Citation Count Prediction: Learning to Estimate Future Citations for Literature

Rui Yan
Dept. of Computer Science and Technology
Peking University
Beijing 100871, P. R. China
r.yan@pku.edu.cn

Jie Tang
Dept. of Computer Science and Technology
Tsinghua University
Beijing 100871, P. R. China
jietang@tsinghua.edu.cn

Xiaobing Liu
Dept. of Computer Science and Technology
Peking University
Beijing 100871, P. R. China
lxb@net.pku.edu.cn

Dongdong Shan
Dept. of Computer Science and Technology
Peking University
Beijing 100871, P. R. China
sdd@net.pku.edu.cn

Xiaoming Li
Dept. of Computer Science and Technology
Peking University
Beijing 100871, P. R. China
lxm@pku.edu.cn

ABSTRACT

In most of the cases, scientists depend on previous literature which is relevant to their research fields for developing new ideas. However, it is not wise, nor possible, to track all existed publications because the volume of literature collection grows extremely fast. Therefore, researchers generally follow, or cite merely a small proportion of publications which they are interested in. For such a large collection, it is rather interesting to forecast which kind of literature is more likely to attract scientists' response. In this paper, we use the citations as a measurement for the popularity among researchers and study the interesting problem of Citation Count Prediction (CCP) to examine the characteristics for popularity. Estimation of possible popularity is of great significance and is quite challenging. We have utilized several features of fundamental characteristics for those papers that are highly cited and have predicted the popularity degree of each literature in the future. We have implemented a system which takes a series of features of a particular publication as input and produces as output the estimated citation counts of that article after a given time period. We consider several regression models to formulate the learning process and evaluate their performance based on the coefficient of determination (R^2). Experimental results on a real-large data set show that the best predictive model achieves a mean average predictive performance of 0.740 measured in R^2, which significantly outperforms several alternative algorithms.

Categories and Subject Descriptors

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

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

General Terms

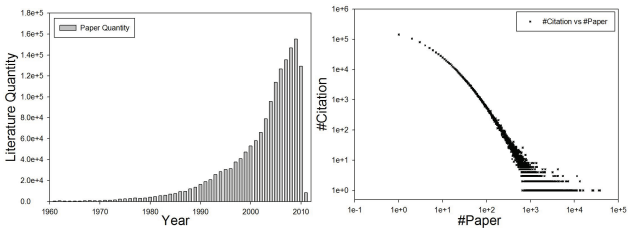
Algorithms, Experimentation, Performance

Keywords

Citation count prediction, regression models, data engineering

1. INTRODUCTION

The rapid evolution of scientific research has been creating a huge volume of publications every year, and is expected to remain in this situation within the foreseeable future. Figure 1 shows statistics on a large literature database in Computer Science.[1] Figure 1.(a) visualizes the explosive increase on the volume of publications in the past years, in particular recent years. For example, the number of publications in 2009 almost triples than that of 10 year before. Effective scientific research requires keeping up with previous literature, but it is not wise, nor possible, for researchers to track all existed publications because the volume of literature collection grows extremely fast as mentioned. Therefore, researchers generally follow, or cite merely a small proportion of publications which they are interested in. An interesting phenomenon is that some of research papers are more likely to attract scientists' response than the others. If we use citation count as the popularity of papers among academia, we have the following observation in Figure 1.(b).



(a). The growing volume of literatures. (b). Distribution of literature citation.

Figure 1: Statistics of literature data from ArnetMiner.

[1] http://arnetminer.org.

It is natural to find that not all publications attract equal attention to academia. We show the citation distribution (the number of papers vs. citation counts) in the log-log plot of Figure 1.(b): the interests toward literature measured by citation counts is highly skewed. Not surprisingly, the plot follows a power law distribution. A power law relationship between two quantities x and y can be written as y=ax^b where a and b are constants. We see that a huge number of research papers attract only a few citations, and a few research papers accumulate a large number of citations.

For the ever-growing literature collection, it is rather interesting to forecast which kind of literature is more likely to attract scientists' response. In this paper, we use the citation counts as a simple measurement for the popularity among researchers and the citation count is calculated by how many times a particular publication is cited by other articles. We study the interesting problem of Citation Count Prediction (CCP) to examine the correlative characteristics for popularity.

As a pilot study on learning to forecast future citations for literature, Citation Count Prediction faces with several challenges:

• The first challenge for CCP is to explore the truly effective features important to future citation counts from several aspects such as paper content, author expertise and venue impact. We introduce a series of features which are correlative with the number of future citations of literature;

• The second challenge for CCP is to combine all relevant features to identify the potentially interesting papers in a unified predictive model, linearly or non-linearly. Given multiple features relevant to popularity, i.e., citation counts in this study, we utilize several regression models to estimate future citations.

Our contributions are manifold by solving these challenges. In Section 2 we first define a series of features which correlate with citation counts. We then formulate citation count prediction as a learning problem and introduce several regression models to unify all possible features for prediction. We describe experiments and evaluations in Section 3, including performance comparisons and feature analysis. We briefly review previous works in Section 4 and draw conclusions in Section 5.

2. CITATION COUNT PREDICTION

2.1 Problem Definition

In this section, we first present several necessary definitions and a formal representation of the citation count prediction problem.

Citations. *Given the literature corpus* ***D****, the citation counts* ($C_T(.)$) *of a literature article* $\mathbf{d} \in D$ *is defined as:*

$$\begin{aligned} citing(d) &= \{d' \in D : d' cites\ d\} \\ C_T(d) &= |citing(d)| \end{aligned} \tag{1}$$

Learning task: Given a set of article features, $\vec{X} = x_1, x_2, \ldots, x_n$, our goal is to learn a predictive function f to predict the citation counts of an article d after a give time period Δt. Formally, we have

$$f(d|\vec{X}, \Delta t) \rightarrow C_T(d|\Delta t) \tag{2}$$

To learn the predictive model, we have investigated multiple relevant factors such as paper content, author expertise and venue impact. It is also important to find unified models which are able to consider all the features simultaneously. We introduce both aspects in the following subsections.

2.2 Feature Definition

2.2.1 Topic Rank

Topics have long been investigated as a significant feature for literature contents [12]. We utilize the unsupervised Latent Dirichlet Allocation [2] to discover topics for our corpus as it has been applied successfully to many content analysis tasks, and implementations are freely available[2]. We empirically train a 100-topic models on our corpus - the top words for a few of the sample topics are shown in Table 1.

Table 1: Top words from selected LDA sample topics after stop-word removal.

Topic	Representative words
00	distribution probability value random probabilistic expected
09	query information search semantic retrieval document
12	mobile network wireless nodes communication device
84	data mining patterns analysis association set
97	programming formal language specification verification logic

Our topic feature works by inspecting the probability distribution over topics assigned to a literature article d. That is, for each of our 100 topics, our topic model calculates $p(topic_i|d)$, the inferred probability of topic i in document d. The topic distribution $\mathcal{T}(d)$ over all topics in document d is then:

$$\mathcal{T}(d) = \{p(topic_1|d), p(topic_2|d), \ldots, p(topic_{100}|d)\}$$

To calculate the total citation counts of a particular topic from article d, denoted by $C_T(topic_i|d)$, we distribute the citations of the article $C_T(d)$ according to the topic distribution $\mathcal{T}(d)$, i.e., $C_T(topic_i|d) = C_T(d) \times p(topic_i|d)$ and hence we obtain the citations of all 100 topics by using:

$$C_T(topic_i) = \sum_{d \in D} C_T(topic_i|d) \tag{3}$$

where D is the whole literature collection. We rank topics by average citation counts, namely topic "popularity".

2.2.2 Diversity

We obtain a notion of the breadth of an article from its topic distributions. This is important for identifying methodology papers, which are often cited by a wider topical range of articles. When an article has a vast range of audience, it is likely to be cited by authors from various research fields, and hence attract high citation counts. To measure the topical breadth of an article, we calculate the entropy of the document's topic distribution:

$$Diversity(d) = \sum_{i=1}^{|\mathcal{T}|=100} -p(topic_i|d) \cdot \log p(topic_i|d) \tag{4}$$

2.2.3 Recency

Temporal dimension has long been proved to be significant in literature studies [1, 21]. Intuitively, the citation counts accumulate as time passes by, thus a measure of the age of an article is assumed to be important. We include as a feature the number of years since the article was published. We expect a positive correlation on temporal recency - the longer an article is published, the more citations it may receive.

2.2.4 H-Index

[2]We use Stanford TMT (http://nlp.stanford.edu/software/tmt/), with default settings for all model parameters.

The h-index is useful which attempts to measure both the productivity and impact of the published work of a scientist [8]. The index is based on the set of the scientist's most cited papers and the number of citations received in others' publications. Besides, h-index has been proved to have predictive power of scientific output and impact of a researcher [9]. Therefore, we consider h-index as a candidate feature to predict citation counts.

2.2.5 Author Rank

We try to identify the correlation between author rank and average citation count. Sometimes, the "fame" of an author's name ensures the amount of citations. Each author has his/her own expectation of citation counts. We calculate all authors according to their average citation counts and assign each of them a unique rank position number.

2.2.6 Productivity

According to [1], authors have inclination to cite papers they have written themselves. Intuitively, the more productive an author is, the larger chances for his/her papers to be cited. We hence assume the productivity of an author is relevant to the citation counts, due to the self-citation behavior analysis from previous studies.

2.2.7 Sociality

From the author social factor studies in [1], researchers tend to cite papers from whom the author(s) have co-authored. Thus, it is natural to assume that the paper from a widely connected author has a larger probability to be cited by his/her wide variety of co-authors. A straightforward and simple measurement is to count the Number of Co-Authors (NOCA) and we assume the correlation between the number of co-authors and average citation counts.

2.2.8 Authority

A unique social network for academia is established from the "citing - cited" relationships among literature articles. Publications carry with author authorities: a widely cited paper indicates peer acknowledgements, and hence indicates authority. We transmit paper authority to all its authors. We first build a graph of $G_a(V, E)$, where V is the set of vertices and each vertex v_i in V represents a literature paper and E denotes the *citing-cited* linkage. The citing-cited graph has directions. The out-degrees measure how many times a paper is cited while in-degrees indicate the references of a particular paper. When there is a citing-cited relationship between two papers, we add a link into the graph. We use standard cosine similarity between two papers to weigh the linkage in the graph, i.e., $aff(v_i, v_j) = sim_{cos}(v_i, v_j)$. The transition probability between v_i and v_j is defined by normalizing the corresponding affinity weight as follows:

$$p(v_i, v_j) = \begin{cases} \frac{aff(v_i,v_j)}{\sum_{k=1}^{|V|} aff(v_i,v_k)} & \text{if } \sum aff \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

We use the row-normalized matrix $M = M_{i,j}{}_{|V|\times|V|}$ to describe G_a with entry corresponding to the transition probability, i.e., $M_{i,j} = p(v_i, v_j)$. In order to make M be a stochastic matrix, the rows with all zero elements are replaced by a smoothing vector with all elements set to $\frac{1}{|V|}$. Based on the matrix M, the authority score of a paper *d* (denoted as $Authority(d)$) can be deduced from those of all other papers linked with it, which can be formulated in a recursive form as in the PageRank algorithm.

$$Authority(v_i) = \mu \sum_{j \neq i} Authority(v_j) \cdot M_{j,i} + \frac{1-\mu}{|V|} \quad (6)$$

where μ=0.85. We define the authority of an author a as:

$$Authority(a) = \sum_{d \in D_a} Authority(d) \quad (7)$$

where $D_a = \{d | author(d) = a\}$.

2.2.9 Venue Rank

Like authors, venues also have academic reputations. Based on our assumption, some venues have larger probability to be highly cited than others. We hereby investigate the venue impact on citations. Similar to the author rank pattern, prestigious venues attract more focus of researchers' attention. The reputation of a venue ensures the amount of citation as well.

2.2.10 Venue Centrality

Venues such as conferences or journals are connected by paper *citing-cited* linkage. We establish a venue connective graph $G_v(V, E)$ where V denotes the venues and the edges E denote the citing-cited relationships between venues. $G_v(V, E)$ also has directions: the out-degrees measure how many times a venue is cited by papers from other venues while in-degrees denote citations. The weight of each edge is calculated by the number of citations between two venues. Hence, the venue centrality can be calculated via a similar PageRank algorithm as Equation (6).

2.3 Predictive Models

2.3.1 Linear Regression (LR)

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. A linear regression line has an equation of the form **Y** = a + b**X**, where **X** is the explanatory variable and **Y** is the dependent variable. In our study, citation features are considered to be explanatory variables, and the predicted citation count is considered to be the dependent variable.

2.3.2 *k*-Nearest Neighbor (kNN)

The *k*-Nearest Neighbor algorithm is a method widely used in statistical estimation and pattern recognition for classifying objects based on closest training examples in the feature space by a majority common vote amongst its *k* nearest neighbors. The same method can be used for regression, by simply assigning the property value (in our case, citations) for the object (i.e., paper *d*) to be the average of the values of its *k* nearest neighbors to predict the value based on a similarity measure (e.g., distance functions such as cosine similarity). The neighbors are taken from a set of objects for which real citation counts are known.

Choosing the optimal value for *k* is best done by first inspecting the data. In general, a large *k* value is more precise as it reduces the overall noise; however, the compromise is that the distinct boundaries within the feature space are blurred. Based on performance tuning on the training set, we set *k*-NN as 5-NN empirically.

2.3.3 Support Vector Regression (SVR)

Statistical Learning Theory has provided a very effective framework for classification and regression tasks involving features. Support Vector Machines (SVM) are directly derived from this framework and they work by solving a constrained quadratic problem where the convex objective function for minimization is given by the combination of a loss function with a regularization term (the norm of the weights). There are two main categories for support vector machines: support vector classification (SVC) and support vector regression (SVR). SVM is a learning system using a high

dimensional feature space. It yields prediction functions that are expanded on a subset of support vectors.

The model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that is close to the model prediction. Support Vector Regression is the most common application form of SVMs. An overview of the basic ideas underlying support vector machines for regression and function estimation has been given in details in [16].

2.3.4 CART Model

We then fit a Classification and Regression Tree (CART) model [3], in which a greedy optimization process recursively partitions the feature space, resulting in a piecewise-constant function where the value in each partition is fit to the mean of the corresponding training data. Folded cross-validation [13] is used to terminate partitioning to prevent over-fitting. Our model included 10 features summarized in the last section as predictors.

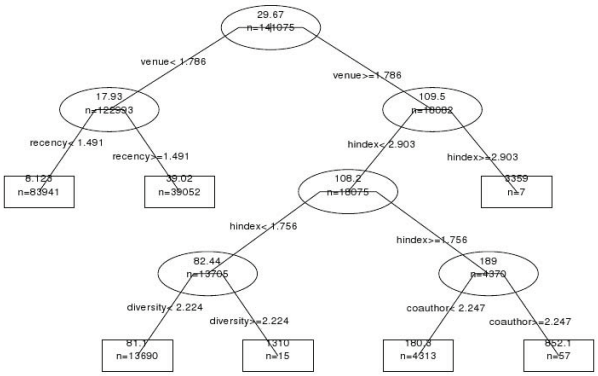


Figure 2: An example of regression tree for citation prediction.

Figure 2 shows the regression tree for one of the folds. Conditions at the nodes indicate partitions of the features, where the left (right) child is followed if the condition is satisfied (violated). Leaf nodes give the function value for the corresponding partition. Thus, for example, one of the leaves indicates that papers with *h-index*$\in$[1.756, 2.903) and *Sociality (NOCA)*$<$2.247 are predicted to have approximately 180 citation counts.

Thorough comparisons among all predictive methods and all features are examined in the experiments and evaluations.

3. EXPERIMENTS AND EVALUATION

3.1 Data Description

We perform citation prediction on the real-world data set[3], which is extracted from academic search and mining platform ArnetMiner [20]. It covers 1,558,499 papers from major Computer Science publication venues and has gathered 916,946 researchers for more than 50 years (from 1960 to 2011). The full graph of citation network contained in this data has 1,558,499 vertices (literature papers) and 20,083,947 edges (citations).

To predict the citation counts after one year, we randomly take 10,000 papers from the literature collection from Year 2009 as the test set, and another random 10,000 papers from the Year 2009 as the development set. Note that for all training and evaluation, we only used features calculated over previous years. For example,

[3]Downloaded from `http://arnetminer.org/citation`

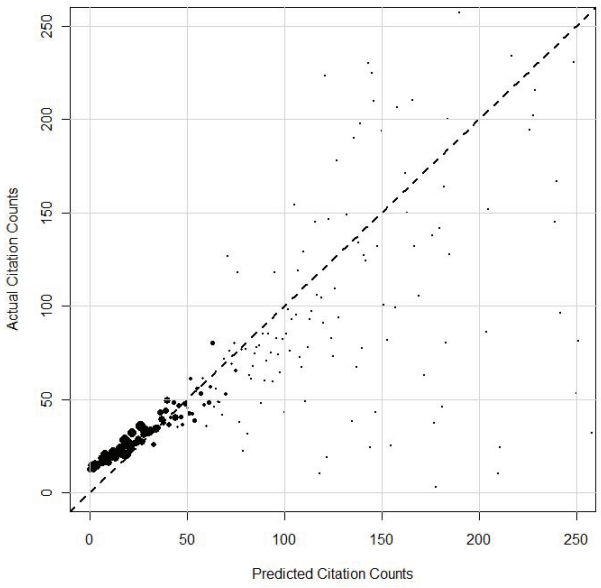


Figure 3: Actual vs. predicted citation counts: the performance for 10-Year CCP for Year 2000 with full features, regression = CART. The dotted line $y = x$ means the best result of predicted citation counts = actual citation counts.

when predicting articles published in Year 2009, all the articles up through Year 2008 are processed, and only the articles from the Year 2009 are available (as test set). Thus, these time dependent features would only include papers published in 2008 and earlier. Structuring the evaluation in this way is more realistic - when presented with new coming articles, the system can only predict possible future citations based on the patterns it has previously observed. We take the same procedure to predict citation counts after 5 (and 10) years with 10,000 test papers and 10,000 development papers from Year 2005 (and Year 2000). For unobserved feature values, e.g., new authors or new venues, we use the minimum feature values instead of N/A: anything has a start. We compare predicted citation counts with actual citations from the test data.

3.2 Evaluation Metric

The coefficient of determination R^2 is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of related features. It is the proportion of variability in a data set that is accounted for by the statistical model, which provides a measure of how well future outcomes are likely to be predicted by the model. The definition of R^2 is:

$$R^2 = \frac{\sum_{d \in D_T} (C_{T\,ccp}(d) - C_T(D_T))^2}{\sum_{d \in D_T} (C_T(d) - C_T(D_T))^2} \quad (8)$$

where $C_{T\,ccp}(d)$ is the predicted citations for article d in the test set D_T and $C_T(D_T) = \frac{1}{|D_T|} \sum_{d \in D_T} C_T(d)$ is the mean of the observed citation counts for an article in D_T. $R^2 \in$[0, 1], and a larger R^2 indicates better performance and hence is desired.

3.3 Performance and Feature Analysis

The best predictive performance of 10-Year citation count prediction is shown in Figure 3, and the detailed results are summarized in Table 2 and 3. The size of the circles in Figure 3 indicates the number of points in each predicted citation counts. Most circles

Table 2: The performance of various prediction techniques for different feature combinations on the test set. "+" indicates the single feature group in isolation while "-" indicates the drop of the feature group from the full combination.

	1-Year CCP (Δt=1)				5-Year CCP (Δt=5)				10-Year CCP (Δt=10)			
Methods	LR	kNN	SVR	CART	LR	kNN	SVR	CART	LR	kNN	SVR	CART
+Content	0.093	0.055	0.097	0.100	0.102	0.061	0.103	0.105	0.122	0.101	0.147	0.155
+Author	0.541	0.515	0.537	0.549	0.572	0.567	0.583	0.571	0.598	0.581	0.603	0.611
+Venue	0.274	0.208	0.322	0.315	0.301	0.296	0.317	0.332	0.321	0.313	0.355	0.373
-Content	0.646	0.628	0.671	0.697	0.679	0.632	0.691	0.705	0.689	0.693	0.711	0.723
-Author	0.279	0.245	0.285	0.303	0.296	0.317	0.364	0.372	0.394	0.387	0.402	0.418
-Venue	0.571	0.551	0.548	0.562	0.582	0.575	0.585	0.589	0.612	0.606	0.631	0.643
Combined	0.664	0.607	0.625	0.683	0.706	0.640	0.719	0.752	0.767	0.725	0.755	0.786

are gathered within in the range of [0, 50], indicating most of the papers have relatively low citations. The predicted citation counts will be overestimated for a short period of years. A possible explanation is that for papers with certain features (such as high *author rank*, high *venue rank*, etc.) are predicted to have high citations. To sum up, the system is not well performed in predicting short term impact but it is still of great significance because it is likely to estimate the long term citation counts for a paper more accurately, but the ultimate citations determine the achievements of literature.

Different predictive models have different performances on these three individual tasks in our experiments. In general, non-linear regression achieves better performance. From Table 2, we notice that kNN has the worst performance. The result is as expected because kNN merely seeks the most similar neighbors and takes the neighbors' citation counts as the predictive citations while utilizes little information from the enormous training data. LR, by linear combination of all features, and CART by non-linear regressions have comparable performances and proves the generality of our extracted features. CART performs best among these regression models in practice.

We then examine the different aspects of feature groups: paper content (feature 1-2), author expertise (feature 4-8) and venue impact (feature 9-10) in Table 2. Author expertise is proved to be the most influential feature group in citation count prediction, with the highest performance of R^2=0.611 in isolation and the lowest performance when left out from full feature combination. It is understandable that authors are likely to cite papers written by reputable and influential authors. Venue impact is also significant. Papers from prestigious venues are likely to be highly cited. Unexpectedly, paper content is proved to have the least significance, with the average performance of R^2=0.12 in isolation for CART. We assume (1) authors have biases to choose their bibliography: they sometimes merely consider author/venue reputation; (2) it seems that paper quality is represented by author/venue which create the paper. Influential authors or venues seem to overwhelm the impact of paper content itself; (3) it might also be due to the insufficient feature distilling for contents, e.g. using abstracts as approximation may not be enough for topic/diversity discovery.

We also conduct to a detailed experiment on all separate features in Table 3. We mark the most prominent performance of single features with asterisks in Table 3. The absence of *Author Rank*, *Venue Rank* and *H-Index* lead to unfavorable decrease.

Table 3: Feature analysis: R^2 result when with the pending feature ("+") in isolation (we mark the top 3 prominent features with asterisks), and the result in R^2 when dropped from the all-features model ("-").

Feature	+	−
Topic Rank	0.079	0.721
Diversity	0.157	0.645
Recency	0.101	0.677
H-Index	0.244*	0.536
Author Rank	0.486*	0.375
Productivity	0.198	0.613
Sociality (NOCA)	0.056	0.731
Authority	0.155	0.647
Venue Rank	0.337*	0.593
Venue Centrality	0.049	0.762

4. RELATED WORK

The measurement of citation count has long been a big concern for academia, and is heavily discussed by fundamental research journals (e.g. *Science*, *Nature* and *PNAS*) as further examinations of scientific achievements to distinguish significant ones. The yearly calculated *Impact Factor*, introduced by Eugene Garfield, is a measurement of citation counts of articles published in science journals and is still pervasive [6]. It is frequently used as a proxy for the relative importance of a journal within its field, with journals with higher impact factors deemed to be more important than those with lower ones and can be combined with other metrics such as popularity [17]. However, impact factor can not reflect the citations of individual papers [5, 15] and hence needs a normalization from the audience of citing sides [22].

As to author aspects, the *h-index* is a useful index that attempts to measure both the productivity and impact of the published work of a scientist or scholar [9, 8]. The index is based on the set of the scientist's most cited papers and the number of citations that they have received in other people's publications. H-index measures the impact of researchers and is directly related to publication citations.

However, both impact factor and h-index reflect the macro characteristics but the attractiveness of a specific collection (all papers from a particular author or venue) may be skewed by individuals. No previous work has focused on manipulation for individual papers, neither does any try to measure future citations of literatures. To the best of our knowledge, we are the first to formally research into future citation counts prediction for literatures.

Citation counts indicate the impact of authors, papers and venues, and several works have conducted to analyze citation behavior [1, 14] and perceive interesting discoveries. Sun *et al*. have investigated different impacts of author, venue and content features for clustering in these heterogeneous networks [18]. Through citation linkage, authors are found to affect to authors and paper contents [21, 19], and as well contents (such as topics) are influential to each other [12, 4]. We conduct to an extended examination of all

these factors correlated with citation counts, with many more new features added. There do exist several prediction works for the literature world based on citation features, such as co-author prediction [11] and citation linkage prediction by collaborative filtering [10]. Other applications include literature search/recommendation system based on features and citation behaviors [1, 7].

Unlike previous studies, we formally research into a new predictive task of citation count prediction and what is more, we add more relevant features into consideration.

5. CONCLUSION AND FUTURE WORK

In this paper we present a novel task of Citation Count Prediction (CCP), which predicts the future citations for publications. Given a particular paper and its corresponding features relevant with citation patterns (such as paper content, author expertise and venue impact), CCP predicts its possible citation counts. We formally formulate CCP task as a learning problem utilizing several regression models, and evaluate the prediction performance by coefficient of determination (R^2).

From our experiments, we find that authors have biases in citing references. Author expertise and venue impact are the distinguishing factors for the consideration of bibliography, among which, *Author Rank*, *Venue Rank* make paper attractive. Content features in isolation are not predictive. In general, the prediction after a longer period can achieve the best accuracy (R^2=0.786 when $\Delta t = 10$). Currently, we consider a particular paper itself without considering any of its audience (citing papers). However, the impact of audience can also be modeled because once a paper is cited by an attractive audience, it is likely to be attractive as well. As considering the audience will result in a *multi-step diffusion* problem and increase the complexity in measurement. In this study, we do not consider the audience's characteristics when measuring the popularity of the cited literature, while it can be further studied in the future.

6. ACKNOWLEDGMENTS

The work was supported by the Natural Science Foundation of China (Grant No. 60933004 and No. 61073073), HGJ Grant No. 2011ZX01042-001-001, Chinese National Key Foundation Research (No. 60933013 and No. 61035004). Particularly, Rui Yan was supported by the MediaTek fellowship.

7. REFERENCES

[1] S. Bethard and D. Jurafsky. Who should I cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 609–618. ACM, 2010.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] L. Breiman. *Classification and regression trees*. Chapman & Hall/CRC, 1984.

[4] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 233–240, New York, NY, USA, 2007. ACM.

[5] J. Dimitrov, S. Kaveri, and J. Bayry. Metrics: journal's impact factor skewed by a single paper. *Nature*, 466(7303):179–179, 2010.

[6] E. Garfield. Impact factors, and why they won't go away. *Nature*, 411(6837):522–522, 2001.

[7] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 421–430, New York, NY, USA, 2010. ACM.

[8] J. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569, 2005.

[9] J. Hirsch. Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, 104(49):19193, 2007.

[10] Z. Huang, X. Li, and H. Chen. Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '05, pages 141–142, New York, NY, USA, 2005. ACM.

[11] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[12] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 199–208, New York, NY, USA, 2010. ACM.

[13] R. Picard and R. Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984.

[14] A. Siddharthan and S. Teufel. Whose idea was this, and why does it matter? attributing scientific work to citations. In *HLT-NAACL*, pages 316–323, 2007.

[15] K. Simons. The misused impact factor. *Science*, 322(5899):165, 2008.

[16] A. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[17] Y. Sun and C. Giles. Popularity weighted ranking for academic digital libraries. *Advances in Information Retrieval*, pages 605–612, 2007.

[18] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 797–806, New York, NY, USA, 2009. ACM.

[19] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 807–816, New York, NY, USA, 2009. ACM.

[20] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 990–998, New York, NY, USA, 2008. ACM.

[21] D. Zhou, X. Ji, H. Zha, and C. L. Giles. Topic evolution and social interactions: how authors effect research. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 248–257, New York, NY, USA, 2006. ACM.

[22] M. Zitt. Citing-side normalization of journal impact: A robust variant of the Audience Factor. *Journal of Informetrics*, 4(3):392–406, 2010.