Predicting Future Citation Count Using Bibliographic and Author Information of Articles

Takashi Matsui, Katsutoshi Kanamori, and Hayato Ohwada

***Abstract*—Electronic journal continues to spread rapidly among academic researchers. Citation count shown of an article in academic journals plays an important role in academia. It helps academic researchers in choosing suitable and reliable references for their research. It is unfortunate that recently published papers tend to be unapparent, since it has less citation count than other former works. This research proposes a new way to predict future citation count of an article using regression analysis by machine learning. As a result, citation count prediction of article is delivered using the new proposed method.**

***Index Terms*—Citation count, article, regression analysis, prediction.**

I. Introduction

Citation count of an article is the number of times an article is referred by another article as its related research. It is also used as an index of the importance and contributing value of an article in one academic field of research In 1927, Gross [1] for the first time used tabulation method to see citation count of chemical and science articles to evaluate the usage of an article by researchers. By then, the method is widely used to evaluate the quality of academic section, articles, or an academic journal.

Recently, way of using citation count has been changed as the electronic journals is developed. Citation count has been one center of attention in academia because of its importance. Without the help of citation count, researcher must read the whole paper to decide what the paper is about and then select the one that is most appropriate. With citation counts, the work has been done.

However, previous research has found some tendency that researchers are more interested in the former work of research than the recently published one, since it has higher citation counts [2]. Recently published paper become not as visible as the former one.

As the problem arises, the purpose of this research is to predict the future citation count of article using bibliographic information and a new method, in order to help researcher to evaluate also the recently published paper. It proposes how the importance of a recently published article could be promoted by discovering the index of importance from the future citation counts of the articles.

Manuscript received September 20, 2013; revised December 10, 2013.

Takashi Matsui, Katsutoshi Kanamori, and Hayato Ohwada are with the Tokyo University of Science, Noda-shi, Chiba-ken, Japan (e-mail: t-matsui@ohwada-lab.net, katsu@rs.tus.ac.jp, ohwada@rs.tus.ac.jp).

II. Related Works

Many research has been done about bibliographic information affects a quoting number by showing correlation of citation count and bibliographic information. For example, Podsakoff found that some features such as paper type, author's reputation, status of a journal, will affect the citation count [3]. Many features, such as the number of authors and the number of pages, have been also analyzed as the influence of citation count. However, research has not generated a prediction model yet.

Lokker [4] applied regression analysis using the index obtained three weeks after an article is published, and predicted the citation count for the next two year from the database publication of the medicine magazine of Britain (BMJ). They verified the accuracy of prediction by generating regression model. Result shows that it can predict the paper in the higher ranks 1/2 of citation count in the data set with sensitivity 83.3% and specificity 71.5%. It also show shows that it can predict the paper in the higher ranks 1/3 of citation count by sensitivity 66.1% and specificity 82.2%. Lawrence predicted by using the data from MEDLINE database [5]. They divided two types of examples, the positive example and the negative example. Citation count after ten years is predicted by support vector machine, and it becomes more than a threshold value. They indicate that the result of prediction was 80% or more of accuracy.

III. Model

A. Data

Data of articles in year 2010 used in this paper is collected from Web of Knowledge, one of academic citation indexing and search engines, where researchers can search and get the detail information of an article. In this research, reference, number of author and page, impact feature, and other article's attributes are used as features to show correlation with citation count in the past. Definition of impact feature used in this paper is an index to measure an influence degree of the magazine for the academic journals of the field of natural science, social science.

This research proposed one way to do prediction using the number of the citation count of the first author's previous article, assuming that other article from one author which has many citations would probably cited in similar times. From the previous research, author's number, sex, nationality, age, affiliated academic society, are examined as prediction feature of author. The new prediction way proposed in this study are limited by the data that is used, only average and maximum citation count of the articles of author in the past.

DOI: 10.7763/IJMLC.2014.V4.401

TABLE I: DETAILS OF THE EXTRACTION DATA

	AVERAGE	MAX	MIN
citations	24.20	207	0
Reference	50.39	828	1
Page	10.98	484	1
People	4.05	1817	1
Belong	389.26	400	12
ifNEW	4.64	153.46	0.04
ifAVE	4.41	102.41	0.03
Ref(5years)	25.98	789	0
autMAX	170.10	999	0
autAVE	17.80	248	0

B. Prediction Model

We used R language using kernlab [6] with in the SVM library. We executed a 10-fold cross-validation using the RBF kernel as the kernel function. SVM is a supervised learning algorithm for performing two values of classification proposed by Vapnik in 1960. It is a learning method for constructing a classifier to identify two classes and to perform determination of unknown data by learning the collectable data. In addition, SVM can deal with a practicable problem (non-linear problem) by using kernel function. In this research, we use SVR as the extension of SVM for regression analysis.

IV. EXPERIMENT

A. Experimental Data

About 6.000 information data of articles which published in 2010 are used in this research. By assuming that their citation count in 2013 are precise, we are able to predict citation count of the next three years. Table I shows the details of each feature extracted from the Web of Knowledge. "ifNEW" is the newest impact feature value, "ifAVE" expresses the average of the value from the past to the present. "Ref(5years)" expresses the number of articles published within five years among references. "autMAX" is the maximum citations of author's past papers, "autAVE" shows average citations of author's past papers. "Belong" is the rank of the institution where author belongs. Data is obtained from THES website, a website that provides up to date rank information, such as universities rank that listed in world top 400. We set the default rank of the value of the institution on 400, so data of rank that is more than 400 cannot be found in the list. We also use the feature "Type of article" other than those shown in Table I. Type of article is divided into five types. General article are the majority. There are other types, such as the proceedings paper, review paper, correction paper, and a software review paper. This feature is treated as a categorical variable.

B. Parameter Setting

"Cost" is a penalty parameter for the error. We verified the error and the correlation by using value between from 1 to 100. As a result, following Table II and Fig. 1 were provided. We can see from Fig. 1 and Table II, correlation and the error in "cost =10" shows a good value in comparison with other cases. Therefore, we sets the penalty parameter into 10 in the next verification.

C. Verification by Feature Deletion

We inspect some tests to check how much each feature used in this study influenced citation count prediction based on the parameter that we set. Table III shows the results. It shows the difference of the value between prediction error and correlation when each feature is removed and not remove (value of Table II as Cost=10). We indicate this value in Fig. 2.

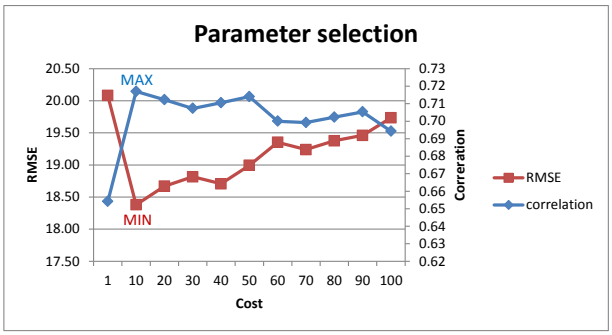

Fig. 1. Correlated with the RMSE in setting cost.

TABLE II: COST CALCULATION RESULT FOR PARAMETER SETTING

C	RMSE	correlation
1	20.08	0.65
10	18.38	0.72
20	18.67	0.71
30	18.82	0.71
40	18.71	0.71
50	18.99	0.71
60	19.35	0.70
70	19.24	0.70
80	19.38	0.70
90	19.46	0.71
100	19.73	0.69

TABLE III: DELETE FEATURE VALIDATION RESULTS

Delete Element	RMSE	correlation
autAVE	1.522	−0.059
autMAX	0.375	−0.012
Ref(5years)	0.357	−0.012
ifAVE	−0.060	0.001
ifNEW	0.130	−0.005
Belong	0.063	−0.003
People	0.213	−0.008
Type	−0.024	0.002
Page	0.360	−0.013
Reference	−0.052	0.002

D. Results and Discussion

We evaluated the predicted value using mean squared error of (1), and the Pearson correlation shown below (2).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(pi - xi)^2}{n}} \quad (1)$$

For equation (1), p_n shows a predicted value of citation count of the *i*-th article, x_n shows the actual measurement value of citation count.

$$Correlation = \frac{\sum_{i=1}^{n} (xi - \bar{x})(pi - \bar{p})}{\sqrt{\sum_{i=1}^{n} (xi - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (pi - \bar{p})^2}} \quad (2)$$

For equation (2), p_n and x_n shows a predicted value and the actual measurement value of citation count of the *i*-th article.

$\bar{p}$ and $\bar{x}$ shows average of each.

As can be seen from Table II, we obtained the result of mean squared error (RMSE) =18.38 and Pearson correlation (correlation) =0.72. Fig. 3 shows the relationship between the actual measurement value and the predicted value obtained using regression analysis. Horizontal axis shows the predicted value and the vertical axis shows the actual measurement. When there is a plot near to the diagonal line, it shows that the result are acceptable. We were able to do to some extent good prediction because in general there is a strong correlation when the correlation coefficient has the value of 0.7 or more. About the error, we should endeavor in order to make an error smaller.

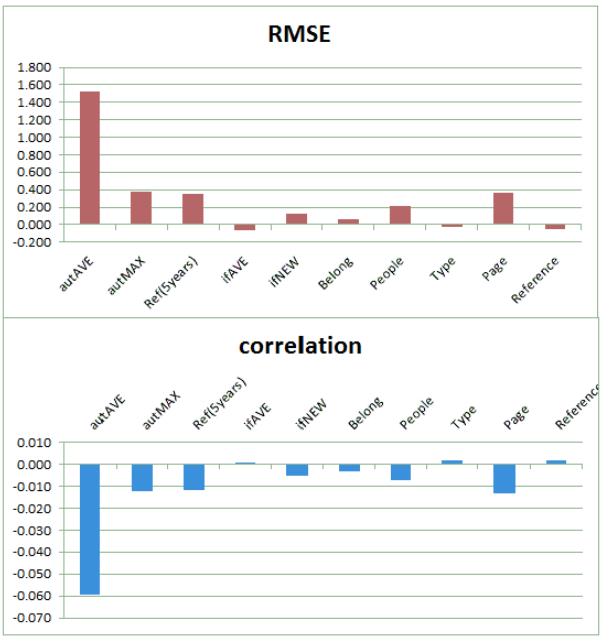

Fig. 2. Change of RMSE and correlation due to feature deletion.

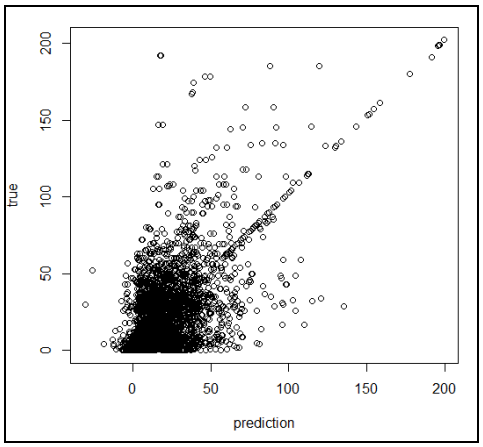

Fig. 3. Scatter plot of the measured and predicted value.

V. Conclusion

This research proposed a new way of future citation count prediction. Future citation count of articles are predicted using information of author and bibliographical in order to evaluate the recently published articles. We obtained the result of mean squared error (RMSE) =18.38 and Pearson correlation (correlation) =0.72. We have the space to improve errors.

However, the relationship between the number of citations of articles and the number of citations of articles of the author in the past has not been analyzed in this paper. For further research, data of the articles should be increased and the research should aim at the improvement of the precision by the examination of a new feature.

References

[1] P. Gross, "College libraries and chemical education," *Science*, vol. 66, pp. 385–389, 1927.
[2] K. Yokoi, *Impact of electronic journals on literature available to researchers: Society for Information and Research Competition Research*, 2010.
[3] P. M. Podsakoff, S. B. Mackenzie, D. G. Bachrach, and N. P. Podsakoff, "The influence of management journals in the 1980s and 1990s," *Strategic Management Journal*, vol. 26, no. 5, pp. 473–488, 2005.
[4] C. Lokker, K. A. McKibbon, R. J. McKinlay, N. L. Wilczynski, and R. B. Haynes, "Prediction of citation counts for clinical articles at two years using data available within three weeks publication: retrospective chort study," *BMJ*, vol. 336, 2008.
[5] D. F. Lawrence and C. F. Aliferis, "Using contentbased and bibleo-metric features for machine learning models to predict citation counts in the biomedical literature," *Scientometrics*, vol. 85, 2010.
[6] A. Karatzoglou, A. Smola, K. Hornik, and Achim Zeileis, "Kernlab – An S4 Package for Kernel Methods in R," *Journal of Statistical Software*, vol. 11, issue 9, November 2004.

T. Matsui graduated from the Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science, Noda City, Japan, in 2013.

He is a student at Tokyo University of Science Graduate School, Division of Science and Engineering Industrial Administration Master's course since 2012, Noda City, Japan. His research interests are in the field of recommendation systems.

K. Kanamori earned a doctorate in information science at the Tokyo University of Science, Noda City, Japan, in 2009.

He is a research associate at Tokyo University of Science, Department of Industrial Administration, Faculty of Science and Technology. His research interests are in the fields of Artificial Intelligence.

H. Ohwada graduated from the Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science, Noda City, Japan, 1983. Then he graduated from Tokyo University of Science Graduate School, Division of Science and Engineering Industrial Administration Doctoral course Completed program with degree, Noda City, Japan, 1988.

He was a research associate (Tokyo University of Science) from 1988 to 1998, lecturer (Tokyo University of Science) from 1999 to 2000, and associate professor (Tokyo University of Science) from 2001 to 2004. Then he is a professor at Tokyo University of Science Faculty of Science and Engineering Department of Industrial Administration from 2005. His research interests are in the fields of Inductive Logic Programming and Bioinformatics.