Citation Count Prediction: Learning to Estimate Future Citations for Literature

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 (R2). 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 R2, which significantly outperforms several alternative algorithms.

在大多数情况下，科学家依靠与他们的研究领域相关的先前文献来发展新思想。然而，追踪所有现存的出版物是不明智的，也不可能的，因为文学收藏的数量增长非常快。因此，研究人员通常只引用他们感兴趣的一小部分出版物。对于这样一个大的集合，预测哪一种文学更有可能吸引科学家的反应是相当有趣的。本文利用引文作为研究人员的受欢迎度的度量方法，研究了引文数预测(CCP)的有趣问题，以检验其受欢迎程度。估计可能的受欢迎程度是非常重要的，是相当具有挑战性的。我们利用了这些论文的一些基本特征，并预测了未来每种文学的受欢迎程度。我们已经实现了一个系统，它以某一特定出版物的一系列特性作为输入，并在给定的时间段内输出该文章的估计引用计数。我们考虑了几种回归模型以确定学习过程，并根据确定系数(R2)对其性能进行评估。实际大数据集的实验结果表明，最佳预测模型的平均预测性能为0.740，远远优于其他几种算法。

Categories and Subject Descriptors

分类和主题描述符

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

H.3[信息存储与检索]:内容分析与索引; I.2[人工智能]:自然语言处理--文本分析

General Terms

一般条款

Algorithms, Experimentation, Performance

算法,实验、性能

关键词

Citation count prediction, regression models, data engineering

引文计数预测、回归模型、数据工程

1. 引言

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 col- lection 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).

科学研究的快速发展每年都在创造大量的出版物，预计在可预见的将来还会继续存在。图1显示了计算机科学中的一个大型文献数据库的静态学。图1 . 1 .(a)在过去几年里，特别是近年来，可视化了公众的爆炸性增长。例如，2009年的出版物数量几乎是10年前的3倍。有效的科学研究需要与以前的文献保持一致，但对于研究人员来说，追踪所有现存的出版物是不明智的，也不可能的，因为文献的数量增长速度非常快。因此，研究人员通常只引用他们感兴趣的一小部分出版物。一个有趣的现象是，一些研究论文比其他论文更有可能吸引科学家的回复。如果我们在学术界使用引文作为论文的受欢迎度，我们就会得到如下图1(b)。

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=axb 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.

发现并不是所有的出版物都能引起学术界的重视，这是很自然的。在图1的log - log图中，我们展示了引文分布(论文和引用计数的数量):用引文计数衡量的文献的兴趣是高度倾斜的。毫不奇怪，情节遵循幂律分布。x和y之间的幂律关系可以写成y = axb其中A和b是常数。我们看到大量的研究论文只吸引了一些引用，一些研究论文积累了大量的引用。

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.

对于不断增长的文学作品，预测哪种文学更有可能吸引科学家的反应是相当有趣的。在本文中，我们使用引文作为一种简单的测量方法来衡量研究人员的受欢迎程度，而引文计数是通过其他文章引用某一特定出版物的次数来计算的。我们研究了引文计数预测(CCP)的有趣问题，研究了受欢迎度的相关特征。

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

作为一项关于学习预测未来文学引文的试点研究，引文数预测面临着几个挑战:

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;

CCP面临的第一个挑战是，从论文内容、作者专业知识和场地影响等几个方面，探索对未来引文重要性的真正有效特征。我们介绍了一系列与未来文学引文相关的特征;

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.

对于CCP来说，第二个挑战是将所有相关的特征结合起来，在一个统一的预测模型中，线性或非线性地识别潜在的有趣的论文。考虑到与流行相关的多个特性，即。在本研究中，我们利用多个回归模型来估计未来的引文。

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.

通过解决这些挑战，我们的贡献是多方面的。在第2节中，我们首先定义了一系列与引用计数相关的特性。然后将引文计数预测作为一个学习问题，引入几种回归模型来统一所有可能的预测特征。我们在第3部分描述实验和评估，包括性能比较和特征分析。我们简要回顾第4部分的前期工作，并在第5节中得出结论。

1. CITATION COUNT PREDICTION

引用计数的预测

2.1 问题定义

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

在这一节中，我们首先介绍了引文数预测问题的几个必要的定义和正式表示。

Citations. Given the literature corpus D, the citation counts (CT (.)) of a literature article d∈ D is defined as:

**引用。**鉴于文学语料库D,引用计数(CT())的文学∈D条被定义为:

citing(d)={d ∈D: d’ cites d}

CT (d) = |citing(d)|

Learning task: Given a set of article features, X⃗ =x1,x2,...,xn, our goal is to learn a predictive function f to predict the citation counts of an article d after a give time period ∆t. Formally, we have

学习任务:给定一组文章特征,X⃗= x1,x2,…xn,我们的目标是学习预测预测文章的引用计数函数f d后给时期∆t。形式上,我们有

f (d|X⃗ , ∆t) → CT (d|∆t)

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.

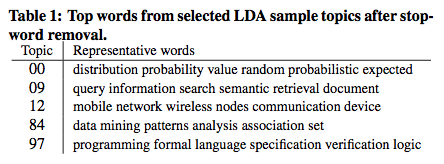
为了学习预测模型，我们研究了多个相关因素，如论文内容、作者的专业知识和场地的影响。同样重要的是要找到能够同时考虑所有特征的统一模型。我们将在以下小节中介绍这两个方面。

2.2 特征定义

2.2.1 话题排名

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 available2. 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.

课题长期以来一直被认为是文学内容[12]的一个重要特征。我们利用无监督的潜在Dirichlet分配[2]来发现我们的语料库的主题，因为它已经成功地应用于许多内容分析任务，并且实现是免费的。我们在我们的语料库上训练了一个100个主题的模型——一些样例主题的顶部单词如表1所示。



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|d), the inferred probability of topic i in document d. The topic distribution T (d) over all topics in document d is then:

我们的主题特征通过检查的概率分布在主题分配给一个文学文章d。也就是说,我们100年的话题,我们的话题模型计算p(主题i| d),话题我在文档的概率推断d。主题分布T(d)/所有文档d然后话题:

T (d) = {p(topic1 |d), p(topic2 |d), . . . , p(topic100 |d)}

To calculate the total citation counts of a particular topic from article d, denoted by CT (topici|d), we distribute the citations of the article CT (d) according to the topic distribution T (d), i.e., CT (topici|d) = CT (d) × p(topici|d) and hence we obtain the citations of all 100 topics by using:

为了计算由CT(topici | d)所表示的某一特定题目的总引文数，我们根据题目分布T(d)来分配文章CT(d)的引文，即。CT(topici | d)= CT(d)×p(topici | d),因此我们获得所有100的引用主题通过使用:

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

D是整个文献的集合。我们按平均引文来排列话题，即主题“受欢迎”。

2.2.2 多样性

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:

我们从其主题分布中获得一篇文章的宽度概念。这对于确定方法论论文是很重要的，这些论文经常被更广泛的文章所引用。当一篇文章有大量的听众时，可能会被来自不同研究领域的作者引用，从而吸引高的引用计数。为了测量一篇文章的局部宽度，我们计算了文档主题分布的熵:

2.2.3 Recency 近因

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.

时间维度长期以来被证明在文献研究中具有重要意义[1,21]。直觉上，引文计数随着时间的流逝而累积，因此，衡量一篇文章的年龄被认为是重要的。我们将这篇文章的出版时间作为一个专题。我们期望在时间上有积极的相关性——文章发表的时间越长，它可能收到的引用就越多。

2.2.4 H指数

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 out-put and impact of a researcher [9]. Therefore, we consider h-index as a candidate feature to predict citation counts.

h - index是用来衡量科学家[8]出版工作的生产力和影响的。该指数是根据科学家们引用最多的论文和其他出版物中引用的引用数量得出的。除此之外，h - index已经被证明具有科学的预测能力和研究人员[9]的影响。因此，我们认为h - index是预测引文计数的候选特征。

2.2.5 作者排名

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.

我们试图找出作者排名与平均引文数之间的相关性。有时候，作者名字的“名气”能保证引用的数量。每个作者都有自己对引文数的期望。我们根据他们的平均引数来计算所有的作者，并给他们每个人分配一个独一无二的排名。

2.2.6 产出率

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 as- sume the productivity of an author is relevant to the citation counts, due to the self-citation behavior analysis from previous studies.

根据[1]，作者倾向于引用他们自己写的论文。直觉上，作者越高产，他/她的论文被引用的机会就越大。因此，由于前人研究的自我引证行为分析，我们将作者的生产力与引文数相关联。

2.2.7 社会性

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.

在[1]的作者社会因素研究中，研究人员倾向于引用作者与作者合著的论文。因此，很自然地认为，来自一个广泛联系的作者的论文有更大的可能性被他/她广泛多样的合著者引用。一个简单而简单的度量方法是计算共同作者的数量(NOCA)，我们假设联合作者的数量和平均引用数之间的相关性。

2.2.8 权威性

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 pa- per authority to all its authors. We first build a graph of Ga(V, E), where V is the set of vertices and each vertex vi 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 be- tween 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., af f (vi , vj ) = simcos (vi , vj ). The transition probability between vi and vj is defined by normalizing the corresponding affinity weight as follows:

一个独特的学术社会网络是由“引用-引用”关系在文献文章中建立的。出版物带有作者权威:一个被广泛引用的论文指出了同伴的致谢，因此表明权威。我们将pa - per权限传送给所有的作者。我们首先构建了一个Ga(V,E)的图形，其中V是顶点的集合，V中的每个顶点vi代表一个文献，E表示引用的链接。引用的图表有方向。外度测量一篇论文被引用的次数，而在学位论文中则显示了某一篇论文的参考文献。当有一个引用的关系是- tween两篇文章时，我们在图中添加一个链接。我们用两篇论文的标准余弦相似度来衡量图中的连杆结构。，af f(vi,vj)= simcos(vi,vj)。vi和vj之间的过渡概率是通过对相应的亲合力权重的正常化来定义的:

We use the row-normalized matrix M = Mi,j|V |×|V | to describe Gawith entry corresponding to the transition probability, i.e., Mi,j = p(vi , vj ). 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 1 . Based on the matrix M, the authority score of a |V | 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.

我们使用row-normalized矩阵M = Mi,j | | V×V | |描述Gawith输入相应的转移概率,即。，Mi,j = p(vi,vj)为了使M成为一个随机矩阵，所有零元素的行被一个平滑向量替换，所有元素集合为1。基于矩阵M，可以从与之相关的所有其他论文中推导出| V |纸d(记为权威(d))的权威分数，可以用递归形式在PageRank算法中给出。

2.2.9 Venue Rank 场地等级

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.

与作者一样，场地也有学术声誉。基于我们的假设，一些场馆有更大的概率被高度引用。我们在此研究场地对引文的影响。与作者的排名模式相似，著名的场馆吸引了更多研究人员的关注。场地的声誉也保证了引文的数量。

2.2.10 Venue Centrality场地中心

Venues such as conferences or journals are connected by pa- per *citing-cited* linkage. We establish a venue connective graph Gv(V,E)where V denotes the venues and the edges E denote the citing-cited relationships between venues. Gv(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).

诸如会议或期刊之类的场所都是通过引用链接来连接的。我们建立了一个场地连接图Gv(V,E)，V代表场地，边E表示场地之间引用的关系。Gv(V,E)也有方向:外度数衡量一个地点被其他地点的文件引用的次数，而in - degrees表示引用。每条边的重量是由两个地点之间的引用次数来计算的。因此，可以通过类似于等式(6)的类似PageRank算法来计算场地的集中度。

2.3 Predictive Models

2.3.1 线性回归（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 + bX, 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.

线性回归尝试用线性方程拟合观测数据来模拟两个变量之间的关系。线性回归线有一个方程Y = A + bX，其中X是解释变量Y是因变量。在我们的研究中，引文特征被认为是解释变量，而预测的引文计数被认为是因变量。

2.3.2 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.

k近邻算法是一种广泛应用于统计估计和模式识别的方法，它以最接近于特征空间中最接近的训练样本为基础，在其k最近的邻居中进行多数共同投票。同样的方法也可以用于回归，只需为对象指定属性值(在我们的示例中是引用)。，论文d)根据相似度度量(例如:，)，使k最近邻居的值的平均值，以预测其值。，距离函数，如余弦相似度。邻居们都是从一组真实的引文计数所知道的对象中获得的。

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.

选择k的最佳值，最好先检查数据。一般来说，大k值更精确，因为它能降低整体噪声;然而，折衷的是特征空间内的明显边界是模糊的。基于训练集的性能优化，我们以5 - nn的经验进行了k - nn的设置。

2.3.3 支撑向量机

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 frame- work 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.

统计学习理论为分类和回归任务提供了一个非常有效的框架。支持向量机(SVM)是直接从这个框架中推导出来的，它们通过求解一个受约束的二次问题，即在正则化项(权值)的组合中得到凸目标函数的最小值。支持向量机主要有两类:支持向量分类(SVC)和支持向量回归(SVR)。SVM是一个使用高维空间的学习系统。它产生的预测函数在支持向量的子集上得到扩展。

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].

SVR生成的模型只依赖于训练数据的子集，因为构建模型的成本函数忽略了接近模型预测的任何训练数据。支持向量回归是SVMs最常见的应用形式。在[16]中给出了回归和函数估计支持向量机基本思想的概述。

2.3.4 CART模型

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.

然后我们拟合一个分类和回归树(CART)模型[3]，其中贪婪的优化过程递归地划分特征空间，导致一个分段常数函数，其中每个分区的值符合相应的训练数据的平均值。折叠交叉验证[13]用于终止标准，防止过度拟合。我们的模型包括在最后一节中总结的10个特征作为预测因子。

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*∈[1.756, 2.903] and *Sociality (NOCA)*<2.247 are predicted to have approximately 180 citation counts.

图2显示了其中一个折叠的回归树。节点的条件表示特征的分区，如果条件满足(违反)，左(右)子就会被跟踪。叶节点为相应的分区提供函数值。因此,例如,一个树叶表明论文的h -指数∈(1.756,2.903)和社会性(NOCA)< 2.247预计将有大约180引用计数。

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

在实验和评估中，对所有预测方法和所有特征进行了彻底的比较。

1. EXPERIMENTS AND EVALUATION 实验与评估
   1. 数据描述

We perform citation prediction on the real-world data set3, 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 net- work contained in this data has 1,558,499 vertices (literature papers) and 20,083,947 edges (citations).

我们在现实世界数据set3上进行引文预测，从学术搜索和挖掘平台ArnetMiner[20]提取。它涵盖了来自主要计算机科学出版物场所的1,558,499篇论文，并收集了916946名研究人员50多年(从1960年到2011年)。在此数据中所包含的引文网的完整图有1,558,499个顶点(文献资料)和20,083,947条边(引文)。

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, 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.

预测引用计数在一年后,我们随机从文献收集从2009年10000篇论文作为测试集,和另一个随机从2009年的10000篇论文,开发集。注意,培训和评估,在前几年我们只用特性计算。例如，当预测2009年发表的文章时，2008年全年的文章都是经过处理的，只有2009年的文章可用(作为测试集)。因此，这些时间依赖的特性只包括在2008年和更早发布的论文。以这种方式构造评估更加现实——当提出新的文章时，系统只能根据之前观察到的模式预测未来可能的引用。我们采用同样的方法，在2005年(和2000年)的5年(和10年)之后，对1万份测试论文和1万份发展论文进行了预测。对于未观测的特征值，例如。新作者或新地点，我们使用的是最小特征值，而不是N / A:任何事物都有开始。我们比较了预测引用和实际引用的测试数据。

3.2 Evaluation Metric

The coefficient of determination R2 is used in the context of statistical models whose main purpose is the prediction of future out- comes 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.

确定R2的系数是在统计模型的背景下使用的，其主要目的是预测未来——这是基于相关特征的。这是统计模型所解释的数据集的变异性的比例，它提供了一种方法，来衡量模型预测未来结果的可能性。

* 1. 表现及特征分析

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 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.

10年引数预测的最佳预测效果如图3所示，具体结果如表2和表3所示。图3中圆圈的大小表示每个预测值的点数个数。大多数的圆都是在[0,50]范围内聚集的，这表明大部分的论文都有相对较低的引用。预测的引文数将会在短时间内被高估。一种可能的解释是，对于具有某些特征的论文(如高作者排名、高地点排名等)，预计会有很高的引用。综上所述，该系统在预测短期影响方面并不是很好，但它仍然具有重要意义，因为它有可能更准确地估计长期的引文量，但最终的引用决定了文献的成就。

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 be- cause 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.

在我们的实验中，不同的预测模型对这三个单独的任务有不同的表现。一般来说，非线性回归具有更好的性能。表2中，我们注意到kNN的性能最差。结果就像预期的那样，因为kNN只是寻找最相似的邻居，并将邻居的引文作为预测引用，同时利用了大量的培训数据的信息。LR，通过所有特征的线性组合，以及非线性回归的CART具有类似的性能，并证明了我们提取的特征的通用性。在实践中，CART在这些回归模型中表现最好。

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 R2=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 R2=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.

然后我们研究专题小组的不同方面:论文内容(特征1 - 2)，作者专长(特征4 - 8)和场地影响(特征9 - 10)在表2。在引文计数预测中，作者的专业知识被证明是最有影响力的特征群，其最高的表现为孤立的R2 = 0.611，而在完整的特征组合中，则表现最差。可以理解的是，作者可能引用由著名作家和有影响力的作家写的论文。场地的影响也很重要。来自著名场所的论文可能会被高度引用。出乎意料的是，论文的内容被证明具有最不重要的意义，而R2的平均性能则为0.12。我们假设(1)作者有偏见选择他们的参考书目:他们有时仅仅考虑作者/地点的声誉;(2)论文的作者/地点似乎代表了纸张质量。有影响力的作家或场所似乎压倒了纸张内容本身的影响;(3)也可能是由于内容的提炼不足，例如使用摘要作为近似值，可能不足以引起主题/多样性的发现。

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.

我们还对表3中所有单独的特性进行了详细的实验。我们在表3中标记了带有星号的单个特性最显著的性能。作者排名、场地排名和h指数的缺失导致了不利的下降。

1. 相关工作

The measurement of citation count has long been a big concern for academia, and is heavily discussed by fundamental re- search 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].

引文计数的测量长期以来一直是学术界关注的焦点，并被基本的研究期刊(如科学、自然和PNAS)进行了大量的讨论，以进一步检验科学成果来区分重要的文献。由尤金·加菲尔德(Eugene Garfield)介绍的每年计算的影响因素，是一项在科学期刊上发表的文章的引用计数的测量，仍然是普遍的[6]。它经常被用来作为在其领域内的期刊的相对重要性的代表，有更高的影响因素的期刊被认为比那些较低的期刊更重要，并且可以与其他指标(如流行的[17])相结合。然而，影响因素不能反映个别论文的引用[5,15]，因此需要从听众中引用[22]。

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].

由尤金·加菲尔德(Eugene Garfield)介绍的每年计算的影响因素，是一项在科学期刊上发表的文章的引用计数的测量，仍然是普遍的[6]。它经常被用来作为在其领域内的期刊的相对重要性的代表，有更高的影响因素的期刊被认为比那些较低的期刊更重要，并且可以与其他指标(如流行的[17])相结合。然而，影响因素不能反映个别论文的引用[5,15]，因此需要从听众中引用[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.

至于作者方面，h - index是一个有用的指标，它试图衡量科学家或学者发表的著作的生产力和影响[9,8]。该指数是根据科学家们引用最多的论文和他们在其他人的出版物中收到的引用数量而建立的。h指数衡量研究人员的影响，直接关系到发表的引用。

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.

然而，影响因素和h指数都反映了宏观特征，但某一特定集合(来自某一作者或地点的所有论文)的吸引力可能会被个人曲解。以前的工作都没有把注意力集中在个人论文的操作上，也没有试图衡量未来的文献引用。就我们所知，我们是第一个正式研究未来引文计数预测的文献。

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].

引文计数显示了作者、论文和地点的影响，以及一些作品对引文行为进行了分析[1,14]，并感知有趣的发现。Sun等人研究了在这些异构网络[18]中，作者、地点和内容特性对集群的不同影响。通过引文链接，发现作者对作者和论文内容的影响[21,19]，以及内容(如主题)对彼此有影响[12,4]。我们对所有这些与引文计数相关的因素进行了进一步的检查，并增加了更多的新功能。基于引文特征的文献世界存在着几种预测方法，如协同过滤[10]的联合预测[11]和引文链接预测。其他应用包括基于特征和引文行为的文献搜索/推荐系统[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.

与以往的研究不同，我们对引文计数预测的一项新的预测任务进行了正式的研究，并在考虑中加入了更多相关的特征。

1. 结论及未来工作

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 (R2).

本文提出了一项新的文献引用预测任务，该任务预测了未来出版物的引用数量。鉴于特定的论文及其与引文模式相关的特征(如论文内容、作者的专业知识和场地影响)，CCP预测其可能的引用计数。摘要利用多元回归模型，将CCP任务作为一个学习问题进行正式表述，并通过确定系数(R2)对预测绩效进行评价。

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 (R2=0.786 when ∆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.

从我们的实验中，我们发现作者在引用参考文献时存在偏见。作者的专业知识和场地的影响是对书目的考虑因素，其中，作者排名、场地排名使论文更具吸引力。隔离的内容特性是无法预测的。一般来说,预测在较长时间内可以达到最好的精度(R2 = 0.786时∆t = 10)。目前，我们考虑的是一篇论文，但没有考虑到它的任何读者(引用论文)。然而，观众的影响也可以被模仿，因为一旦一篇论文被吸引的观众引用，它也很可能是有吸引力的。在考虑受众的同时，会导致多步扩散问题，增加了测量的复杂性。在本研究中，我们不考虑受众的特征，在测量引证文学的受欢迎程度时，在未来可以进一步研究。