Predicting high impact academic papers using citation network features

Daniel McNamara¹, Paul Wong², Peter Christen¹, and Kee Siong Ng¹,³

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...
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].