Citation Networks as a Multi-layer Graph: Link Prediction and Importance Ranking

CS224W Project Report, Group 5

Jingyu Cui
Electrical Engineering
Stanford CA 94305
jycui@stanford.edu

Fan Wang
Electrical Engineering
Stanford CA 94305
fanw@stanford.edu

Jinjian Zhai
Computer Science
Stanford CA 94305
jameszjj@stanford.edu

ABSTRACT

In academia, to represent the relationships between researchers or papers, the traditional way is to analyze the citation network constructed by paper citation relationships. Our major contribution is that we incorporated multiple networks from the publication dataset, such as the paper citation network, author citation network, author collaboration network, etc. These networks can provide very useful and interesting information for analyzing the citation behavior of researchers. In this paper, we first analyze how people are citing papers by applying logistic regression model to paper citation network. With the additional information from the author networks, we can get improved results of link prediction in citation network. Then we run a modified PageRank algorithm on paper citation network and author citation network, which can provide us with a more accurate sense of the paper and researcher impact.

Keywords

Citation network, Regression, Link prediction, PageRank, Impact metric.

1. INTRODUCTION

When researchers are writing a paper, usually they will cite other publications as references. As digital libraries become more and more popular recently, almost all the research publications have been made available to the general public. Although there might be a lot of related works available for citation, researchers usually don't just randomly pick some of them; it is believed that they are following some specific rules. It might be interesting and meaningful for us to investigate the citation network and try to find these rules. Our goal of this project is to analyze the static and dynamic properties of citation network, to further obtain some insights about the measurements of research quality, collaboration behavior, even the evolution of science and technology.

Citation network analysis has been a hot research topic, and there have been many publications in the related field. In [15], the authors investigated how patterns of citations varied between the scientific disciplines and how such patterns related to the impact of a paper. A citation projection graph was defined, and several metrics and statistics were proposed to capture the network property. In [13], although the network analyzed was a Blog network, the proposed Susceptible-Infected-Susceptible model was very useful for us to analyze citation network. In [12], how to find the most important nodes to obtain information from the network was discussed as an optimization problem.

The authors of [8] proposed important indices on arcs to identify the important parts of the citation network. The authors of [1] proposed an efficient algorithm for determining the arc weights in the SPLC and SPNP weights. The authors of [9] tried to analyze a citation network constrained in a specific attainability science domain. The authors of [10] proposed three methods to analyze the large scale dynamic network using EM algorithm, modularity optimization, and eigenvector centrality.

Inspired by the publications above, we are trying to analyze the citation network in a more accurate and complete manner, and we state the problems that we want to investigate in the following sessions.

The rest of the paper is organized as follows: Section 2 describes some basic analysis of the citation network characteristics. Section 3 depicts our methods of modeling and predicting citation behavior in the network. Section 4 gives the method and results for predicting impact of items in the network.

2. BASIC ANALYSIS OF CITATION NETWORK

We consider two datasets in our following analysis and experiments. The first one is DBLP dataset, which includes the information on computer science publications listed in the DBLP Computer Science Bibliography (`www.informatik.uni-trier.de/~ley/db/`). The second one is Hep-Th dataset,

Each entry of the dataset will contain a piece of publication record, which includes publication id, title, author list, citations, proceeding, year, pages, etc. First we construct the traditional paper citation network. Note that in DBLP dataset, all of the publications have not been cited, nor

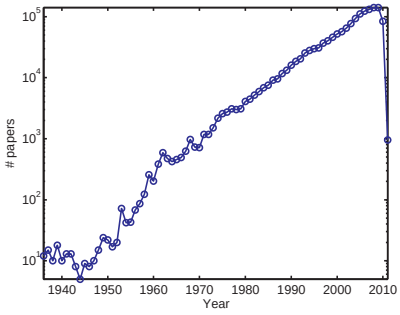

Figure 1: The distribution of paper number in each year for DBLP data set.

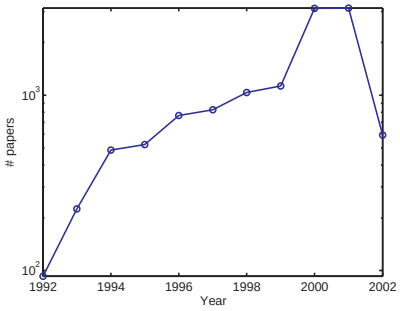

Figure 2: The distribution of paper number in each year for the HEP-TH data set.

all publications have cited other publications in the same dataset. There are in total 1,466,034 papers in DBLP data set, while only 22,138 papers which are either citing other papers or being cited by others. That is to say, the DBLP data set doesn't contain the full citation information; however, we can regard it as sufficient to reflect the true citation pattern. The reason here is that the characteristics and the citation pattern we found for this network and the Hep-Th network are similar to each other, which will be shown in the following parts.).

2.1 Data Analysis

In this part, we analyzed the characteristics of the traditional paper citation network. Several interesting points have been discovered here, as described in the following sections.

2.1.1 More papers are produced as years go by

We first plot the distribution of the number of papers published in each year as Fig.1 and Fig.2. The x-axis is the year, and the y-axis is the log scale of the number of papers published in the corresponding year. The log-scale paper numbers almost increase linearly as time goes by (the point corresponding to the last year can be ignored since the data might be incomplete), which means, the actual number of papers is increasing exponentially! Each field is a really booming field, and this could be also because publishing papers is becoming easier as there are more journals and conferences.

2.1.2 The degree distribution follows power law

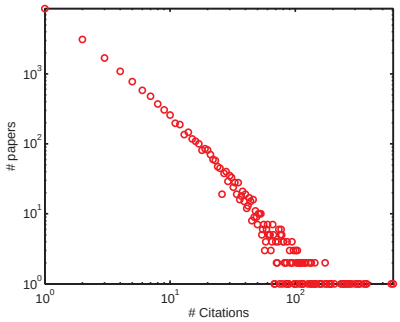

Figure 3: The in-degree distribution of DBLP data set.

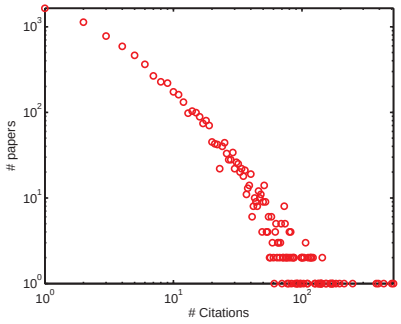

Figure 4: The in-degree distribution of the HEP-TH data set.

We construct a citation graph from each dataset, which is a directed graph. For a paper which is cited by another paper, it has an incoming edge; for a paper which is citing another paper, there is an outcoming edge.

The in-degree distribution and out-degree distribution are plotted in Fig.3,Fig.4, FIg.5, Fig.6 respectively. From the figures we can see that, in- and out- degrees of both data sets follow power law distribution.

3. CITATION MODEL AND LINK PREDICTION

To model a network, the most important thing is to determine how the graph evolves and how to determine there is

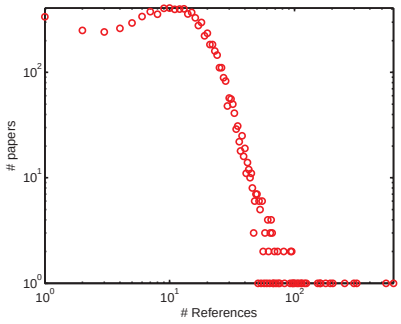

Figure 5: The out-degree distribution of DBLP data set.

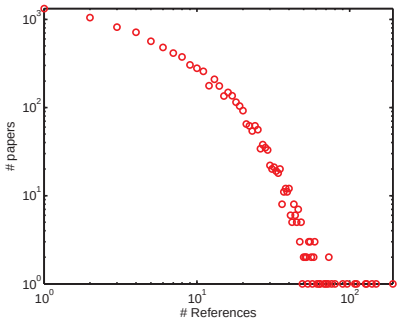

Figure 6: The out-degree distribution of the HEP-TH data set.

link between two nodes. Here we want to decide the probability of connecting two nodes through a function of many factors. These factors are those we think might have influence on an author's decision of citation. If we can model this function accurately, we may explicitly express the rule of citation, and explicitly state some properties of the citation network.

3.1 Citation Model

In many pervious literatures, how the information is propagated or how a link will be created in the network was simulated by a simple but powerful probabilistic model, in which each node spreads its information to (or link to) its neighborhood nodes with probability β. Thus the choice of β is very critical. However, it is somewhat limiting to have one value of β to describe the behavior of all nodes. In our problem, when researchers are thinking about citing other publications, they are not making decisions based on a simple and fixed parameter. Many factors will have important influence in this process, such as how famous the target paper is, whether or not the target paper was the research's own previous work, whether the cited paper is recent work or more than 10 years ago, etc. Therefore, it is necessary and meaningful to adapt the probability value β according to the properties of the edge and the two nodes the edge connects, i.e., the probability that a link is made depends on the properties of the source node and the destination node and the properties of these two together.

Generally speaking, we try to model the relationship as a linear regression model [2]:

$$z(s,d) = \sum_{i=1}^{M} \alpha_i f_i(s) + \sum_{j=1}^{N} \gamma_j f_j(d) + \sum_{k=1}^{L} \xi_k f_k(s,d)$$

where $f_i(s)$ are M properties of the potential source node s, $f_j(d)$ are N properties of the potential destination node d, and $f_k(s,d)$ are L properties decided by the two nodes jointly, i.e. the properties of the edge. α_i, γ_j, and ξ_k are weights to be determined by the regression algorithm. This part is basically a linear combination of all features.

To get a probability parameter β, we transform the real-valued $z(s,d)$ to the $(0,1)$ interval using the logistic function:

$$\beta(s,d) = \frac{1}{1+e^{-z(s,d)}}$$

Note that these regressors contain both the properties decided by single node ($f_i(s)$ and $f_j(d)$), and the properties decided by two nodes of an edge ($f_k(s,d)$).

Specifically, given two papers in our experiment, we first determine the direction of the potential edge. The edge direction can be simply determined by the publishing year of the two papers due to the natural causality of citation. The paper published recently is called the source paper A, and the paper published earlier is regarded as a potential destination B for citation edge.

1. In-degree of B, meaning how many citations B has received;
2. Out-degree of B, meaning how many papers B has cited;
3. Year difference, i.e. The difference between publication year of A and B;
4. Author overlap: The number of overlapping authors of the author lists of A and B;
5. Title similarity: the semantic similarity between two paper titles defined by "WordNet";

3.2 Multi-Layer Graph

Although we are analyzing the citation model, it is the authors who make the decision of citations. Therefore, we think it is also useful to incorporate the authors' properties in our citation model.

Given the publication dataset, two more networks could be constructed besides the paper citation network:

1. Author citation network: if an author X's papers have been cited by another author Y for k times in total, there is a directed edge from Y to X with weight k.
2. Author collaboration network: if an author X has co-authored k papers with author Y, there is an undirected edge between X and Y with weight k.

The three networks together constitute a multi-layer hypergraph structure as shown in Fig.7. These two newly constructed networks are related to the paper citation network; however, we believe that they could provide extra information about the authors, who are the creators of papers and their personal characteristics might affect the citations.

Therefore, several new regressors could be added to our regression model:

1. "MaxACollaberateWithB": the maximum number of collaborations happened between each author of paper A and each author of paper B;
2. "MeanACollaberateWithB": the average number of collaboration between authors of paper A and authors of paper B;

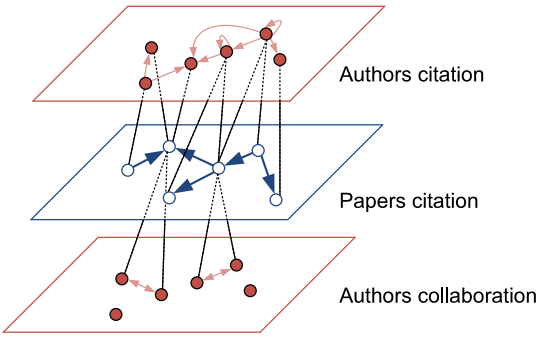

Figure 7: Multi-layer structure of the network containing paper citation network, author citation network, and author collaboration network.

3. "MaxACiteB": the maximum number of citations each author of paper B have received from each author of paper A;

4. "MeanACiteB": the average number of citations the authors of paper B have received from authors of paper A;

5. "AverageCitedNumberB": the average citations the authors of B has received, i.e. the average weight of B's in-edge in author citation network;

6. "AverageReferenceNumberB": the average citations the authors of A has made, i.e. the average weight of B's out-edge in author citation network;

7. "AverageCitedNumberA": the average citations the authors of paper A has received;

8. "AverageCitedNumberA": the average citations the authors of paper A has made;

3.3 Experimental Results

3.3.1 Experiments Setup

Each sample in our experiments represents an edge, and the value of features (regressors) extracted for this edge is calculated based on the network constructed at the time when the source paper is published. That is to say, we cannot use future information when finding citations for current paper.

We generate two training sample sets from DBLP dataset, one is called "DBLP '90-'95", in which the source papers of all edges were published between the year 1990 and 1995; the other is "DBLP '95-'00", meaning the edges are started from papers from 1995 to 2000. The Hep-Th dataset is used as a whole.

For each source paper A, all the actual out-edges can be used as positive training samples, and we random select n non-existing edges starting from A as negative samples, in which n is the out-degree of A. Therefore, we have the same number of positive and negative training samples.

The regression problem can then be solved using the standard Newton-Raphson algorithm [2] to maximize the log-likelihood given the observations.

3.3.2 Performance Evaluation

After fitting the model, given two papers, we can calculate the probability that the later one cites the earlier one, i.e. there is a directed edge between them. Since we know the true edges in the network as the ground truth, we can evaluate prediction precision, as follows:

$$p = \frac{\#\text{correctly predicted edges}}{\#\text{predicted links}}$$

which means, if our model predicts there is an edge, how likely it is a true edge. We can also have the true edge recall as:

$$r = \frac{\#\text{correct predictions}}{\#\text{all true edges}}$$

which means, in all the true edges, how many of them have been found by our model.

We can first see the performance when only using the regressors from the paper citation network like described in Sec.3.1. The results are shown in Fig.8.

After the 8 regressors described in Sec.3.2 are included, we have improved performance shown in Fig.9.

In these two tables, each column means the model is trained with the sample set indicated at the top of the column, each row means the models are applied onto the sample sets indicated at the left end of the row. The upper part of each table is precision and the bottom part is recall.

If we look at the four experiments about the models trained on samples sets "DBLP '90-'95" and "DBLP '95-'00" and applied on these two sets, the performance has been greatly improved by using all regressors from multi-graph compared with only using the regressors on paper citation network.

When a model was trained with one sample set and applied onto the other sample set, if we still get good results, the model should have good generalization ability, which can be shown by the values in the off-diagonal entries.

To visualize the performance better, we have the F-1 score (harmonic mean of precision and recall) plotted in Fig.10.

The small difference between one model applied onto different sample sets shows the generalization ability of our model.

We have also included the results trained by Decision Tree algorithm, in the right half of Fig.9 and Fig.10. Generally speaking, decision tree has a better performance than regression model. However, the generalization ability is a little weaker, and the tree structure is too complicated to be interpreted. Therefore, we will only try to interpret the meaning of regression model in next section.

3.3.3 Citation Model Interpretation

We obtained 3 regression models trained on three different sample sets and those models are compared in Fig.11.

Those weights have been normalized according the scale of each feature, so they can reflect the actual importance.

	Logistic Regression		Decision Tree		
	DBLP '90-'95 Model	DBLP '96-'00 Model	DBLP '90-'95 Model	DBLP '96-'00 Model	
DBLP '90-'95 Data	73.90	69.62	84.42	71.22	Precision
DBLP '96-'00 Data	75.91	72.58	72.07	84.11	
DBLP '90-'95 Data	64.62	71.51	73.88	62.48	Recall
DBLP '96-'00 Data	57.52	64.20	54.14	70.88	

Figure 8: Precision and recall for the regression model with regressors only from paper network.

	Logistic Regression			Decision Tree			
	DBLP '90-'95 Model	DBLP '96-'00 Model	Hep-Th Model	DBLP '90-'95 Model	DBLP '96-'00 Model	Hep-Th Model	
DBLP '90-'95 Data	91.47	90.07		98.63	94.80		Precision
DBLP '96-'00 Data	92.84	91.89		96.15	98.79		
Hep-Th Data		80.95	91.13		97.41	99.45	
DBLP '90-'95 Data	73.01	80.10		98.61	96.00		Recall
DBLP '96-'00 Data	66.64	74.03		91.22	98.86		
Hep-Th Data		90.59	76.07		97.44	99.53	

Figure 9: Precision and recall on different training sets and testing sets.

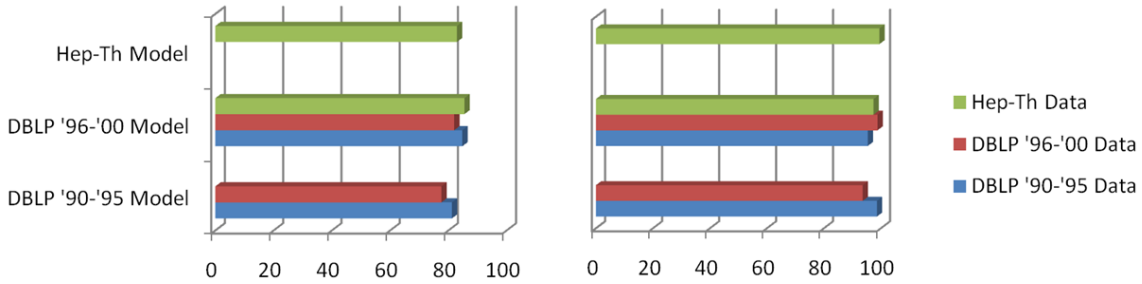

Figure 10: F1 score for the regression model and decision tree model on different training sets and testing sets.

Regressor	DBLP '90-'95	DBLP '95-'00	Hep-Th
In degree of B	0.12	0.04	0.48
Out degree of B	0.18	0.21	0.11
Year difference	0.09	0.11	0.16
Author overlap	0.00	0.00	0.00
Title similarity	0.39	0.39	0.37
MaxACollabrateWithB	0.03	0.02	0.10
MeanACollabrateWithB	0.05	0.03	0.12
MaxACiteB	0.17	0.16	0.26
MeanACiteB	1.05	1.21	1.31
AverageCitedNumberB	0.00	0.02	0.34
AverageReferenceNumberB	0.00	0.03	0.24
AverageCitedNumberA	0.02	0.02	0.02
AverageReferenceNumberA	0.19	0.13	0.27

Figure 11: Regression model trained on different dataset. The length of each bar represents the absolute value of weight as the number shown near the right end of each bar; the color of each circle indicates the sign of the weight: red is negative, green is positive, yellow is almost zeros. sets.

Our first observation is that, these three models, although trained on different dataset, show similar patterns. This means the model makes sense, and the dataset does capture the actual citation characteristics though it's not a complete dataset.

According to the weights of each regressor, the citation patterns could be interpreted as:

(1) "MeanACiteB" has the largest weight: People tend to cite papers written by authors that they have cited before

(2) "Title similarity" has almost the second largest weight: Papers with similar content tend to cite each other.

(3) "In-degree of B" has large weight in Hep-Th: Popular papers tend to get more citations in physics field (rich gets richer).

(4) "Year difference" has negative weight: People tend to cite recent papers.

(5) "Author overlap" has almost zero weight: We originally thought that researchers prefer to cite their own papers, but the experiments told us that this effect is not significant.

4. MODIFIED PAGE RANK FOR IMPACT PREDICTION

We'd like to investigate the importance of each node (paper or author) in the citation network using the structure of the graph. This will enable many interesting applications including ranking the items according to impact, examining the correlation with popular impact indicators, discovery of trends in the graph, etc.

Inspired by the page rank algorithm to rank web pages according to popularity, we propose to rank the entities in the multi-layer hypergraph according to their interconnection relationships. This could be better than just using the node degree (number of citations) as a measurement of impact since it takes into account more information.

The original page rank algorithm assumes that the rank of a web page is the sum of contributions from pages that link to it, discounted by their out degrees:

$$r(n_i) = \sum_{j \to i} \frac{r(n_j)}{d(n_j)}$$

where $j \to i$ means entity j refers entity i.

A damping factor was further introduced to the model to improve robustness:

$$r(n_i) = \frac{a}{N} + (1-\alpha) \sum_{j \to i} \frac{r(n_j)}{d(n_j)}$$

This problem is an eigenvector problem of a modified adjacent matrix, which can be solved by iteration.

We cannot directly use this model for predicting importance of authors or papers, since the the number of references of a paper is not equivalent to the number of out links of a web page. The contribution of paper A to paper B by citing it should be proportional to the impact of paper A, but should not be inverse proportional to the number of references of paper A. As a result, we modify the page rank model as follows:

$$r(n_i) = \frac{a}{N} + (1-\alpha) \sum_{j \to i} r(n_j)$$

The algorithm converges very fast with our sparse matrix

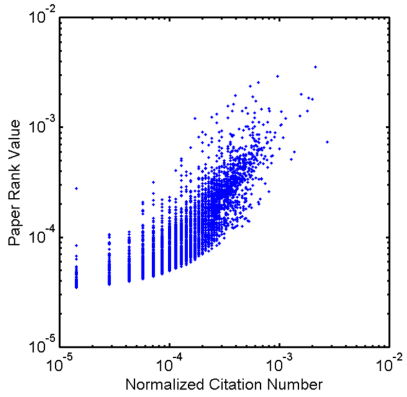

Figure 12: Log-log plot of paper rank vs. normalized citation number.

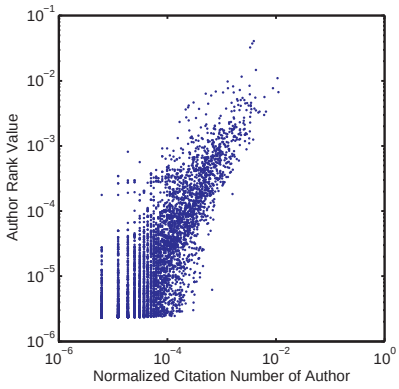

Figure 13: Log-log plot of author rank vs. normalized citation number.

implementation, and it's relatively robust in terms of parameter choice. The relative ranking of the papers stay almost the same was we vary the parameter α in the range between 0.1 to 0.4.

We compared the proposed paper rank with the normalized citation number of the investigated papers, as shown in Figure 12. We can see that the two criterion correlates well.

We also carried out the experiment on the citation network of the authors, the proposed author rank are plotted against the normalized citation number of the authors, and shown in Figure 13. We can see that the two criterion correlates well, too.

Lastly, we compared the proposed author rank with the h-score [5], which is a popular indicator for author impact. An author has h-score h if he/she has at least h papers that are cited at least h times each. We can see from figure 14 that the proposed rank correlates with h-score well.

We can also see from the figures that the correlation behavior looks different for different indicators, which means that maybe combining indicators will give better performance.

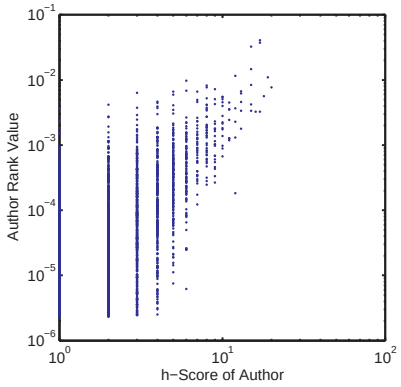

Figure 14: Log-log plot of author rank vs. h-score.

4.1 What can we infer from the citation network

We have already investigated the static property of the citation network, mostly in the macro scale. Furthermore, we can model the time varying properties of the graph. More specifically, we can look at the evolution of the citation for a specific paper over time, or look at how new scientific discoveries (e. g., SVM, Compressive Sensing, FFT, etc.) propagate in the academic community.

Take "database" for example, we plot in Figure 15 the number of papers with the word "database" in the title for different years, and also the average impact per paper for each year. We can see that although there is an exponential growth in the number of papers on the topic of "database", the actual impact per paper is diminishing after the foundational papers were published around 1976. The three papers in 1976 that have highest impact are:

1. System R: Relational Approach to Database Management.
2. The Notions of Consistency and Predicate Locks in a Database System.
3. Differential Files: Their Application to the Maintenance of Large Databases.

They've received 960, 1750, 259 citations until now according to Google Scholar search.

5. ACKNOWLEDGMENTS

The Proximity DBLP database is based on data from the DBLP Computer Science Bibliography with additional preparation performed by the Knowledge Discovery Laboratory, University of Massachusetts Amherst.

The authors would like to thank Professor Leskovec and the TAs for their kind assistant and valuable discussion.

6. REFERENCES

[1] V. Batagelj. Efficient algorithms for citation network analysis. *Arxiv preprint cs/0309023*, pages 1–27, 2003.

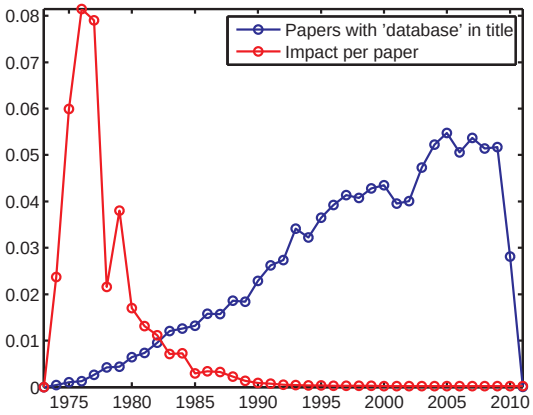

Figure 15: Trend of the word "database" in the titles of the papers.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. 2007.

[4] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[5] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69:131–152, 2006.

[6] B. H. Hall, A. B. Jaffe, and M. Trajtenberg. The nber patent citation data file: Lessons, insights and methodological tools. *NBER Working Paper 8498*, 2001.

[7] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):pp. 16569–16572, 2005.

[8] N. Hummon and P. Doreian. Connectivity in a citation network: The development of dna theory. *Social Networks*, 11:39–63, 1989.

[9] Y. Kajikawa, J. Ohno, Y. Takeda, K. Matsushima, and H. Komiyama. Creating an academic landscape of sustainability science: an analysis of the citation network. *Sustainability Science*, 2(2):221–231, July 2007.

[10] E. a. Leicht, G. Clarkson, K. Shedden, and M. E. Newman. Large-scale structure of time evolving citation networks. *The European Physical Journal B*, 59(1):75–83, Oct. 2007.

[11] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 915–924, New York, NY, USA, 2008. ACM.

[12] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429, New York, NY, USA, 2007. ACM.

[13] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *In SDM*, 2007.

[14] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[15] X. Shi, J. Leskovec, and D. A. McFarland. Citing for high impact. In *JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*, pages 49–58, New York, NY, USA, 2010. ACM.