Proceedings of the 2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design

Predicating Paper Influence in Academic Network

Yi Xie
Shandong University
Jinan, China 250100
Email: heilongjiangxieyi@163.com

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

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

***Abstract*—It is meaningful to recommend appropriate works to a researcher. One important consideration is the relatedness to one's interests. Although it can be expressed by one's query on an academic dataset, there often exists some semantic ambiguity in relatedness computation that are caused by personalized vocabularies of authors and queriers. Another considered aspect is the quality of a publication, which is often justified by the number and quality of its citations. But it is difficult to estimate the potential influence of a new publication when it has few citation. In this paper, we try to solve the two problems in academic recommendation. To reduce the semantic ambiguity, domain knowledge is created by learning the inherit relativity of word usage from an academic dataset. To compute the potential influence of a new publication,we taking into account the contents and venue of a paper, as well as the reputation of its authors. A recommendation algorithm is designed to find the top k related and influential papers for a query from new publications. We verify the proposed method on real dataset.**

I. Introduction

An academic dataset, such as $DBLP$, is often represented as a dynamic and complex network, where authors, papers or venues are represented as nodes, co-authorships or paper citations are in form of links. Generally a respected work often influences a lot of people and is cited by their publications. It is very helpful for researchers to find a good related work at its first published time so as to improve their work as quick as possible. Generally, people often search such related works by a query on academic network. A search engine takes several keywords as an input and returns the top k papers by calculating the similarity between papers and queries. To recommend the personalized results, some search algorithms adaptively adjust the rankings of papers relatedness by learning from user hitting behaviors. However, since the current methods compute papers relatedness by text comparison, there often exists the semantic ambiguity that are caused by different vocabularies of both authors and queriers. It is desired to find the intrinsic semantics behind an academic query, as well as the semantic relativity between papers. Another considered aspect is the quality of a publication, which is often justified by the number and quality of its citations. But it is difficult to estimate the potential influence of a new publication when it has few citation, which is called as the Paper Influence Prediction Problem in this paper.

A. Related Work

To solve the semantic relativity problem, the topic dynamic model provides a useful means to discover the relatedness between documents from a large collection. Mohamed et al. proposed an LDA-based topic model for analyzing topic-sentiment evolution over time by modeling time jointly with topics and sentiments, and derived inference algorithm based on Gibbs Sampling process[5]. Hu et al.constructed a dynamic LDA model for mining of text topics, and the evolution of dynamic topics of text contents was achieved from the aspects of topic similarity and intensity[6]. Minghui Qiu et al.[7] proposed an LDA-based behavior-topic model (B-LDA) which jointly models user topic interests and behavioral patterns. However, the models are user-dependent. It is impractical to train thousands of models for all individuals in a large digital library. Mehmood et al.[4] proposed a community-level social influence index to analyze information propagation and social influence at the granularity of communities. All of the above methods cannot incorporate the implicit aspect of relatedness between text words in an academic network.

To predict the influence of new publications, some methods try to find hot topics such that the papers covering the topic may be popular. Topic evolution concerning the evolution of inter-community influence, as well as the evolution of influence relationships between communities in dynamic academic networks[3]. Song M et al.[2] proposed three novel technique to conduct topic evolution analysis: Markov Random Field-based Topic Clustering (MRFTC), Automatic Clustering Labeling, and Meta Term Mapping. Liu et al.[9] simulated the constructed features of previously observed topic dynamics with the PreWHether model to predict whether a topic will become prevalent, and further modeled the distributions of time intervals from the emergence of the topic to its prevalence by using the Gamma distribution with the $PreWHen$ model, which predicates when a topic becomes popular. However, a paper is not necessarily influential if it researches a hot topic.

Another related work for predicating the influence of a paper is to compute the authority of authors. Moreira et al.[1] proposed several rank aggregation algorithms in the expert finding task, which located the authoritative authors in academic network. They proposed two frameworks for combining multiple estimators of expertise. These estimators are derived from textual contents, from graph-structure of the citation patterns for the community of experts and from profile information about the experts. It is a reasonable choice to justify the quality of a new publication by its author authority. But even written by the same author, the paper qualities may be different. So it is desired to consider more other

978-1-5090-1915-1/16/$31.00 ©2016 IEEE

issues that related to the quality of a paper, such as the venue of publication which is the basic measurement by the professionals.

In this paper, we solve the above two problems, the potential influence of a new publication and the semantic ambiguity on an academic dataset. To reduce the ambiguity of different publications and queries, we establish domain knowledge with representative keywords, and find the most related domain of the given search words. With regard to the influence, we predict the potential influence of a new publication using author authority. Based on documents that describe authors' activities, paper textual contents, and citation dataset, we search the papers that are influential and related to a given retrieval problem, which is referred to as *Influential Related Paper Prediction*, *IRP* for short. The task takes several keywords as a search requirement, and returns a list of papers which are sorted by their level of authoritative in what concerns the query topic. Better than the existing search engines, that only return the most relevant answers, our system returns the most potentially influential and relevant papers.

The paper is organized as follows. Section II presents the formal definition of the problem. In section III, we present our approach to compute the relatedness to fine grained domain and papers for an academic query. Section IV introduces our model for computing the potential influence of a new published paper. In Section V, we present and discuss the experimental results on influence prediction and the performance of our model. We conclude the paper in Section VI.

II. Problem Statement and Framework

A *Dynamic Academic Network*, *DAN* for short, is a sequence of snapshots on academic network at several time points. For example, the DBLP dataset contains the academic papers on computer science. In a DAN, nodes are authors and papers, links are co-authorships and citation relations. A snapshot of a dynamic academic network at time t is denoted by $G^t = (V, E, M, B)$, where V is the node set and E is the edge set, M and B denote the attributes of V and E, $t \in [1..T]$, where T is the maximum time window on *DAN*. We further partition the node set into two subsets, namely $V = A \cup P$, where A and P represent author set and paper set, respectively.

An academic query is used to express the search interests on some academic domain, which is used to compute the relatedness between papers and interests. For example, "machine learning" is a search word within an academic query. For a given query on an academic dataset, our purpose is to find the most related and potentially influential new papers that has few citation.

Given a DAN G^t, $t \in [1..T]$, a positive number $k \in N^+$ and a set of keywords as an academic query $Q = \{w_1, w_2, \cdots, w_l\}, l \in N^+$, $w_i, i \in [1..N]$ is a word, the Influential Related Papers Prediction Problem is to find k papers for Q with the most influence on the related domain at $t+1$ from G^t.

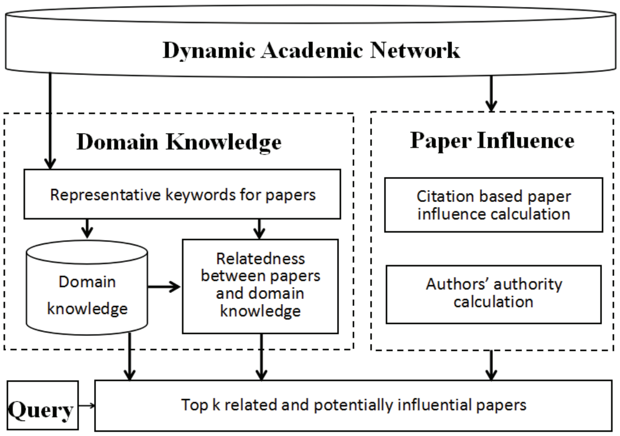

Dynamic Academic Network
Domain Knowledge
Representative keywords for papers
Domain knowledge
Relatedness between papers and domain knowledge
Paper Influence
Citation based paper influence calculation
Authors' authority calculation
Query
Top k related and potentially influential papers

Fig. 1. Framework on Influential Related Papers Prediction

The solution architecture is shown in Fig.1, where there are two key points: the relatedness between query Q and domain and the influence of a new publication.

It is a natural choice to compute the relatedness between query Q and domain by the domain knowledge categories (like ACM Categories), but the method only considers the explicit connections. Also categories are only a coarse classification, which can not describe the more detailed research directions, especially for dynamically generated research points. So, we introduce fine grained semantic extraction method (SEM for short) to create the domain knowledge from academic dataset, which consider both explicit and implicit relatedness. This process is shown as the part $Domain\ Knowledge$ in Fig. 1. The representative keywords of a paper are extracted from the paper textual contents. From the whole scientific dataset view, each keyword is represented by the statistic vector for all papers and their relations can be got by computing the distance of two keywords. Then the fine grained research domains are classified by cluttering these key words, which are called $Domain\ Knowledge$ (**DK** for short). Based on **DK**, the relatedness between papers and domains can be computed. For any query Q, the related domains and most related papers can be calculated also.

Considering the second point, we predict the potential influence of new publications by authors' authority and more other issues that related to the quality of a paper, such as the venue of publication. Since doing research is continuous process, only after commutative investigation one can have some achievements. So the reputation of an author can be computed from one's historical scientific activities, which can be used to predicate a new publication influence. which are shown as the part $Paper\ Influence$ in Fig. 1. In the following sections, we would present the details.

III. Paper Relatedness

There are two kinds of relatedness: explicit and implicit (or hidden) relatedness. If a paper contains a query keyword, this paper is called explicitly related to that keyword. But if another paper does not contain a given query keyword, it does not mean it is not related to that keyword. Maybe actually this paper is highly related to the query. Since in practice, people often have their own habits in writing papers, such as frequently used parses or presentations. Similarly, people are used to their own vocabulary different words for the same search requirement. For example, "touch screens" and "haptic devices" are totally different phrases that are relevant to the same small academic domain, which refers to "Tactile and Hand-based Interfaces".

We try to reveal the semantics from academic dataset and represent the implicit relatedness as domain knowledge. To solve this problem, we calculate related domains of a certain query, rather than calculate related papers directly. Firstly, we need to find the hidden closeness between words, which actually exist in the content of each paper . For example, a paper with keyword "touch screens" always contains "interface design" and "virtual keyboards" etc keywords. So, the concept of domain knowledge is introduced to represent the intrinsic semantics of the related words in a small specific domain. Then, for each paper, the relatedness to the domain knowledge is calculated so that the semantic ambiguity can be reduced. Similarly, the relatedness of a query to papers is computed against the domain knowledge.

A. Domain Knowledge

To select the representative words of a paper, we extract key words from the abstract of a paper instead of using its Key Words directly. The Key Words of a paper are sematically limited, while an abstract contains more semantic information. We adopt the TF-IDF approach, which is widely adopted in information retrieval to quantify the importance of a word in a text [10]. For the efficient computation purpose, a representative word set can be chosen instead of the whole set. For paper p_i, the set of selected keywords are represented as K_i. The TF-IDF score of keyword w_i in paper p_j is calculated as follows:

$$x_{i,j} = \begin{cases} 1 + \log f_{i,j} \times \log \frac{|P|}{n_i}, & f_{i,j} > 0 \\ 0, & otherwise \end{cases} \tag{1}$$

where $f_{i,j}$ is the frequency of keyword k_i in paper p_j, $n_i = |\{p_j | k_i \in K_j\}|$.

For the academic network, the set of keywords are defined the union of K_i, denoted by $K = \bigcup_{p_i \in P} K_i$. For each keyword w_i, the vector x_i is created as the TF-IDF scores against all papers in P in a given DAN. Namely, $x_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,|P|})$. To find the fine grained research points, we need to cluster the representative keywords into groups, which based on the closeness between words. There are many candidate functions to compute the distance between two vectors.For example, the Manhattan Distance[1] is an efficient choice. For two words w_i and w_j, their distance $dis(i,j)$ is:

$$dis(i,j) = \|x_i - x_j\|_1 = \sum_{k=1}^{|P|} |x_{i,k} - x_{j,k}| \tag{2}$$

Then we cluster keywords into domains. To provide users a flexible choice on the retrieval grain, the clusters should be hierarchically organized so that a lower level reflects a closer relatedness. Also such clustering algorithms should be able to detect overlapped clusters, and to reflect how much a keyword belongs to each domain. In a hierarchical domain distribution, different level maps to different extent. Such requirements are also consistent with the traditional academic classification.

Both hierarchical clustering and *CESNA* [14] clustering algorithms satisfy the above properties. Besides, they are efficient in computations. In this paper, we adopt hierarchical clustering algorithm to cluster keywords into domains.

After clustering, each cluster maps to a specific domain Θ_m and is represented by a set of keywords associated with belonging weights. Formally, $\Theta_m = \{(w_i, \xi_i) | w_i \ is \ a \ keyword, \xi_i \in R^+\}$, where ξ_i is a weight which describes the loyalty of keyword w_i to DK_m.

B. Domain Related Papers

In order to compute which domains a paper belongs to, we calculate the similarity between domains and papers. There are quite a few functions can be used to compute the similarity of two vectors. Such functions should be capable of managing different standardization degrees between two vectors. In this paper, we select Spearman correlation coefficient[2], which is a nonparametric measure of statistical dependence between two variables.

Having a domain vector Θ_m, we compute the rank of each keyword according to the loyalty. For each paper, the vector $\Lambda_n = \{(w_i, tfidf_i) | w_i \ is \ a \ keyword, \ tfidf_i \in \{x_{i,n}, 0\}\}$ is created against K, where each score for the corresponding word is the TF-IDF score in paper and other elements not in K_i are set to 0. Obviously, it is different from the vector x_i.

For two vectors Θ_m and Λ_n for DK_m and p_n, respectively, we convert them to ranks θ_m and λ_n. The similarity $sim(m,n)$ between the ranked variables is calculated as:

$$sim(\Theta_m, \Lambda_n) = 1 - \frac{6 \sum_{i=1}^{|K|} d_i^2}{(|K|)[(|K|)^2 - 1]} \tag{3}$$

where $d_i = \theta_{m,i} - \lambda_{n,i}$ is the difference between ranks. A paper may related to more than one domains in different degrees. For example, a paper related to Data Mining domain by eighty percent, related to Security domain by twenty percent.

C. Query Related Papers

To recommend the most related papers of a query Q, we need to calculate the similarity between a query and academic domains. Having the domain knowledge **DK** and a given query Q, we firstly finds the l most related domains. The relatedness between Q and **DK** is computed against Equation 2, denoted by ξ, based on which we have the top $l \in N^+$ related domains $D_Q^l = \{(DK_1, \xi_1), (DK_2, \xi_2), \cdots, (DK_l, \xi_l)\}$. Then we can create an $l \times |P|$ sized matrix P_D, in which the rows are

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

[2] https://en.wikipedia.org/wiki/Spearman' s_rank_correlation_coefficient

DKs and the columns corresponding to papers. The element $P_D(i,j)$ is the loyalty of paper p_j to DK_i. QP denotes the relatedness between Q and papers in different domains, which is calculated as $(D_Q^l \times P_D)$. The relatedness between a specific paper p_m and Q is calculated as:

$$rel_Q(m) = \sum_{i=1}^{l} QP(i,m) \tag{4}$$

IV. Paper Influence

To compute the potential influence of a new publication, we taking into account the authority of the author and the venue information. The historical academic activities of authors are used to compute authorities. We take time into account so as to express the time decay of paper influence prediction. We adopt *PageRank* method to judge the influence of papers with citations, and *Contemporary Hirsch index* to predict the influence of new publications.

A. Justification of Influential Papers

There are two features to judge a paper influential with citation dataset: a paper is cited by many papers, or it is cited by influential papers. Apparently, papers with a lot of citations are more influential than those with few citations. With regard to papers with same citation counts, the papers cited by influential ones are more likely to be influential than those cited by regular papers.

Citation counts give no weighting towards papers of greater influence. Naturally, definition of such influence is a subjective task. Therefore, we propose a *PageRank* method of objectively weighting paper influence. A PageRank results from a mathematical algorithm based on the papergraph, created by all academic papers as nodes and citations as edges. The rank value indicates an influence of a particular paper. A citation to a paper counts as a vote of support. The PageRank of a paper is defined recursively and depends on the number and PageRank metric of all papers that cite to it (“incoming links”). A paper that is cited by many papers with high PageRank receives a high rank itself. This method is originally used for link analysis by search engines, and it proposes that a web page itself carries a greater importance if linked to by other high importance pages. We adopt the PageRank algorithm to calculate the influential of academic papers.

Given a DAN G^t, $t \in [1..T]$, $p_i \in P$ is an academic paper, the PageRank values of paper p_i is calculated as:

$$PR(p_i) = \frac{1-d}{|P|} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{5}$$

where d is the dampening factor ($d = 0.85$ here), $M(p_i)$ is the set of all inbound citations to p_i, $L(p_j)$ is the number of outbound citations of p_j.

B. Predication of Paper Influence

In this section, we predict the potential influence of new publications, that have few or no citations to show their influence directly. Papers are more likely to be influential that written by authoritative authors. The ultimate aim of predicting paper influence is to obtain an orderly sequence instead of getting accurate citation counts. Therefore we translate the *Paper Influence Prediction* problem into an *Author Authority Judgement* problem.

There are three features to estimate the authority of authors based on citation information, respectively, PaperRank scores, academic indexes, and graph centrality[1]. With regard to the first feature, the papers’ PaperRank scores of a specific author are calculated mean value, which is used to predict her new publication’s PaperRank.

With regard to academic impact indexes, the *Hirsch index*[3] is widely adopted by the academia to measure the scientific productivity and the scientific impact of an author, denoted by $\hbar$. For a given author who has published N_p papers, her *Hirsch index* is $\hbar \in N^+$ if there are at least $\hbar$ of her papers that each one is cited by at least $\hbar$ papers. Authors who have a high $Hirsch\ index$ are more likely to be considered experts. Adding a temporary weight to each cited article, we give less weight to older articles. We introduce time element into the quantification of $Hirsch\ index$ and name it by *Contemporary Hirsch index* $\hbar^c$. An author has a contemporary Hirsch index $\hbar^c$ if $\hbar^c$ of his N_p articles have a score of $S^c(i) \geq \hbar^c$ each, and the remaining $(N_p - \hbar^c)$ articles have a score of $S^c(i) \leq \hbar^c$. For a paper p_i, the score $S^c(i)$ is

$$S^c(i) = \gamma * (T - t(i) + 1)^{-\delta} * |CitationsTo(i)| \tag{6}$$

In this formula, T is the year we make predictions, $t(i)$ refers to the year of publication for paper i, $t(i) < T$, $CitationsTo(i)$ is the citations of paper p_i. The γ and δ parameters were set to 4 and 1, respectively, which means that the citations for a paper that was published during the current year are accounted for as four times, the citations for an article that was published 4 years ago are accounted for as only one time and the citations for a paper that was published 6 years ago are accounted for as four/six times, and so on.

With regard to graph centrality features, we adopt *AuthorRank* method to estimate the authority of authors. Given a graph with authors as nodes, they are connected through citation links. The AuthorRank of an author $a_i \in A$, $AR(a_i)$ is defined in *Equation* 7.

$$AR(a_i) = \frac{(1-d)}{|A|} + d \sum_{a_j \in inLinks_i} \frac{AR(a_j)}{outLinks_j} \tag{7}$$

In *Equation* 7, the sum is over all authors a_j that cite author a_i, denoted by $inLinks_i$. The term $outLinks_j$ corresponds to all citations made by an author a_j, and the term $1-d$ corresponds to the dumping factor, which can be seen as a decay factor. Under the expert finding scenario, the dumping factor can be seen as an interest over a different author, instead of the search process only being interested in authors that are cited. Most of the implementations in the literature set the parameter d to 0.85.

[3] https://en.wikipedia.org/wiki/H-index

In comparison, judging authors' authority by Contemporary Hirsch index in Equation 6 is more credible. An author is not necessarily authoritative if she has published influential papers many years ago, while she is very likely to be authoritative if she has published influential papers in recent years.

V. Recommendation Algorithm and Experiment

This section describes the validation of the main hypothesis behind this work, which states that either learning to rank approaches or learning to rank integration methods can combine domain relatedness and paper potential influence with respect to queries in a principled way, in this way improving over the current paper retrieval system.

A. Integration and Metric

To validate the relatedness calculating approaches that have been proposed in this work, we required a set of queries that have the corresponding paper's relevance judgements. We proposed a set of 13 query topics from the Computer Science domain. Table I(a) shows the nearest keywords that are associated with each topic, which are calculated by Equation 2. Table I(b) shows the indexes of the most related papers that are associated with each domain which is related to a query.

To validate the paper influence prediction algorithms, we used two different performance metrics, namely, the Precision at k (Ps). The Precision at rank k is used when a user wishes to look only at the first k retrieved domain papers. The precision is calculated at that rank position as:

$$Ps = \frac{|CH(k) \cup GT(k)|}{|GT(k)|} \tag{8}$$

where $CH(k)$ is the set of top k potentially influential papers from our method, $GT(k)$ is the set of top k papers that have the most citations according to ground truth.

B. Experimental Setup and Baseline Method

In this work, we use a dataset for evaluating paper searches in the Computer Science domain, which corresponds to an enriched version of the DBLP[4] database that was made available through the Arnetminer project. This dataset contains the papers published in conferences from 1980 to 2010, totally counting 1397240 papers. The normal information of papers contains index, title, author, published year, conference and, some papers, citation and abstract.

There are 818457 papers contain the abstract and the total number of citations is 3011489. So pre-processing and cleaning of the data is necessary. In order to make full use of the citations, we kept the papers which were lacked of the information of citation or abstract but were cited by other papers.

The existing algorithms in estimating the importance of the contributions of specific publications can be used to handle the paper influence prediction problem discussed in this paper. As our proposed approach is based on time series analysis, we compare our method with the $a\text{-}index$, which measures the magnitude of the most influential papers. For an author that has an a-index of $a = N_{c,tot}/\hbar^2$, when he has a Hirsch index of $\hbar$ and a total of $N_{c,tot}$ citations towards his papers.

TABLE I
Some example of query words

(a) Query words and 4 nearest hot keywords

QUERY WORDS	4 NEAREST NEIGHBORS
parallel + cloud	process, application, web, software
bridge + switch	control, web, space, security
packet + router	simulation, control, quality, web
power + energy	efficient, parallel, linear, dynamic
memory + storage	space, environment, mobile, power

(b) Related papers of query words

QUERY WORDS	Rel. PAPERS
parallel + cloud	12837 16346 17155, 20322,···
bridge + switch	17515, 3024, 20686, 21841,···
packet + router	3943, 9483, 10694, 13797,···
power + energy	12880, 10694, 18136, 8027,···
memory + storage	16246, 17115, 23003, 23015,···

C. Experimental Results

This section presents the results that were obtained in influential prediction that were tested in this work. We extract the data from 1980 to 2010, which is divided into 6 time windows and the length of every time window is 6 years. Papers of 5 years are treated as the training set to predict the influence of papers that published in the sixth year. First, we use the data set of 1980 to 1984 as the training set to predict the influence of papers published in 1985. We adopt the data set of 1985 to 1989 is used as the training set to predict influence of papers in 1990 and compute the precision between the predicted value and true value. The paper influence of 1990 to 2010 are similar to the above-mentioned settings.

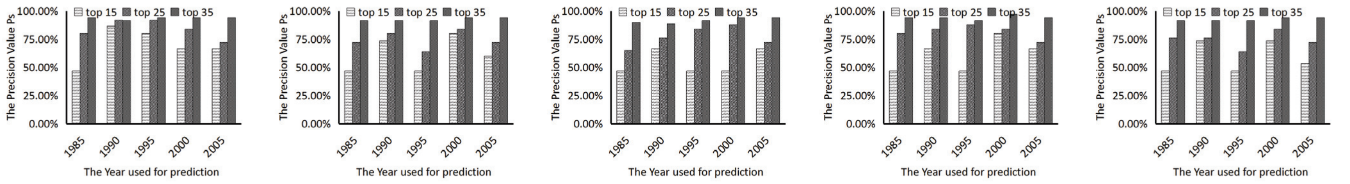
Take the query "bridge and switch" as an example, we choose top 40 domain related papers that published in every five years. Then we use the citation dataset in earlier five years to rank the influence of those 40 papers, and evaluate the performance of our method against the ground truth by Equation 8. The experiment results are shown in Figure 2(a), in which the curve marked by different colors means different k values that set to rank the top k influential papers.

As shown in Figure 2, with the increase in number of year from 1980 to 2005, the value of P_s subsequently increase under different k settings. These observed results demonstrate that the larger size of observed data will lead to better prediction results. Figure 3 shows the prediction performance of different approaches on the top 25 influential papers. We can see that our model consistently achieves better performance than the baseline method. However, the training data cannot be very large, because too many latent features will cause the over-fitting problem which will do harm to the prediction accuracy.

VI. Conclusion

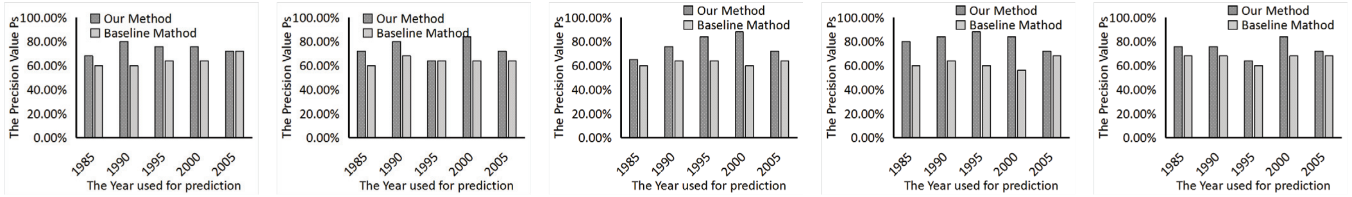
In this paper, we solved this problem from both the relatedness and influence for a new published paper. To make

[4] http://www.arnetminer.org/citation



(a) Bridge and Switch (b) Parallel and Cloud (c) Packet and Router (d) Power and Energy (e) Memory and Storage

Fig. 2. The Performance of Influence Prediction of Top 40 Related Papers published.



(a) Bridge and Switch (b) Parallel and Cloud (c) Packet and Router (d) Power and Energy (e) Memory and Storage

Fig. 3. The Performance Evaluation for Our method and Baseline Method.

a recommendation more meaningful, domain knowledge is learned from academic dataset and each paper is computed its relatedness with different domains. We computed the potential influence of a new paper, by taking into account the contents of paper, the authors' previous publications and activities. An academic query is desired to express the academic interests on different domain, which is used to compute the relatedness between papers and interests. The top k related and potentially influential papers are recommended. In the near future, it would be practical to extend our model with other implicit information such as social network information, expecting for better prediction performance.

VII. Acknowledgements

This research is supported by the National Natural Science Foundation of China (61173140), Special Program on Independent Innovation & Achievements Transformation of Shandong Province (2014ZZCX03301) and Science & Technology Development Program of Shandong Province (2014GGX101046).

References

[1] Moreira C, Calado P, Martins B. Learning to rank academic experts in the DBLP dataset[J]. Expert Systems, 2013.

[2] Song M, Heo G E, Kim S Y. Analyzing topic evolution in bioinformatics: investigation of dynamics of the field with conference data in DBLP[J]. Scientometrics, 2014, 101(1): 397-428.

[3] Chikhaoui B, Chiazzaro M, Wang S. A New Granger Causal Model for Influence Evolution in Dynamic Social Networks: The Case of DBLP[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[4] Mehmood Y, Barbieri N, Bonchi F, et al. Csi: Community-level social influence analysis[M]//Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2013: 48-63.

[5] Mohamed Dermouche, Julien Velcin, Leila Khouas, and Sabine Loudcher.A Joint Model for Topic-Sentiment Evolution over Time.In ICDM, pages 773-778,2014.

[6] Hu Jiming,Chen Guo. Mining and Eolution of Content Topics Based on Dynamic LDA [J]. Library and Information Service, 2014,58(2): 138-142. (in Chinese)

[7] Minghui Qiu,Feida Zhu,Jing Jiang .It Is Not Just What We Say, But How We Say Them: LDA-based Behavior-Topic Model. In SIAM. 2015. Page 794-802

[8] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the Web[J]. 1999.

[9] Deng Z H, Gong X, Jiang F, et al. Effectively Predicting Whether and When a Topic Will Become Prevalent in a Social Network[J]. 2015.

[10] Plansangket S, Gan J Q. A query suggestion method combining TF-IDF and Jaccard Coefficient for interactive web search[J]. Artificial Intelligence Research, 2015, 4(2): p119.

[11] Mohammadi E, Thelwall M. Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows[J]. Journal of the Association for Information Science and Technology, 2014, 65(8): 1627-1638.

[12] Ray Choudhury S, Giles C L. An Architecture for Information Extraction from Figures in Digital Libraries[C]//Proceedings of the 24th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2015: 667-672.

[13] Priem J, Piwowar H A, Hemminger B M. Altmetrics in the wild: Using social media to explore scholarly impact[J]. arXiv preprint arXiv:1203.4745, 2012.

[14] Yang J, McAuley J, Leskovec J. Community detection in networks with node attributes[C]//Data Mining (ICDM), 2013 IEEE 13th international conference on. IEEE, 2013: 1151-1156.