Learning Approaches to Predicting Future Citation Counts

Sree Ram Sreekar
B.Tech. 4th year
CSE, IIT Hyderabad
cs13b1008@iith.ac.in

Akshita Mittel
B.Tech. 4th year
CSE, IIT Hyderabad
cs13b1040@iith.ac.in

Surya Teja Chavali
B.Tech. 4th year
CSE, IIT Hyderabad
cs13b1028@iith.ac.in

ABSTRACT
In this paper, we survey different approaches that have been suggested in the past to solve the problem of predicting the *future citation* count of a scientific article after a given time interval of its publication. To this end, we adopted the Aminer dataset consisting of citation data extracted from DBLP, ACM, and other sources. We compare and evaluate the different solutions proposed. Further, we intend to present a comparative evaluation of the popular h-index and p-rank(a PageRank inspired model for assessing author impact) metrics.

CCS Concepts
H.3[**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2[**Artificial Intelligence**]: Natural Language Processing - *Text Analysis*

Keywords
Citation count prediction; regression models; data engineering

1. INTRODUCTION
The explosive increase in the volume of scientific publications in recent years is expected to continue for the foreseeable future. Citation count of a paper is widely used to indicate the popularity/impact of a scientific papers.

Accurate and automatic prediction of citation counts would provide a powerful method for evaluating articles and faster identification of promising articles could accelerate research and dissemination of new knowledge. Accurate models for citation count prediction would also improve our understanding of the factors that influence citation.

Citations are a noisy, indirect quality measure, and accumulation rates vary unpredictably between articles. Breakthrough papers can stop receiving citations after review articles replace them or the subject matter becomes common knowledge. Predictions based on current data assume that citation behavior will not change in the future, and this assumption may be violated in fast-paced research fields such as Computer Science.

The very limited number of solutions suggested have mostly modeled the problem as a learning task – given a set of features and a particular time interval, a regression model is trained on the entire set of the training population, and accordingly, the future citation count of a query paper is estimated.

2. WORK DONE
Surveyed four[1, 2, 3, 4] related papers. The broad approach in all papers is the same. A set of features is selected. These features are extracted from each paper. A combination of regression models are used to get CCP(paper, Δt).

[1] is by far the most mature and justified approach of the four. 6 author-centric, 3 venue-centric and 4 paper-centric features comprised the feature set for each paper. A two-stage prediction model was used. Using a rule-based classifier,, each paper was classified into six categories. An SVR model was learned on each of the six categories.

We have implemented [1] completely. The training of SVR model on each category could not be accomplished due to computational limitations.

[2] has compared different regression models and concluded that SVR performs significantly better than Logistic Regression, CART.

[3] adopts a different learning approach. Threshold values of 20, 50, 100, 500 citations were used as response variables. The response variables were binary and indicated if the predicted number of citations exceed their corresponding thresholds. The paper proposed that error metrics such as this were more intuitive to the CCP problem than traditional metrics such as mean-square error or percent variation.

[4] has essentially crafted fairly non-trivial features and used an SVR to train the model.

The features for each of 2, 3, 4 have been extracted as well.

3. METHODOLOGY
We include the features and learning models for each of the four papers studied.

3. 1 FEATURES
[1] Towards a Stratified Learning Approach to Predict Future Citation Counts:

1. Author centric features
    a. Author productivity
    b. Author H-index
    c. Author diversity
    d. Sociality of author
2. Venue centric features
    a. Long term venue prestige
    b. Short term venue prestige
    c. Venue diversity

3. Paper centric features
    a. Team size
    b. Reference count
    c. Reference diversity
    d. Keyword diversity
    e. Topic diversity

[2] Citation Count Prediction: Learning to Estimate Future Citations for Literature

1. Topic Rank
2. Diversity
3. Recency
4. H-Index
5. Author Rank
6. Productivity
7. Sociality
8. Authority
9. Venue Rank
10. Venue Centrality

[3] Models for Predicting and Explaining Citation Count of Biomedical Articles

1. Article title
2. Article abstract
3. MeSH terms
4. Number of articles for first author
5. Number of citations for first author
6. Number of articles for last author
7. Number of citations for last author
8. Publication type
9. Number of authors
10. Number of institutions
11. Journal impact factor
12. Quality of first author's institution

[4] Predicting Citation Counts Using Text and Graph Mining

1. Author features:
    a. log of sum of citations over papers
    b. log mean citations over papers
    c. log of max citations over papers
    d. H-index
    e. G-index
2. Venue features:
    a. log of sum of citations over papers
    b. log mean citations over papers
    c. log of max citations over papers
    d. H-index
    e. G-index
3. References network:
    a. H-index
    b. G-index
    c. log mean citations over papers
    d. log median citations over papers
    e. Graphical features such as:
        i. Graph density
        ii. Clustering coefficient
        iii. Connectivity
4. Content Similarity:
    a. language model published
    b. language model referenced

3.2 LEARNING MODELS

[1] Towards a Stratified Learning Approach to predict future Citation Counts

Two-stage prediction model

In the first stage, the model maps a query paper into one of the six categories, and then in the second stage a regression module is run only on the subpopulation corresponding to that category to predict the future citation count of the query paper.

[2] Citation Count Prediction: Learning to Estimate Future Citations for Literature:

Linear regression(LR), k-NN, SVR, CART models were all used and compared against each other. The SVR model gave the best performance.

[3] Models for Predicting and Explaining Citation Count of Biomedical Articles

The response variable is defined by a set of citation thresholds to determine if an article is labeled positive or negative. For a given threshold, a positive label means that an article received at least that number of citations within 10 years of publication. Thresholds of 20, 50, 100, 500 were chosen. Predictions were made for a binary repose variable. SVM models were used as the learning algorithm.

[4] Predicting Citation Counts Using Text and Graph Mining

SVR model was used.

4. WORK PLAN FOR FINAL PRESENTATION

4.1 Comparative Study of models

A comparative study of the performance of the four models on shall be done.

4.2 Comparative Study of features

A comparative study of the feature set comprising all the feature sets of the four papers shall be done

4.3 Discussion of h-index and p-rank

An analysis discussing the comparative advantages and disadvantages of h-index and p-ranks shall be presented.

4.4 Cross Validation

A 10-fold cross validation analysis of the approaches will be presented.

5. EVALUATION

5.1 The Dataset

We use the well-studied Arnet Miner Dataset[5], found here. The citation data is extracted from DBLP, ACM, and other sources. The first version contains 629,814 papers and 632,752 citations. Each paper is associated with abstract, authors, year, venue, and title. It has been extensively used for the purposes of clustering with network and side information, studying influence in the citation network, finding the most influential papers, topic modeling analysis, etc.

There are eight versions of this dataset, as shown below.

Data set	#Papers	#Citations
Citation-network V1	629,814	>632,752
Citation-network V2	1,397,240	>3,021,489
DBLP-Citation-network V3	1,632,442	>2,327,450
DBLP-Citation-network V4	1,511,035	2,084,019
DBLP-Citation-network V5	1,572,277	2,084,019
DBLP-Citation-network V6	2,084,055	2,244,018
DBLP-Citation-network V7	2,244,021	4,354,534
DBLP-Citation-network V8	3,272,991	8,466,859
ACM-Citation-network V8	2,381,688	10,476,564

We shall use V1 and V2 as training sets, V3 for validation, and V4 for testing.

5.2 Evaluation Metric

We seek to compare the approaches to this problem by way of using the Coefficient of Determination(R^2) statistic. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. This is defined by the equation:

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

The reason for this choice is that all papers we've used have chosen this metric for evaluation.

5.3 Results and analysis

We're yet to run our models and cross-validate them - we've only completed data cleaning, feature extraction, and serialized the dataset into an assimilable form.

6. REFERENCES

[1] Towards a Stratified Learning Approach to Predict Future Citation Counts, Tanmoy Chakroboty et al, JCDL '14, Pages 351-360

[2] Citation Count Prediction: Learning to Estimate Future Citations for Literature, Yan et al, CIKM '11, Pages 1247-1252

[3] Models for Predicting and Explaining Citation Count of Biomedical Articles,Lawrence D. Fu, Aliferis, AMIA Annual Symposium Proceedings. 2008; 2008: 222–226.

[4] Predicting Citation Counts Using Text and Graph Mining, Avishay Livne et al, Modern Advances in Applied Intelligence, Volume 8482, Pages 109-119

[5] ArnetMiner: Extraction and Mining of Academic Social Networks, Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998