Collaboratively Learning Latent Factors and Correlations for New Paper Influence Predication

Lei Shen
Shandong Unversity
Jinan, China 250100
Email: shenlei162@163.com

Yuqing Sun
Shandong Unversity
Jinan, China 250100
Email: sun_yuqing@sdu.edu.cn

Xin Li
Shandong Unversity
Jinan, China 250100
Email: lx@sdu.edu.cn

***Abstract*—There are an increasing number of papers published every year. It is desired for researchers to find the new high-quality papers, which is also a challenging task due to the lack of citation information. In this paper, we propose a novel method to predicate a new paper influence by collaboratively learning the latent vectors of paper features and correlations. We propose the concept *topic related authority* to integrate the dynamic topic model with paper citations so as to learn how content and authors influence a paper quality. We adopt the Factorization Machine method to collaboratively learn the latent vectors of correlations between different paper features. Comparing with traditional methods, it does not require the citation information to evaluate a paper quality, which is appropriate for new published papers. We conduct extensive evaluation against a real dataset crawled from ACM Digital Library. The results show that our method outperforms the other methods.**

I. Introduction

There is an increasing number of papers published every year. It is important for researchers to find the related and high quality papers. However, according to the report by Garfield[11], only about 20% papers received more than 80% citations, which are the acknowledgment by the academia, and the other papers are rarely cited or even never cited. So how to recognize the promising new papers before it receives a lot of citations is a meaningful task.

The current related works on paper influence predication focus on two aspects: the related features extraction and the predication methods. The often adopted factors include authors, topics, publication venues, citations and etc. The citation information are regarded as the most important element for judging the quality of a paper and user authority. For example, the number of citations is often chosen as an intuitive quantitive metric, such as H-index proposed by Chakraborty et al.[4] being used to measure both the productivity and the impact of an author. A modified method considered the citation network, where the node set are papers and the edge set are the paper citations. Xie et al.[13] and Yan et al.[14] compute the influence of a paper by the PageRank algorithm and an author's authority is as the sum influence of all her publications. Although this method well justifies paper quality and authority, it is not appropriate for predicating a new published paper influence since there is not available citation information. Another shortcoming is that the authority is quantified as one score, which can not reflect an author's ability on different disciplines. For example, a professional is famous in physics does not imply he could write a good quality paper on chemistry if he/she never published any related papers.

The often adopted prediction methods include the classification and regression. For example, Chakraborty et al.[4] take the Support Vector Machine method to classify papers into six categories on citation trend and then employ the Support Vector Regression method to predicate the citations if a paper is in an ascend trend. Yan et al. [14] adopt two methods for predicting paper citation number, namely Gaussian Process Regression as well as Classification and Regression Tree, respectively. But they do not learn the latent correlations among the features. Another representative method is to use the post-publication information as training data and to use reinforced Poisson processes for predicating future citation counts [10]. Xiao et al.[12] propose a self-triggering process, namely Hawkes process, for long-term paper citation prediction. They take the citations of a paper in first period of several years after it was published as the training window, and predicate a paper's citation in later period. Obviously, it does not work for a new paper without any citation.

To tackle the changeling problem of a new paper influence predication, we propose a collaborative learning method for the latent vectors of paper features and correlations. The dynamic topic model is adopted to extract publication content and predicate a topic hotness. The topic related authority is proposed to justify how a paper content and user authority influence a paper quality, which is learned from one's historical publication focusing on topics and citations. We also take into account social theory to analyze the coauthor network so as to learn inner-clique and inter-clique citation behaviors. Each factor is represented as a latent vector and the Factorization Machine method is adopted to collaboratively learn the latent correlations between them. Then a new paper influence is predicated based on the author information, the paper content, and the publication venue. We conduct extensive evaluation against a real dataset crawled from ACM Digital Library. The results show that our method outperforms the other methods.

The paper is organized as follows. Section II presents the related works of influence prediction on papers. In section III, we present the formal definition of the problem and framework. Section IV introduces the feature extraction of a

978-1-5386-1482-2/18/$31.00 ©2018 IEEE

paper. Section V discusses the prediction model on a new paper influence. In section VI, we discuss the experimental results and make comparison with other methods. We conclude the paper in Section VII.

II. Related Work

A. Features Engineering

1) Author-centric Features: An author with high reputation in certain domain have many high quality papers, which are more likely to be cited by others. Bjarnason et al.[1] evaluated the authority by citation count. They thought that highly cited authors are more likely to achieve citations, documenting the "rich-get-richer" phenomenon. H-index measuring both productivity and citations was also taken for evaluating authors' reputation[4], [6], [14], [15]. However, these methods can't measure the topic-related authority and how authority affect a paper's quality. In this paper, we introduce the topic related authority in a fine-gained way and justify how a paper content as well as authority influence a paper's influence.

Another factor affecting a paper's popularity is the author's social relations. Tahamtan et al.[11] found that researchers used to cite the paper published by his cooperative partners. People who often cooperate with each other tend to form a clique and share similar interests. Thus, the stronger the social ties are, the higher probability they cite each other in some clique. The number of co-authors is taken to evaluate social features[4], [6], [14], [15], which is unable to measure strength of social ties. In this paper, we apply the social theory to analyze social factors from inner-clique and inter-clique aspect.

2) Content-centric Features: A paper's popularity is highly related to the content. Latent Dirichlet Allocation(LDA), Dynamic Topic Model(DTM) are widely taken to extract content-centric features[4], [6], [14] . LDA is a Bayesian network learning the static topic-specific word distribution. DTM is a LDA variant learning dynamic topic distribution. In this paper, we take DTM to learn the dynamic topic distribution due to the evolution of contents and word usage.

It's a common sense that popular topic catch much citation counts. Expectation citation number of the topic is taken as the indicator to estimate the hotness of the topic[5], [6], [14]. However, on one hand, the less popular topic may be mistaken for popular ones because of its outdated citation, on other hand, no all papers concerning hot topic have high quality. In this paper, we leverage topic distribution to estimate topic hotness in time and combine hotness with topic related authority to learn paper influence.

3) Venue-centric Features: The papers published in prestigious venues tend to be with high quality since all these papers have been evaluated by professionals. Yan et al.[14]implemented the PageRank algorithm in the venues' citation network for venues' reputation evaluation. We take the H5-index as the indicator of the venue, which is the H-index for venue published in the last 5 years.

B. Influence Prediction

Castillo et al.[3] implemented linear regression and C4.5 decision tree by combining priori and posteriori features. Chen et al.[5] proposed content, author features and introduced Random Forests(RF), the Gradient Boosted Regression Trees(GBRT), and the initialized Gradient Boosted Regression Trees(iGBRT) to predict the citation counts of papers. Yan et al.[15] combined all relevant features to identify the interesting papers in several regression models, including Linear Regression(LR), k-Nearest Neighbor(kNN), Support Vector Regression(SVR) and Classification and Regression Tree(CART). In this follow-on work, Yan et al.[14] integrated the factor of future influence into Gaussian Processes Regression. Chakraborty et al.[4] developed two-stage prediction model-Support Vector Machine(SVM) identified the papers' category and then SVR predicted future citation counts. With the help of post-information, Shen et al.[10] used reinforced Poisson processes and Xiao et al.[12] proposed a self-triggering process namely Hawkes process for long-term paper citation prediction.Yu et al.[16] took this problem as a link prediction problem, which is suitable for personalized recommendation rather than influence prediction.

All these method don't take topic-related authority and latent correlations among features into consideration. In this paper, we evaluate authority in a fine-gained aspect and learn latent correlations among features for predicting new papers' influence effectively.

III. Problem Definition and Framework

In this section, we would present the formal definition of the problem and the architecture of our model.

Problem 1: Given a paper corpus **P** as a training set, a set of new papers(N_P), an evaluation metric φ and a parameter k as the preference, the new paper influence prediction problem is to select the top-k influential papers $I_p \subset N_P$ such that $\forall p \in I_p$, $\forall p' \in N_P \cap p' \notin I_p$, $\varphi(p) > \varphi(p')$.

The proposed framework is illustrated in Fig.1, which includes two parts: feature extraction and collaborative learning process. In the first part, we take into account the information available in a paper and extract four important kinds of features: the hotness of topics, topic-related authority, author social features and venue-centric features.

Considering a paper context, the popularity of topics is an important indicator. A new paper concerning a hot topic is likely to attract more attention. We adopt the Dynamic Topic Model(DTM)[2] to learn dynamic topic distribution and learn its hotness by paper-specific distribution over these topics.

Different with previous works, we evaluate an author's authority in a fine-gained aspect. PageRank(PR) algorithm is adopted to quantify a paper reputation over the created citation network. Papers' reputation along with topic distribution are combined to learn topic-related authority. Then we introduce the topic-related authority as a distributed vector to project the overview qualification of one's subsequent paper quality against topics. The combination of hot topics learned by

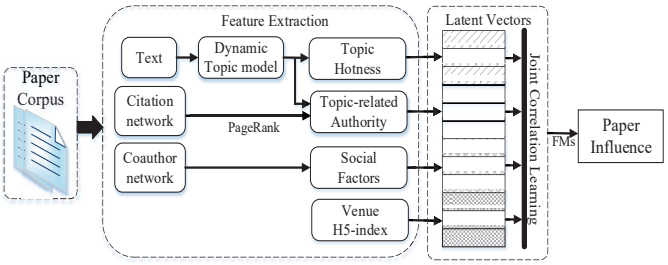

Fig. 1. The proposed framework for influence prediction

DTM and topic related ability well illustrate how one's pre-existing authority affect his/her following papers. Comparing with related works that assess the authority as citations only by a paper topic distribution and author expertise, we distinguish the ability on different topics, which better understand the correlations.

Taking into account the social theory, researchers tend to cite the papers produced by their partners or coauthors, we analyze the authors' social factors from the inner-clique and inter-clique aspects. Since the reputation of a journal or a conference is an important indicator of the publications quality, we adopt the H5-index, which is taken as an venue feature in predicating a paper influence.

In the second part, we use the latent vectors to represent each dimension of a paper features and collaboratively learn their correlations by Factorization Machines(FM).

IV. Features Engineering

A. Dynamic Hotness of Topics

To learn how a paper content influences a paper popularity, we adopt the topic model to represent a paper content as the latent vector of topics. In a topic model, the words of each document are assumed to be independently drawn from a topic-specific proportions. Considering that the popularity of topics is an important indicator to attract researchers' attention, we adopt Dynamic Topic Model(DTM)[2] to learn its topic evolution and hotness. It specifies a statistical model of topic evolution and develops efficient approximate posterior inference techniques for determining the evolving topics.

Generally, a sequential paper corpus **P** is temporally organized by year. Hence, we separate the corpus into *T* periods by year. The papers published in period $t \in [1, T]$ are denoted as $P^t \in \mathbf{P}$. DTM considers how topics related to period *t+1* evolve from topics related to period *t*. In period *t*, DTM models papers in P^t with *Z* topics over *W* words, where *Z* is the number of topics and *W* is the number of words over **P**. Let $\vec{\alpha}^t$ denote prior parameters of paper-specific topics distribution in period *t*. The *z*th-element α_z^t in $\vec{\alpha}^t$stands for prior parameter specific to topic *z*, $z \in [1, Z]$. $\vec{\beta}^t$ represents the topics-specific word distribution. Its *z*th-element β_z^t in $\vec{\beta}^t$ stands for topic distribution specific to topic *z*. DTM chains these periods and parameters sequentially in a state space model that evolves with Gaussian noise(δ^2, σ^2) which satisfies $\alpha_z^t|\alpha_z^{t-1} \sim N(\alpha_z^{t-1}, \delta^2)$ and $\beta_z^t|\beta_z^{t-1} \sim N(\beta_z^{t-1}, \sigma^2)$. The generative process for P^t is as follows:

1) Draw topics $\vec{\beta}^t|\vec{\beta}^{t-1} \sim N(\vec{\beta}^{t-1}, \sigma^2 I)$.
2) Draw $\vec{\alpha}^t|\vec{\alpha}^{t-1} \sim N(\vec{\alpha}^{t-1}, \delta^2 I)$.
3) For each document:
   a) Draw $\eta \sim N(\vec{\alpha}^t, a^2 I)$
   b) For each word:
      i) Draw $z \sim Mult(\pi(\eta))$
      ii) Draw $w_{p_i,n}^t \sim Mult(\pi(\beta_z^t))$

where η is the paper-specific topic distributions drawing from Gaussian distribution with α^t as mean and a hyper parameter a^2 as variance. $w_{p_i,n}^t$ is *n*th word in p_i, $n \in [1, N_{p_i}]$, N_{p_i} is the number of words in paper p_i, $p_i \in P^t$.

For topic distribution $\vec{tp}_i$ of paper p_i, its *z*th-element $tp_{i,z}$ is considered to be related to p_i when the probability of *z*th topic is higher than *1/Z*. Thus, the hotness of a topic in period *t* can be computed as follow:

$$ht_z = \frac{|\{p_i | p_i \in P^t, tp_{i,z} \geqslant \frac{1}{Z}\}|}{|P^t|} \tag{1}$$

B. Influence of Topic-related Authority on Paper

For a given paper corpus *P*, we construct the citation network *GC=(P,E)*, where *E* stands for citation links between the papers in *P*. An author with high reputation in a certain area is more likely to be followed by other researchers. Let P_{a_j} denotes the set of papers published by a_j. For each $p_i \in P_{a_j}$, PageRank algorithm is adopted to learn the influence of p_i, denoted as pr_{p_i}. An author's overall authority is the summary of his/her publications' influences. Thus, the topic-related authority of an author a_j is denoted as $\vec{A}_{a_j} = (A_{a_j,1}, ., A_{a_j,Z})$, *z*th-element $A_{a_j,z}$ in $\vec{A}_{a_j}$ represents author a_j's authority on topic *z*, which can be calculated as follow:

$$A_{a_j,z} = \sum_{p_i \in P_{a_j}} tp_{i,z} * pr_{p_i} \tag{2}$$

The quality of a new paper is affected by an author's pre-existing authority and the paper's topic distribution. We use the cross product to evaluate the latent effect of an author's topic-related authority on his/her new paper. So the influence of the a_j's topic-related authority on his/her new paper p_i ($\overrightarrow{IA}_{a_j}$ for short) as follow:

$$\overrightarrow{IA}_{a_j,p_i} = \vec{tp}_i \times \vec{A}_{a_j} \tag{3}$$

For the rule of cross product, we select top 3 topic probability of papers and authority relating to these topic. For a paper p_i, we measure the expectation of p_i's authors' features.

C. Author-related Social Features

To analyze author social features, we construct the co-author network denoted by $GP = (A, E)$, where *A* is the set of author nodes and *E* is the co-authorship between nodes.

1) Triadic Closure: In sociology, if two people in a social network have a common friend, then they are more likely to become friends at some period in the future, which is called as triadic closure. We use the *Local Clustering Coefficient(LCC)*[17] to quantitatively measure the closeness of neighbors to a clique as follow:

$$LCC_a = \frac{2 * |\{e_{b,c} : b, c \in N_a\}|}{|N_a| * (|N_a - 1|)} \tag{4}$$

where $a, b, c \in A$, N_a is the neighbors of author a, $e_{b,c} \in E$ is the edge between author *b* and author *c*. *LCC* can be used to evaluate the inner-clique citing behavior.

2) Embeddedness: In sociology, a high embeddedness score represents trust and confidence, and the presence of mutual friends reduces the chance of misbehavior[7]. It can be used to estimate how likely an author's papers will be cited by his co-authors. We define the embeddedness of an author a as:

$$Emb_a = \frac{1}{|N_a|} \sum_{b \in N_a} \frac{|N_a \bigcap N_b|}{|N_a \bigcup N_b|} \tag{5}$$

3) Structural Hole: An individual who acts as a mediator between two or more closely connected clique of people often gain important comparative advantages[1]. It can be used to mine the inter-clique citing behavior. We use the HIS algorithm [8] to estimate the structural hole score. **C**=$\{C_1, .., C_l\}$ denotes l clique in co-author network. $I(a, C_i) \in [0, 1]$ is the importance of *a* in clique C_i and $H(a, S) \in [0, 1]$ is the structural hole score of *a* in *S*, $S \subseteq \mathbf{C}$, $|S| \geq 2$.

$$I(a, C_i) = \max_{\substack{e_{a,b} \in E, \\ C_i \in S}} \{I(a, C_i), \alpha_i I(a, C_i) + \beta_s H(b, S)\} \tag{6}$$

$$H(a, S) = \min_{C_i \in S} \{I(a, C_i)\} \tag{7}$$

where α_i, β_s are two tunable parameters.

D. Venue-centric features

The papers accepted by top venues imply that these papers with high quality are all recognized by experts. Thus, we also take venue factor into consideration and use h5-index to evaluate the venues, which is defined that a venue with an index of *h* has contained *h* papers each of which has been cited in other papers at least *h* times within the last 5 years.

V. Learning and Prediction

For $p_i \in \mathbf{P}$, we use the aforementioned features to form a *k*-dimensional feature vector $\vec{p_i} \in \mathbf{R}^k$. Factorization Machine(*FM*) is adopted to learn latent vectors and estimate interactions between each dimension of features[9]. In detail, for each pair of variables $p_{i,j}, p_{i,j'}$ in $\vec{p_i}$, two *h*-dimensional latent vector $\vec{S_j}$, $\vec{S_{j'}} \in \mathbf{R}^h$ are used to represent latent space respectively. $< \vec{S_j}, \vec{S_{j'}} >= \sum_{l=1}^{h} s_{j,l} s_{j',l}$ models the interactions between two variables with the dot product and **S**$\in \mathbf{R}^{k*h}$. In order to predict a paper p_i's influence in period *t*, the prediction function is computed by:

$$\begin{aligned} \hat{y}(\vec{p_i}) &= \varphi(\vec{p_i}|\Phi, t) \\ &= \omega_0 + \sum_{j=1}^{k} \omega_j p_{i,j} + \sum_{j=1}^{k-1} \sum_{j'=j+1}^{k} p_{i,j} p_{i,j'} < \vec{S_j}, \vec{S_{j'}} > \end{aligned} \tag{8}$$

where $\omega_0 \in \mathbf{R}$ is the global bias and $\omega_i \in \mathbf{R}$ is the importance of *i*th-dimension of new paper's features. $\Phi = \{\omega_0, \omega_1, ..., \omega_k, s_{1,1}, .., s_{k,h}\}$ denotes the parameter set of prediction function.

By applying dot product between two latent vectors into Eq.8, the equation can also be rewritten as:

$$\hat{y}(\vec{p_i}) = \omega_0 + \sum_{j=1}^{k} \omega_j p_{i,j} + \frac{1}{2} \sum_{l=1}^{h} [(\sum_{j=1}^{k} s_{j,l} p_{i,j})^2 - \sum_{j=1}^{k} s_{j,l}^2 p_{i,j}^2] \tag{9}$$

In order to minimize the error between real value and predicted value, least-square loss function is adopted and L2 regularization is employed to overcome the overfitting brought by the large number of parameters. The object of model is defined as follow:

$$\Phi^* = arg \min_{\Phi} \{ \sum_{p_i \in P} (\hat{y}(\vec{p_i}|\Phi) - y_i)^2 + \lambda \sum_{\theta \in \Phi} \theta^2 \} \tag{10}$$

where y_i is the paper citations, λ is a parameter that controls the regularization value.

The partial derivative of $\hat{y}(\vec{p_i})$ can be written as:

$$\frac{\partial \hat{y}(\vec{p_i})}{\partial \theta} = \begin{cases} 1 & \theta = \omega_0 \\ p_{i,j} & \theta = \omega_j \\ p_{i,j} \sum_{j \neq l} s_{j,l} p_{i,j} & \theta = s_{j,l} \end{cases} \tag{11}$$

The learning algorithm is given in Algorithm 1. Stochastic gradient descent(SGD) is adopted and performs updates on the model parameters:

$$\theta \leftarrow \theta - 2\eta[(\hat{y}(\vec{p_i}) - y_i) \frac{\partial \hat{y}(\vec{p_i})}{\partial \theta} + \lambda \theta] \tag{12}$$

$\eta \in \mathbf{R}$ is the learning rate for gradient descent.

Algorithm 1 Learning Algorithm for FM

Input: a paper corpus **P**, co-author network GP=(A,E), citation network GC=(P,E)
Output: Φ=$\{\omega_0, \vec{\omega} \in R^{1*k}, \vec{S} \in R^{k*h}\}$

```
ω⃗ ∼U(0,2), S⃗∼N(0,σ)
Learning each paper p ∈ P topic distribution using DTM
for p ∈ P do
    Calculate paper's features according to Eq.1-Eq.7
end for
repeat
    for p ∈ P do
        p_i ← a random paper drawn from P
        Compute gradients of ω_0, ω⃗, S according to Eq.11
        Update the above parameters according to Eq.12
    end for
until Convergence
```

[1] https://en.wikipedia.org/wiki/Structural_holes

VI. Experiments

A. Dataset

We crawled the publicly available dataset from ACM Digital Library[2]. We select *Databases & Information Systems(DB)* and *Human Computer Interaction(HCI)* as candidate areas. Three representative top conferences are selected from DB and HCI area respectively according to google scholar category[3](SIGMOD, SIGIR and CIKM for DB. CSCW, UIST and UbiComp for HCI). The crawled dataset contains about 10505 papers on DB, 4184 papers on HCI and 23261 authors.

The co-author network and the citation network are constructed according to the crawled dataset. The graph of co-author network has 79897 links(co-authorship) and 23261 nodes(authors). The graph of citation network has 222103 links(citation) and 10794 nodes(papers). Moreover, each paper contains a title, an unique index, author(s), publication year, venue, an abstract and reference of a paper. Each author includes author's name, a unique index, publication counts, citation counts, average citations per paper.

B. Methods for Comparison

- *kINP*. Our model finding the top-k influential new papers(*kINP*) integrates aforementioned features into FM model.
- *kINP-HT(AT,SO)*. *kINP* model drops the feature of hotness of topics(topic-related authority, social features) from all combination.
- *YAN*. Yan et al.[14] analyze the paper's features with the same three categories features to us. Gaussian Process is adopted for prediction.
- *TC*. Chakraborty et al.[4] develop a two-stage prediction model by using SVM and SVR together and inputting features extracted in [4] to predict the future citation count of the papers.
- *SVR*. SVR stands for Support Vector Regression model. We take the features extracted by our method as inputs to *SVR*.

C. Evaluation Metric

Two metrics are taken to justified the results of the experiments, which are often adopted by other works[4], [14].

- *Coefficient of determination* R^2. R^2[14] is used in the context of statistical models whose purpose is the prediction of future outcomes on the basis of paper related features. It is defined as follows:

$$R^2 = \frac{\sum_{p\in P_T}(\hat{c}^t(p) - c^t(P_T))^2}{\sum_{p\in P_T}(c^t(p) - c^t(P_T))^2} \tag{13}$$

  where $\hat{c}^t(p)$ is the predicted citations for article p until time *t* in the test set P_T and $c^t(P_T) = \frac{1}{|P_T|}\sum_{p\in P_T} c^t(p)$ is the mean of the observed citation counts for a paper until time *t* in P_T. $R^2 \in$[0,1], and R^2 approaching to 1 indicates a better performance.

- *Normalized Discounted Cumulative Gain(NDCG)*. This is widely used in information retrieval and it measures the quality of ranking. We take the *NDCG* value of top k I_p selected by model as the qualitative evaluation metric. *NDCG* is computed as following:

$$NDCG = \frac{DCG_p}{IDCG_p} \tag{14}$$

$$DCG_p = \sum_{p\in P_K} \frac{1}{\log_2(rank_p + 1)} \tag{15}$$

$$IDCG_p = \sum_{i=1}^{k} \frac{1}{\log_2(i+1)} \tag{16}$$

  where P_K is the top k I_p selected by models from N_P and $rank_p$ is the true position of paper p.

In order to predict papers' influence, we respectively predict a new paper's influence in period *t* after they're published, $t \in \{1, 2, 3, 4, 5\}$ and take 5-fold cross validation into consideration, *t* for year after publication. For example, to predict papers' citation when *t*=1, the paper set published in year 2010 are randomly partitioned into 5 subset. We randomly take a single subset as the validation set of test model and the remaining 4 subsets along with the papers published before 2010 are taken as the training set. The citation that papers in validation set receive in 2011 are taken as the ground truth. Besides, top k I_p in period *t* are selected according to prediction of citation. We take the same strategy when *t*=2(3,4,5).

D. Parameters Setting and Performance Analysis

In our experiments, we set the global bias ω_0=2 and the importance of *i*th-dimensional features ω_i=2, dimensionality for latent vector *h*=1000, learn rate η=0.005, and standard deviation for initialization of latent vector σ=0.01.

We conduct experiments to compare *kINP* model with baseline models. Table I shows the comparison between *kINP* and three baseline methods in the terms of R^2 and *NDCG* in different time period. Our model has a good performance in this dataset. With regard to *NDCG*, our model is more likely to find I_p from N_P. The comparison results between *SVR* and *kINP* show that learning latent relationship is helpful to this task.

In Fig.2, we further analyze *kINP* in different periods and how much a feature contributes to paper influence prediction in terms of R^2. As time goes on, the R^2 is gradually close to 1 and has a slight growth after *t*=4. The values in $t = 1$ shows that social features and hotness of topic are the main factor to new paper's influence and topic-related authority has the least impact(*kINP-SO*:0.105, *kINP-AT*:0.127, *kINP-HT*:0.112). However, topic-related authority becomes more and more important to papers' influence and social features' and hotness of topics' impact decline as the time going on.

In the terms of *NDCG* k=20, different features contribute more or less in different time period as showing in Fig.3. For example, social features have less contribution between $t = 2$

[2]http://dl.acm.org/

[3]https://scholar.google.com.hk/citations?view_op=top_venues&hl=en&vq=eng

TABLE I
PREDICTION PERFORMANCE COMPARISON AGAINST BASELINES.

Metric	R^2				*NDCG*							
					k=5				k=10			
	TC	YAN	SVR	kINP	TC	YAN	SVR	kINP	TC	YAN	SVR	kINP
$\Delta t = 1$	0.077	0.158	0.115	**0.443**	0.177	0.201	0.245	**0.354**	0.253	0.331	0.279	**0.373**
$\Delta t = 2$	0.208	0.308	0.117	**0.550**	0.200	0.301	0.241	**0.374**	0.261	0.410	0.272	**0.423**
$\Delta t = 3$	0.178	0.269	0.178	**0.811**	0.192	0.242	0.282	**0.321**	0.255	**0.342**	0.299	0.335
$\Delta t = 4$	0.161	0.242	0.142	**0.920**	0.181	0.237	0.274	**0.308**	0.251	0.351	0.308	**0.362**
$\Delta t = 5$	0.206	0.287	0.163	**0.965**	0.201	0.291	0.306	**0.340**	0.231	0.340	0.306	**0.345**

and $t = 4$, while they have opposite effects in $t = 1$ and $t = 5$. The property of three categories features are deserving of being learned in future work.

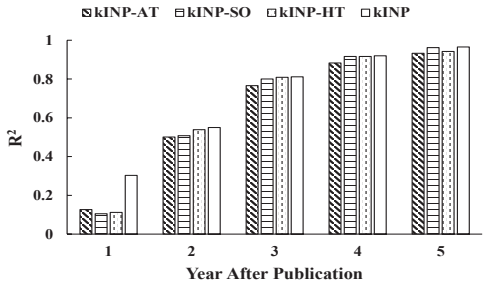

Fig. 2. Performance on Different Features(R^2).

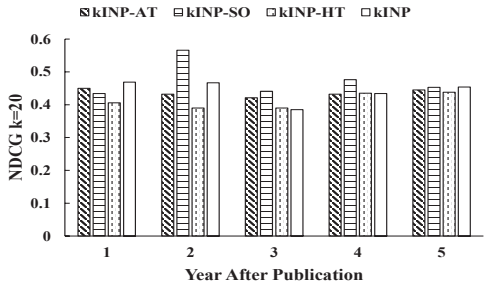

Fig. 3. Performance on Different Features(*NDCG k=20*).

VII. CONCLUSION

In this paper, we introduce the topic related authority to justify how a paper content and user authority influence a new paper's popularity. We adopt the Factorization Machine method to collaboratively learn the latent correlations between factors and top k influential papers are selected. Comparing with traditional methods, it does not require the citation information to evaluate a paper quality, which is appropriate for new published papers.

VIII. ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (91646119),the Key Research and Development Program of Shandong Province (2015GGX106002, 2017GGX10114), Shandong Provincial Natural Science Foundation under Grant No. ZR2014FM014 and SAICT Expert Program.

REFERENCES

[1] T. Bjarnason and I. D. Sigfusdottir. Nordic impact: Article productivity and citation patterns in sixteen nordic sociology departments. *Acta Sociologica*, 45(4):253–267, 2002.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[3] C. Castillo, D. Donato, and A. Gionis. *Estimating Number of Citations Using Author Reputation*, pages 107–117. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[4] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. Towards a stratified learning approach to predict future citation counts. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 351–360, Piscataway, NJ, USA, 2014. IEEE Press.

[5] J. Chen and C. Zhang. Predicting citation counts of papers. In *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on*, pages 434–440. IEEE, 2015.

[6] Y. Dong, R. A. Johnson, and N. V. Chawla. Will this paper increase your h-index?: Scientific impact prediction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 149–158, New York, NY, USA, 2015. ACM.

[7] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[8] T. Lou and J. Tang. Mining structural hole spanners through information diffusion in social networks. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 825–836, New York, NY, USA, 2013. ACM.

[9] S. Rendle. Factorization machines with libFM. *ACM Trans. Intell. Syst. Technol.*, 3(3):57:1–57:22, May 2012.

[10] H. Shen, D. Wang, C. Song, and A.-L. Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 291–297. AAAI Press, 2014.

[11] I. Tahamtan, A. S. Afshar, and K. Ahamdzadeh. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, 107(3):1195–1225, 2016.

[12] S. Xiao, J. Yan, C. Li, B. Jin, X. Wang, X. Yang, S. M. Chu, and H. Zhu. On modeling and predicting individual paper citation count over time. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2676–2682. AAAI Press, 2016.

[13] Y. Xie, Y. Sun, and L. Shen. Predicating paper influence in academic network. In *Computer Supported Cooperative Work in Design (CSCWD), 2016 IEEE 20th International Conference on*, pages 539–544. IEEE, 2016.

[14] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, pages 51–60, New York, NY, USA, 2012. ACM.

[15] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: Learning to estimate future citations for literature. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1247–1252, New York, NY, USA, 2011. ACM.

[16] X. Yu, Q. Gu, M. Zhou, and J. Han. Citation prediction in heterogeneous bibliographic networks. In *SDM*, 2012.

[17] Y. Zhao, G. Wang, P. S. Yu, S. Liu, and S. Zhang. Inferring social roles and statuses in social networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 695–703. ACM, 2013.