODUCTION

In many areas, major achievements often require steep learning curves and long training. In today ś society where information is transmitted on a large scale, people are eager to identify potential scientists as early as possible.
The most representative indicator system used to evaluate the impact of scientific research is the "H index" proposed by Jorge Hirsch, a statistical physicist at the University of California, San Diego (UCSD) in 2005. The "H-index" is defined as "there are H articles that have been cited not less than H times", that is, when a scientific researcher publishes N papers, each of the H papers is quoted at least H times. A study on Nature is based on the "H index", using linear regression and elastic net regularization in machine learning to fit a predictive model. Although the concept of "H index" is simple and easy to calculate, it also has certain limitations, so it is necessary to design a more comprehensive evaluation index.

There are two ways to improve: 1) In order to balance the contribution of the essay co-authors to the article, the contribution rate of the author is introduced to solve the problem that different authors contribute differently to the paper due to co-authorship, so as to more accurately evaluate the academic influence of the researcher. 2) We do not consider the citation from the perspective of the author but from the perspective of the article, to get a relationship between high-cited papers and factors such as author, publication venue, title, abstract, and publication year, etc. The citation of an author can be represented by the total number of his articles ' citations.

2 METHODS

2.1 theoretical analysis

CO-AUTHORSHIP Contribution weight
In order to more reasonably allocate the contribution of authors of different signature ranks, the academic community has proposed a number of allocation methods based on the author's signature rank. Research finds that Arithmetic algorithm is better than other methods to some extent.
For papers with multiple authors, the position weight design should satisfy that the first author gives greater weight than the second author, and the second author is given more weight than the third author, namely $w_i > w_j, \forall i < j$. The sum of the weights of the authors of all papers is equal to 1. Assuming that there are k co-authors in a paper, the contribution weight w_j of the j-th author should be satisfied

$$w_j = \frac{2(k-j+1)}{k(k+1)} \tag{1}$$

Therefore, the improved "H index" is the value of H multiplied by the authorś contribution weight, which is named "H_w index".

NETWORK ANALYSIS:

DIFFUSION MODEL Threshold-based diffusion model
The structure of the network plays a decisive role in the dissemination of information on the network.
To predict the influence of a researcher in the future, we should study how his paper is spread in academic networks. An optimal network diffusion model should be an improved threshold model. Each node v has an information acceptance threshold θ_v. Each node v is affected by all its active neighbor nodes. The impact value is represented by weight w. Assume node u diffuses node v with influence weight $w_{v,u}$,and $w_{v,u}$ can also be allowed to take negative values. If successful, node v is converted to active node; If failed, consider attenuation coefficients γ. After several steps, the effect of node u on node v is not then $w_{v,u}$, but $\gamma w_{v,u}$. Once a node becomes infected, it remains infected for ever.

PAGE RANK ALGORITHM Two main Academic Impact Assessment and Prediction methods are based on Page Rank Algorithm, which is used to evaluate web pages. Each page will have a Page Rank (PR) value, to indicate importance of the page. Assume that the PR value of a web page P1 is PR(P1). The value of weight pointed to P2 is PR(P1)/L(P1), where L(P1)

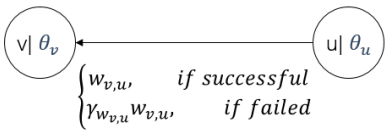


Figure 1: diffusion from node u to node v

represents the total number of links of P1.The final PR value of web page P2 is the sum of link weight. Improved Page Rank algorithm adds the concept of damping coefficient ρ to measure the probability of stopping at current node. For a N-point network, a complete equation of Page Rank Algorithm is shown as follows.

$$PR(P_i) = \frac{1-\rho}{N} + \rho \sum \frac{PR(P_i)}{L(P_i)} \tag{2}$$

2.2 machine learning

1. preprocessing
2. Use data set for fitting
3. Use the fitted formula to predict future influence

The main challenging is to extract features from title, authors, year, venue, citation count, and abstract and turn into numeric values.

PREPROCESSING:

1. Non-numeric value One method to handle text documents is word2vec (word embeddings), which is to turn the words in natural language into Dense Vectors that the computer can understand, and words with similar meanings will be mapped to similar positions in the vector space.
The main steps of word embeddings:
1).Word Segmentation / Stemming and Morpheme Reduction
2).Construct a dictionary, count word frequency

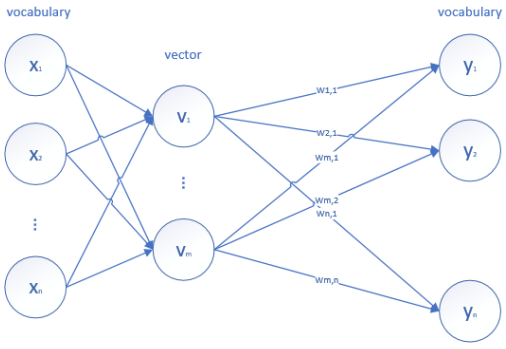


Figure 2: Word2Vec training model

3).Construct a tree structure

2. numeric value

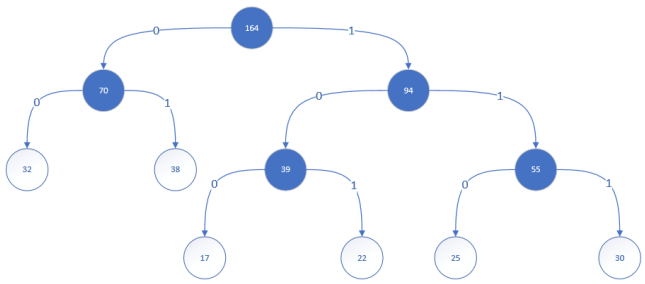


Figure 3: huffman tree based on word frequency

Algorithm 1 author citations

Input: A dictionary of author names , {}; A dictionary of records, {"author name": name, "citation": citation}
Output: A dictionary of author citations , {"author citation": int(citation) };

```
for each name in author names do
    if name in author citation then
        author citation[name] += int(citations)
    else
        author citation[name] = int(citations)
    end if
end for
```

author citations: calculated the total number of citations an author has in whole dataset.
publication venue: (similar to author citation) calculate the total number of citations for the papers in a specific publication venue in the whole dataset.
publication year: directly extracted from the dataset.

3. plots
The data set here contains only about 500,000 articles. We can intuitively see the relationship between citation and its numeric features (Figure 4).

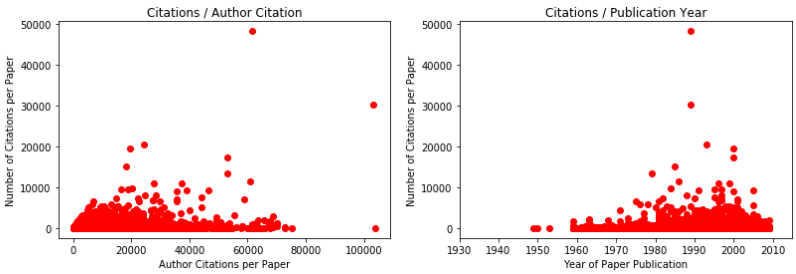
REGRESSION: we aim to predict (article) citation, and count the total citation of a certain author (same as algorithm 1).

3 RESULTS AND DISCUSSION

Table 1 shows the comparison of different algorithm.

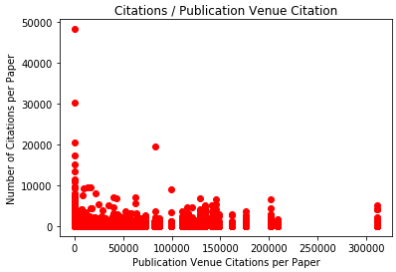
Table 1: Table of MMAE & MMSE

algorithm	minimum mean absolute error	minimum mean square error
linear regression	18.96	136.80
non-linear regression	18.48	136.32
SVM	17.74	138.42
regression tree	20.27	147.54



(a) citations-author citations

(b) citation-publication years



(c) citations-publication venues citation

Figure 4: citation and its influencing factors

3.1 conclusion

Linear methods may not lead to a best solution in this problem. Meanwhile, in the case of uniform dataset, KNN may have a good performance. However, KNN has its drawbacks of algorithm complexity.

3.2 Future Work

Current method of extract numeric features from an factor, such as "author", is to count the total number of his citation. However, we consider it necessary to add an attenuation factor to express the publication year of the paper.