# Predicting High Impact Academic Papers Using Citation Network Features

Daniel McNamara[1], Paul Wong[2], Peter Christen[1], and Kee Siong Ng[1,3]

[1] Research School of Computer Science
[2] Office of Research Excellence
The Australian National University, Canberra, Australia
[3] EMC Greenplum
dpmcna@gmail.com, {paul.wong,peter.christen}@anu.edu.au,
keesiong.ng@emc.com

**Abstract.** Predicting future high impact academic papers is of benefit to a range of stakeholders, including governments, universities, academics, and investors. Being able to predict 'the next big thing' allows the allocation of resources to fields where these rapid developments are occurring. This paper develops a new method for predicting a paper's future impact using features of the paper's neighbourhood in the citation network, including measures of interdisciplinarity. Predictors of high impact papers include high early citation counts of the paper, high citation counts by the paper, citations of and by highly cited papers, and interdisciplinary citations of the paper and of papers that cite it. The Scopus database, consisting of over 24 million publication records from 1996-2010 across a wide range of disciplines, is used to motivate and evaluate the methods presented.

## 1 Introduction

This paper seeks to produce a method which, given a database of academic publications and citations between them, can predict future high impact papers. The topic of this paper is a part of an effort to provide ongoing analytical support to decision and policy development for the Commonwealth of Australia [1,2,3]. One aspect of this effort is to develop an 'early warning system' to predict, anticipate and respond to emerging research trends.

It is amply clear that R&D operates in an increasingly competitive environment, where the traditional US and Europe dominance is under direct challenge by a number of Asian countries. Australia, with a small population base and slightly more than 2% GDP spend on R&D [2], will need to compete and stretch its investment dollar in more creative and efficient ways. Decision and policy makers thus need to marshal all available resources and intellectual capital to develop sound strategies to remain competitive on a global scale. The utilisation of data mining techniques to make predictions about citations of scholarly publications, taken as a proxy for the onset of research breakthroughs, when used in combination with other relevant leading indicators, can potentially provide

J. Li et al. (Eds.): PAKDD 2013 Workshops, LNAI 7867, pp. 14–25, 2013.

competitive intelligence for strategy development. While Australia may not be able to invest in R&D to the same extent as other economic powerhouses to take advantage of being 'the first mover', with the development of insightful predictive analytics over a range of data sources, it can become an 'early adopter' and develop national research capabilities in an agile and timely manner. The motivation behind this paper is to develop useful predictive models to empower decision and policy making.

This paper is organised in the following way. Section 2 reviews related work, and the Scopus database is presented in Sect. 3. Section 4 covers the methods used in this paper, including a suitable measure of paper impact, predictive features from the paper's citation network neighbourhood, and prediction algorithms. The results of applying these methods to the Scopus database are shown in Sect. 5. Section 6 presents the conclusion and future work.

## 2 Related Work

There is a rich literature on the topics of defining and predicting the impact of academic papers. Citation counts are the traditional and most straightforward way of measuring the impact of an individual paper. Citation counts have been used to distinguish between 'classic' papers which continue to be cited long after publication, and 'ephemeral' papers which rapidly cease to be cited [4]. We seek to formalise the notion of a classic or high impact paper.

Raw citation counts vary significantly between disciplines, making it a challenge to find an impact measure which is fair to papers from all fields. One approach has been to divide a paper's citations by its disciplinary average [5,6]. A critique found that dividing by disciplinary average still generates different distributions across disciplines [7]. Other studies have instead worked with the disciplinary percentile rank, for example proposing that the top 1% of papers in each discipline should be considered classics [8,9]. As detailed in Sect. 4.1, this paper builds on the percentile rank approach, but explicitly considers the possibility of multiple disciplinary classifications for a single paper, and favours papers with enduring influence using exponential discounting favouring more recent citations.

There are a range of features that can be used as predictors of a paper's future impact. These include citations of a paper soon after it is published [10,11]; measures of network centrality such as average shortest path length, clustering coefficient and betweenness centrality [12]; the paper's authors' previous work [13,14]; and keywords from the text of the paper [15]. The framework of information diffusion emphasises that ideas, like epidemics, spread through networks [16,17]. We therefore expect that a paper's position in the network will be a determinant of the impact of its ideas. The theory of 'preferential attachment' suggests that in evolving networks, new nodes favour connections to existing highly connected nodes [18]. It has been proposed that when nodes span boundaries or 'structural holes' between previously disparate parts of intellectual networks, they induce structural variation and hence become influential [19,20,21].

This paper draws upon and examines these arguments by evaluating whether the number and interdisciplinarity of citations by and of a paper are predictive of its future impact.

Previous research has investigated the effect on future citation counts of paper interdisciplinarity, measured by the proportion of citations made by a paper outside its own discipline [22,23]. This study builds on this approach but additionally distinguishes between closely and distantly related disciplines, allows multiple disciplines per paper, and considers the interdisciplinarity of citations of papers citing and cited by the original paper.

The experiments presented in previous studies using network features to predict academic impact often use datasets from individual fields [12,20] or institutions [24]. This paper is unusual in presenting results over a dataset as large and broad as the Scopus database. Additionally, it incorporates the dynamic nature of the citation network by considering citations disaggregated by year.

## 3 The Scopus Database

Scopus is a proprietary database of metadata records of academic papers. The database is owned by the publisher Elsevier and is one of a small number of major multidisciplinary bibliometric databases along with Thomson's Web of Science and Google Scholar. The version of Scopus used in this paper contains metadata records for 24,097,496 papers published during the years 1995-2012. The years 1996-2010 are complete, with more recent records yet to be comprehensively added. The records include title, authors including their countries and institutional affiliations, journal, document type, abstract, keywords, subject areas, and citations of and by the paper.

Figure 1 shows the disiplinary coverage of the Scopus database, which focuses on medicine and science. The All Science Journal Classification (ASJC) system is used, with papers hierarchically grouped into 334 disciplines at the 4-digit level and 27 disciplines at the 2-digit level [25]. A given paper may have zero, one, or multiple disciplinary classifications.

## 4 Methods

We consider the task of predicting the future impact of papers over a horizon of $\tau$ years from the present. We assume that citations by the paper of papers published up to $\kappa$ years before its publication are available. The parameter $\delta$ is the number of years of citations of the paper available at the time of prediction.

The database of academic papers considered can be represented as a set $N$, and an individual paper is represented by $n \in N$. $N_t \subset N$ refers to the set of all papers published in year $t$. Citations are represented by $a_{mn}$, which is equal to 1 if paper $m$ cites paper $n$, and 0 otherwise. The paper impact vector of length $|N|$ is represented as $\mathbf{y}$, where $y_n = y(n)$ is the impact of paper $n$.

We assume that each paper is classified as belonging to one or more disciplines $k \in K_0$, where $K_0$ is the set of disciplines. Further, we assume that the elements

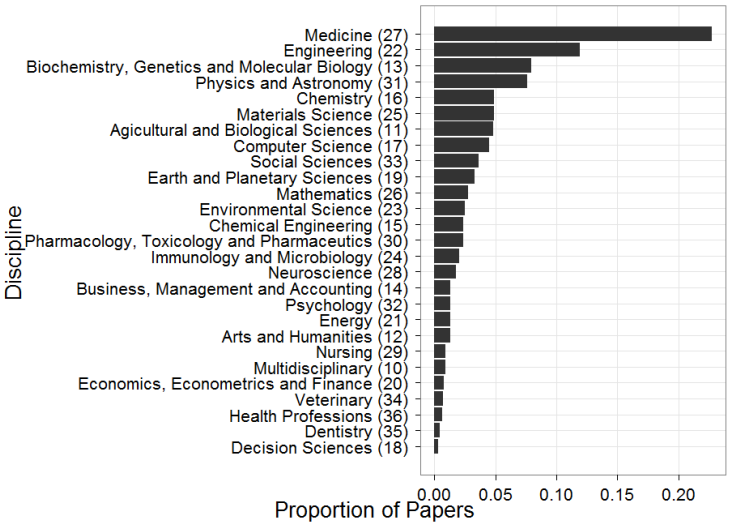


**Fig. 1.** Scopus coverage by discipline. The 2-digit ASJC codes of each discipline are shown in brackets.

of $K_0$ may be hierarchically grouped at levels of discipline similarity in the range $i \in [0, \omega]$, where $K_i$ is the set of groups at level $i$ in the hierarchy. At level 0, each code is assigned its own group; at level $\omega$, all codes are in the same group; and at intermediate levels, codes are assigned to groups containing some but not all other codes. In the case of Scopus, $\omega = 2$, $K_0$ contains a group for each 4-digit ASJC code, $K_1$ contains the 2-digit ASJC code discipline groups, and $K_2$ contains all disciplines in one group. Disciplinary classifications are represented by $c_{nk}$, the proportion of classifications of paper $n$ as discipline $k$.

### 4.1 Measuring Paper Impact

Our goal is to predict the impact of a given paper. To do this we must first determine how to measure impact, a topic discussed in Sect. 2. The number of citations of a paper is a good starting point.

We would like to take into account citations over several years, favouring recent citations. This is to find papers that have a lasting influence, rather than those that are popular for only a brief time. We do this using exponential decay in (1). The parameter $r \in [0, 1]$ controls the rate of decay, and can also be called the discount factor.

$$y'(n) = \sum_{t=\delta+1}^{\tau+\delta} r^{\tau+\delta-t} \sum_{m \in N_t} a_{mn} \tag{1}$$

Some disciplines cite more frequently than others. We accommodate this by finding the percentile rank of $n$ across all papers in its discipline(s), including papers

from multiple years. This is shown for an individual discipline in (2). We use the indicator function $I(a,b) = 1$ if $a > b, 0$ otherwise. These ranks are combined to a single rank in (3), where $y$ is the paper impact metric. Using percentile rank makes the paper impact distributions of all disciplines approximately uniform in the range $y(n) \in [0,1]$.

$$y''(n,k) = \frac{\sum_{m \in N} I(y'(n), y'(m)) I(c_{mk}, 0)}{\sum_{m \in N} I(c_{mk}, 0)} \tag{2}$$

$$y(n) = \sum_{k \in K} c_{nk} y''(n,k) \tag{3}$$

We propose fixing a threshold $\lambda$, such that for a set of papers $N$, the high impact or classic papers $N^*$ are defined according to (4). A similar approach has previously been suggested [9], identifying high impact or classic papers as the top 1% most highly cited papers in each discipline. Using this 1% threshold corresponds to setting $\lambda = 0.99$. The paper impact $y(n)$, referred to as the target variable in the context of prediction, has the additional advantages that it takes into account papers with multiple classifications, and weights later citations more heavily to measure the ongoing effect of a paper. Note that this definition of classics is relative to the set of papers being considered, so that every set of papers will always have a fixed proportion of classics.

$$N^* = \{n \in N | \frac{1}{|N|} \sum_{m \in N} I(y(n), y(m)) \geq \lambda\} \tag{4}$$

### 4.2 Predictive Features

There are many potential predictors of future impact. In this paper only properties of the paper's neighbourhood in the citation network are considered. As described in Sect. 2, this is motivated by the framework of information diffusion which states that a node's position in a network impacts its ability to have intellectual influence. The features $f$ used are specified in Table 1.

The paper's disciplinary classifications and the annual citations of and by the paper are the base citation network neighbourhood features considered. In the case $\delta = 0$, we only have information about papers cited by the paper, whereas if $\delta > 0$ we also have information about papers that cite the paper.

Previous work has proposed that interdisciplinary work is likely to be more influential [20,21,26], since it fills in 'structural holes' in the network. This paper seeks to quantitatively evaluate this hypothesis, extending previous work which measures the interdisciplinarity of a paper using the proportion of its citations that are of papers in other disciplines [22,23]. In this study, individual papers may have multiple disciplinary classifications, and the classifications may be hierarchically grouped at levels in the range $i \in [0, \omega]$. The interdisciplinarity type $i$ means that at least one pair of classifications of the cited and citing

**Table 1.** Summary of features used for predicting the impact of individual papers. 'b' stands for citations *by* a paper, 'o' stands for citations *of* a paper, and moving outwards from the original paper these citation types are added to the feature set name. $K_\nu$ is the set of discipline groups at the hierarchy level $\nu$, $\omega$ is the number of levels in the hierarchical grouping of disciplines, $\kappa$ is the years of citations by the paper available, and $\delta$ is the year of prediction relative to the paper's publication.

| **Feature Set** | **Feature Set Size** | **Feature Set Description** |
|---|---|---|
| **c** | $\lvert K_\nu \rvert - 1$ | $c_k$ is the proportion of paper's disciplinary classifications in discipline group $k$ |
| **b** | $\kappa$ | $b_t$ is the number of citations by paper in year $t$ |
| **B** | $(\omega + 1)\kappa$ | $b_{it}$ is the proportion of cited papers of interdisciplinarity type $i$ published in year $t$ |
| **o** | $\delta$ | $o_t$ is the number of citations of paper in year $t$ |
| **O** | $(\omega + 1)\delta$ | $o_{it}$ is the proportion of citing papers of interdisciplinarity type $i$ published in year $t$ |
| *bo* | 1 | Average number of citations of cited papers |
| **bo** | $\omega + 1$ | $bo_i$ is the average proportion of citations of cited papers of interdisciplinarity type $i$ |
| *oo* | 1 | Average number of citations of citing papers |
| **oo** | $\omega + 1$ | $oo_i$ is the average proportion of citations of citing papers of interdisciplinarity type $i$ |

papers are in the same group at hierarchy level $i \in [0, \omega]$, but not at any lower hierarchy level. In the context of Scopus, interdisciplinarity type 0 indicates that the two papers share a 4-digit ASJC code, type 1 indicates that they share a 2-digit ASJC code but no 4-digit ASJC code, and type 2 indicates that they share no 2-digit ASJC code. The proportions of citations of and by the paper of each interdisciplinary type for each year are used as predictive features.

Going one level further out in the neighbourhood of the paper, the number and interdisciplinarity of citations of those papers cited by and citing the paper are considered. These 'higher order' features are of interest since they measure the effect of citing and being cited by 'authorities'.

### 4.3 Prediction Algorithms

Several algorithms are used for making predictions of the target variable based on the features outlined in Sect. 4.2. These are linear regression, decision trees and random forests [27]. These were chosen since they are known to be effective prediction algorithms with readily available implementations [28,29,30].

## 5 Experiments and Discussion

The Scopus Database detailed in Sect. 3 was used to evaluate the methods presented in Sect. 4. A training set with predictors and response variables completely

available before the year of prediction is required to train the prediction algorithm. In our experiments, the training set consists of Scopus database papers published in 2000 and the test set consists of papers published in 2005.

Furthermore, the papers considered are restricted to those with at least one ASJC disciplinary classification, and to citations of and by those papers where the other paper also had at least one ASJC disciplinary classification. This is the case in more than 98% of the dataset and eliminates the complexity of dealing with missing data. The final training set consists of 1,184,842 papers and the test set of 1,704,624 papers.

We use the following parameter settings: the prediction horizon $\tau = 3$, a common timeframe for decision-makers; citations of papers up to $\kappa = 4$ before the paper's publication are included to fit into the data available; experiments where $\delta = 0$ and $\delta = 2$ are tried to assess the impact of varying the year of prediction relative to the paper's publication; $\omega = 2$ so that citation interdisciplinarity can be measured using 2-digit and 4-digit ASJC codes; $\nu = 1$ so that the 2-digit ASJC codes of papers are made available to the prediction algorithm; and the discount rate $r = 0.9$ to reward papers with enduring influence.

### 5.1 Feature Ranking Using Spearman Coefficient

Spearman's rank correlation coefficient $\rho$, a standard measure of the dependence of two variables using a monotone function, was taken for each of the features described in Sect. 4.2 and the target variable $\mathbf{y}$. The top features ranked by their $\rho$ value with the target variable $\mathbf{y}$ are shown in Table 2. Figure 2 shows a dendogram of the top features, which are hierarchically clustered using the distance metric defined in (5). The unsupervised feature clusters correspond closely to the groupings defined in Table 1.

$$dist(f_1, f_2) = 1 - |\rho(f_1, f_2)| \tag{5}$$

The variables not known at the time of the paper's publication are shown as NA in the $\rho_0$ column. The feature sets **B**, **O**, **bo** and **oo** are the proportions of citations of a particular interdisciplinarity type (see Table 1 for details). For each of these feature sets, the papers for which there are no such citations are excluded from the Spearman coefficient calculations, since these proportions are not meaningful for these papers. In the prediction algorithms these features are given a value of 0 in these cases, to avoid the problem of missing data.

Table 2 shows that the most predictive variables are $o_2$ and $o_1$, the number of citations of the paper 2 years and 1 year after publication respectively, which are also clustered together in Fig. 2. This is intuitive since we would expect citations in early years to have a strong positive correlation with those in later ones.

The next most predictive variables are those in **b**, the number of citations made by the paper, which also form a cluster in Fig. 2. This suggests that papers which cite more are themselves more highly cited. A high number of citations may suggest that the paper is thoroughly researched, or may be a review paper.

*bo*, the average number of citations of cited papers, and *oo*, the average number of citations of citing papers, are both positively correlated with the target

**Table 2.** Top 10 features, ranked by absolute value of Spearman coefficient $\rho$ for the prediction task where $\delta = 2$. The subscripts 0 and 2 refer to the value of $\delta$ used.

| Feature | Rank | $\rho_2$ | $\rho_0$ | Description |
|---|---|---|---|---|
| $o_2$ | 1 | 0.6757 | NA | Citations of paper at $t = 2$ |
| $o_1$ | 2 | 0.5887 | NA | Citations of paper at $t = 1$ |
| $b_{-3}$ | 3 | 0.4361 | 0.4463 | Citations by paper at $t = -3$ |
| $b_{-2}$ | 4 | 0.4327 | 0.4489 | Citations by paper at $t = -2$ |
| $b_{-4}$ | 5 | 0.4264 | 0.4325 | Citations by paper at $t = -4$ |
| $b_{-1}$ | 6 | 0.3733 | 0.3919 | Citations by paper at $t = -1$ |
| $oo$ | 7 | 0.2735 | NA | Average number of citations of citing papers |
| $bo$ | 8 | 0.2266 | 0.2594 | Average number of citations of cited papers |
| $oo_2$ | 9 | 0.1346 | NA | Proportion of citations of citing papers of most interdisciplinary type |
| $o_{21}$ | 10 | 0.1341 | NA | Proportion of citing papers of most interdisciplinarity type published in year $t = 1$ |

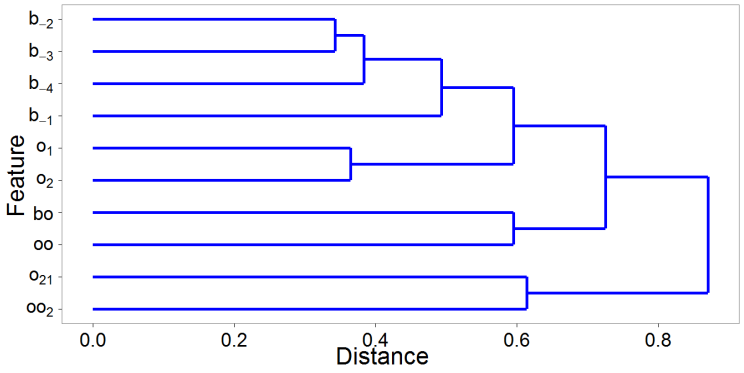


**Fig. 2.** Dendogram of top 10 features as described in Table 2. The distance between features is given by (5).

variable, and form a cluster in Fig. 2. The first result suggests that citing papers that are 'authorities' is advantageous for future citations. The second suggests that being cited by 'authorities' is also advantageous.

There is also evidence that interdisciplinarity is a predictor of future citations. $oo_2$, the proportion of citations of citing papers which are most interdisciplinary, is positively correlated with the target variable. So is $o_{21}$, the proportion of citations of the paper of the most interdisciplinarity type published in year $t = 1$. Other features indicating citations of the most interdisciplinary type fell just outside the top 10 and showed positive correlations. Previous studies have found that interdisciplinarity has a mix of both positive and negative correlations with paper impact depending on the paper's discipline [22,23], and no clear correlation overall [23]. While individual disciplines are not studied here, there are weak positive correlations between features indicating interdisciplinarity and impact overall. A possible reason for this discrepancy is that in this study features of interdisciplinarity are

disaggregated by year, and include citations of the paper and citations of cited and citing papers, in addition to citations by the paper as in [22,23].

The correlations with impact calculated from the year of publication follow a similar pattern overall to those with impact calculated from two years after publication. However, citations by a paper matter more to its citations soon after publication than several years after, when other factors become more dominant.

It is possible to test significance of the Spearman coefficients using the null hypothesis that there is no correlation between the target variable and the feature [31]. A test statistic can be generated for a Student's $t$-distribution with $|N| - 2$ degrees of freedom. The values of this test statistic showed that each of the top 10 features shown in Table 2 were statistically significant.

### 5.2 Prediction Results

Root mean square error (RMSE) is a standard measure of the accuracy of predictions in a regression context. Linear regression, decision trees and random forests, as implemented here, all learn parameter values which minimise the sum of squares error (and hence RMSE) over the training set. In order to get a sense of how well our prediction algorithms are performing, it is helpful to have a baseline. A simple baseline is the mean target variable of the training set. This is also the optimal constant value which minimises the RMSE over the training set. This baseline achieved RMSE scores of 0.3645 for the training set and 0.3797 for the test set. We evaluate prediction performance by calculating the percentage improvement on this baseline.

The test set score of each feature set and algorithm combination is shown in Fig. 3. As expected, all the algorithms found predicting a paper's future citations from two years after publication ($\delta = 2$) much easier than predicting its citations from the year of its publication ($\delta = 0$).

The best performing algorithm was random forest. For the prediction task where $\delta = 0$, it achieved an 18.38% improvement on the baseline, and for $\delta = 2$, it achieved a 34.44% improvement. It is not surprising that as an ensemble method it performed better than the individual regression methods. It is noticeable that adding more features, particularly in the task predicting from two years after publication, actually made its performance slightly worse. This is likely related to the fact that each split only uses a sample of the features. When more features are added in, it may miss the most important features.

Other metrics offer further insights into the algorithm's performance. Using $R^2$, which can be interpreted as the proportion of variation in the target variable explained by the prediction, random forest's best test set results were 0.3342 for the $\delta = 0$ task, and 0.5697 for the $\delta = 2$ task. A classification approach, using the definition of classic papers from (4), showed that 8.28% of test set classic papers were successfully predicted for $\delta = 0$, and 38.73% for $\delta = 2$.

In the case of an individual decision tree, its results were not quite as strong as random forest, but were in similar ranges for the two tasks. Linear regression did not perform as well as the other algorithms, though it showed improvement when information about the interdisciplinarity of citations was included.

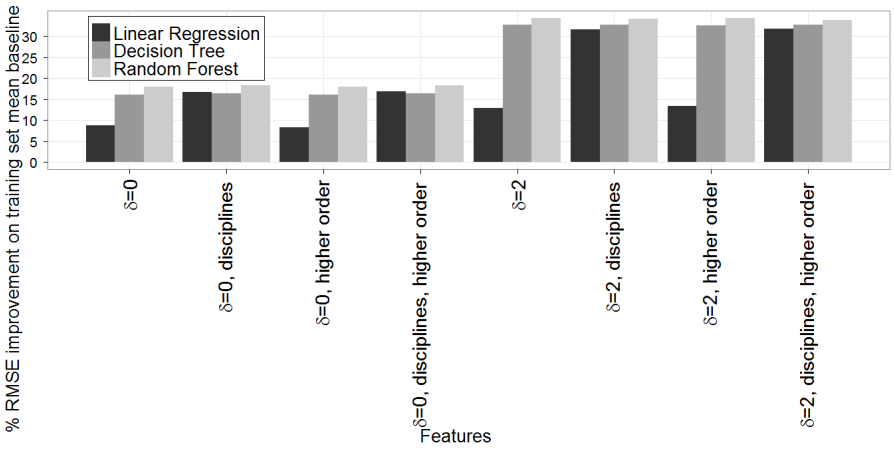


**Fig. 3.** Performance of prediction algorithms with a range of features, as described in Sect. 4.2

## 6 Conclusion and Future Work

This paper presented a new method for the prediction of the future impact of individual papers. Predictive features based on a paper's position in the citation network were used, drawing upon and evaluating previous research on information diffusion in networks, which suggests that nodes which are highly connected [18] and span network boundaries [19,20,21] are likely to be more influential. The method was implemented and evaluated using an exceptionally large and broad academic database, Scopus, comprising over 24 million papers from 1996-2010.

The notion of a classic or high impact paper was formalised using a novel metric of paper impact. This is a weighted average of the percentile ranks of citations of a paper across its disciplinary classifications, with an exponential discount rate favouring more recent citations to identify papers with enduring influence. The number of citations of the paper in the early years after publication, the number of citations by the paper, the average number of citations of citing and cited papers, and more interdisciplinary citations of the paper and of citing papers, were found to positively correlate with the paper's future impact.

Three prediction algorithms - linear regression, decision trees and random forest - were proposed to predict the future impact of individual papers. The percentage of RMSE improvement over the training set mean baseline was used to evaluate prediction performance. The results found that random forest was most predictive, achieving an 18% improvement predicting from the year of a paper's publication, and a 34% improvement predicting from two years after it.

This predictive capacity can assist universities, governments and investors by alerting them to future high impact papers, as well as to researchers, institutions and fields producing such papers. There is exciting potential for such an analytical tool to assist policy development and decision making.

Improved prediction can be achieved using a longer time window; adding other features such as author, journal and article text; and employing more sophisticated prediction algorithms such as support vector machines. Another option is the collective classification approach, simultaneously making predictions for individual papers and allowing these predictions to influence each other [32]. While in this paper the task is predicting citation counts, link prediction in the citation network [33] would provide the user with more detail.

The predictions about individual papers may be aggregated at the field level using co-citation analysis [34]. A co-citation graph can be constructed, where predicted classic papers are nodes, and edges occur when the citation behaviours of two papers are sufficiently similar using a metric such as weighted cosine similarity. Emerging fields of research can be predicted using community detection in the co-citation network of predicted high impact papers, for example by extracting the maximal cliques or components of the network. The authors of this paper anticipate a forthcoming publication on this topic, with the goal of creating a powerful tool to aid strategic research investment.

## References

1. Australian Government: Australia in the Asian Century White Paper (2012)
2. Department of Industry, Innovation, Science, Research and Tertiary Education: 2012 National Research Investment Plan (2012)
3. Office of the Chief Scientist of Australia: Health of Australian Science (2012)
4. Price, D.: Networks of scientific papers. Science 149(3683), 510–515 (1965)
5. Castellano, C., Radicchi, F.: On the fairness of using relative indicators for comparing citation performance in different disciplines. Archivum Immunologiae et Therapiae Experimentalis 57(2), 85–90 (2009)
6. Radicchi, F., Fortunato, S., Castellano, C.: Universality of citation distributions: Toward an objective measure of scientific impact. Proc. Natl. Acad. Sci. USA 105(45), 17268–17272 (2008)
7. Waltman, L., van Eck, N.J., van Raan, A.F.: Universality of citation distributions revisited. J. Am. Soc. Inf. Sci. Technol. 63(1), 72–77 (2012)
8. Small, H.: Tracking and predicting growth areas in science. Scientometrics 68(3), 595–610 (2006)
9. Upham, S., Small, H.: Emerging research fronts in science and technology: patterns of new knowledge development. Scientometrics 83(1), 15–38 (2010)
10. Adams, J.: Early citation counts correlate with accumulated impact. Scientometrics 63(3), 567–581 (2005)
11. Manjunatha, J.N., Sivaramakrishnan, K.R., Pandey, R.K., Murthy, M.N.: Citation prediction using time series approach KDD cup 2003 (task 1). SIGKDD Explor. Newsl. 5(2), 152–153 (2003)
12. Shibata, N., Kajikawa, Y., Matsushima, K.: Topological analysis of citation networks to discover the future core articles. J. Am. Soc. Inf. Sci. Technol. 58(6), 872–882 (2007)

13. Castillo, C., Donato, D., Gionis, A.: Estimating number of citations using author reputation. In: Ziviani, N., Baeza-Yates, R. (eds.) SPIRE 2007. LNCS, vol. 4726, pp. 107–117. Springer, Heidelberg (2007)
14. Yan, R., Tang, J., Liu, X., Shan, D., Li, X.: Citation count prediction: learning to estimate future citations for literature. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 2011, pp. 1247–1252 (2011)
15. Yogatama, D., Heilman, M., O'Connor, B., Dyer, C., Routledge, B.R., Smith, N.A.: Predicting a scientific community's response to an article. In: EMNLP 2011, pp. 594–604 (2011)
16. Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chávez, C., Wojick, D.: Population modeling of the emergence and development of scientific fields. Scientometrics 75(3), 495–518 (2008)
17. Goffman, W., Newill, V.A.: Generalization of epidemic theory: An application to the transmission of ideas. Nature 204(4955), 225–228 (1964)
18. Barabási, A., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)
19. Burt, R.S.: Structural holes: the social structure of competition. Harvard University Press, Cambridge (1992)
20. Chen, C.: Predictive effects of structural variation on citation counts. J. Am. Soc. Inf. Sci. Technol. 63(3), 431–449 (2012)
21. Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., Pellegrino, D.: Towards an explanatory and computational theory of scientific discovery. J. Informetr. 3(3), 191–209 (2009)
22. Adams, J., Jackson, L., Marshall, S.: Bibliometric analysis of interdisciplinary research. Report to Higher Education Funding Council for England (2007)
23. Larivière, V., Gingras, Y.: On the relationship between interdisciplinarity and scientific impact. J. Am. Soc. Inf. Sci. Technol. 61(1), 126–131 (2009)
24. Nankani, E., Simoff, S.: Predictive analytics that takes in account network relations: A case study of research data of a contemporary university. In: Proceedings of the 8th Australasian Data Mining Conference, AusDM 2009, pp. 99–108 (2009)
25. Scopus: Scopus custom technical requirements, Version 2.0 (2009)
26. Guo, H., Weingart, S., Börner, K.: Mixed-indicators model for identifying emerging research areas. Scientometrics 89(1), 421–435 (2011)
27. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
28. Liaw, A., Wiener, M.: Package 'randomForest': Breiman and Cutler's random forests for classification and regression (2012)
29. R Documentation: Fitting linear models (2012)
30. Therneau, T.M., Atkinson, E.: An introduction to recursive partitioning using the RPART routines (2011)
31. R Documentation: Test for association/correlation between paired samples (2012)
32. Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., Eliassi-Rad, T.: Collective classification in network data. AI Magazine 29(3), 93–106 (2008)
33. Shibata, N., Kajikawa, Y., Sakata, I.: Link prediction in citation networks. J. Am. Soc. Inf. Sci. Technol. 63(1), 78–85 (2012)
34. McNamara, D.: A new method for the prediction of emerging fields of research. Honours thesis, Australian National University (2012)