Modeling Time-dependent and -independent Indicators to Facilitate Identification of Breakthrough Research Papers

Holly N. Wolcott[1], Matthew J. Fouch[1], Elizabeth Hsu[2], Catherine Bernaciak[1], James Corrigan[2], and Duane Williams[1]

holly.wolcott@thomsonreuters.com
[1]Intellectual Property & Science, Thomson Reuters, Rockville, MD 20850 (USA)

corrigan@mail.nih.gov
[2]Office of Science Planning and Assessment, National Cancer Institute, Bethesda, MD 20892 (USA)

Abstract

Research funding organizations invest substantial resources to stay current with important research findings within their mission areas to identify and support promising new lines of inquiry. To that end, we continue to pursue the development of tools to identify research publications that have a strong likelihood of driving new avenues of research. This research-in- progress paper describes our work incorporating multiple time-dependent and -independent features of publications into a model that aims to identify candidate breakthrough papers as early as possible following publication. We used multiple Random Forest models to assess the ability of indicators to reliably distinguish a gold standard set of breakthrough publications as identified by subject matter experts from among a comparison group of similar *Thomson Reuters Web of Science*™ publications. These indicators will be selected for inclusion in a multi-variate model to test their predictive value. Prospective use of these indicators and models is planned to further establish their reliability.

Conference Topic

Indicators

Introduction

The National Cancer Institute (NCI) of the US National Institutes of Health (NIH) continues to show a commitment to encouraging transformative research, which the NIH recognizes on its Transformative Research Award website as “unconventional research projects that have the potential to create or overturn fundamental paradigms.” Key requirements for identifying and nurturing these potential scientific breakthroughs are an enhanced understanding of the research landscape and awareness of novel approaches with great potential.

Defining Breakthrough Publications

The term "breakthroughs" has been used in prior work by Thomson Reuters (Ponomarev et al., 2014) and operationally, breakthrough publications have previously been defined as those that are highly cited and result in a change in research direction. The body of literature addressing breakthrough publications also uses the term “transformative research.” Here, we define a breakthrough publication as an article that results from transformative research. In 2007, the National Science Board (NSB) defined transformative research as “research driven by ideas that have the potential to radically change our understanding of an important existing scientific or engineering concept or leading to the creation of a new paradigm or field of science or engineering. Such research also is characterized by its challenge to current understanding or its pathway to new frontiers” (NSB, 2007).

Prior Work Identifying Breakthrough Publications

Much of the research literature on breakthroughs focuses on retrospective identification of breakthroughs or pivotal points within a specific topic or field (Chen, 2006; Compañó & Hullmann, 2002; Fujita et al., 2012; Huang et al., 2013; Klavans et al., 2013; Ponomarev et al., 2014). In addition, many of the current approaches require manual selection or curation of all data analysed (Chen, 2006; Klavans et al., 2012). Ponomarev et al. (2014) used variations of a single indicator, citation velocity, to predict highly cited papers while other groups made use of multiple indicators, full-text data and/or co-citation analysis to identify and characterize breakthrough publications in retrospective analyses (Chen, 2006, 2012; Klavans et al., 2012; Klavans et al., 2013). Other efforts focused on the development of analysis and visualization tools for quick visualization and assessment of potential turning points and breakthroughs (Boyack & Börner, 2003; Dunne et al., 2012).

Here, we aim to establish automated and semi-automated approaches to provide early indicators of published research with great potential. The goal is to provide program staff with a robust methodology that highlights pockets of breakthrough research, thereby enabling more informed program management. The methodology leverages an array of indicators to identify work that may contribute significantly to progress in its field. Here we describe work done to identify time-dependent and -independent publication indicators for differentiating breakthrough papers.

Data and Methods

Creating a Gold Standard Data Set

The first challenge in testing the importance of various publication features in predicting research breakthroughs is defining a core set of publications to be used as a gold standard. For our gold standard set of breakthroughs, we selected research articles from the following sources that highlight advances in cancer research:

1. The American Association of Cancer Research (AACR) publishes the AACR Cancer Progress Report annually (176 articles from the 2011-2014 reports).
2. The American Society of Clinical Oncology (ASCO) reports on key research in their annual Report, ASCO Clinical Cancer Advances. (58 articles from the 2009-2013 reports).
3. *Nature Medicine* 2011 special edition focused on advances in cancer research (74 articles spanning publication years 2008-2010).

Using these three sources we identified 287 distinct breakthrough publications that were indexed in the *Web of Science.* Table 1 shows the frequency by *Web of Science* Journal Subject Category. The inclusion of older publications (e.g., publication years of 2008 and 2009) enabled the curation of a dataset that included papers mature enough to have a range of breakthrough characteristics.

Table 1. Top 10 Web of Science Journal Subject Categories by Frequency for the Breakthrough Gold Standard Set (N=287).

Journal Subject Category	Count
Oncology	118
Medicine, General & Internal	109
Multidisciplinary Sciences	31
Cell Biology	17
Biochemistry & Molecular Biology	11
Public, Environmental & Occupational Health	7
Hematology	7
Genetics & Heredity	6
Immunology	6
Medicine, Research & Experimental	5

227 of the 287 breakthrough publications (81.7%) were published in journals in either the Oncology or Medicine, General & Internal *Web of Science* Journal Subject Categories.

Comparison Group Publication Set

We chose a comparison group of publications from a similar set of *Web of Science* Journal Subject Categories. We retrieved 647,879 publications from the 1) Oncology and 2) Medicine, General and Internal categories published between 2008 and 2014. We selected 2,500 publications at random from this dataset for use as the comparison group. We chose to select our control group by matching on the distribution of journal subject categories between the gold standard and comparison sets. However, we did not match the control group on publication year distribution due to the uneven publication year distribution resulting from the gold standard selection criteria.

Publication Indicators- bibliographic, citations, and altmetrics

We collected data from *Web of Science* to generate indicators for inclusion in our assessment. The majority of indicators were derived from the individual *Web of Science* citation records. These indicators were at the publication level (Table 2) and were collected in January 2015. While using a field-normalized Journal Impact Factor (JIF) would have been preferable, some publications in the gold standard set do not have JIFs determined for the publication journal, so we chose to use JIF best quartile as the best available alternative. Npayoffs reflects the inclusion of altmetrics gathered from *Web of Science* usage.

Table 2. Publication-level Indicators Considered For Inclusion in Random Forest Models.

Indicator level	Variable	Description
publication	TimesCitedTotal	total cites
	TimesNSCitedTotal	total cites (non-self)
	TimesCited2y	total cites in past 2 years
	TimeNSCited2y	total non-self cites in past 2 years
	NPages	total number of pages in an article
	NCitedRefs	number of references
	NAuthors	number of authors
	PubYear	publication year
	NCitedJSC	number of JSCs present in cited references
	NCountries	number of countries associated with publication authors
	NOrgs	number of institutions associated with publication authors
	CitVel6m	Citation velocity of specified time period (or maximum number of days since the article was published)
	CitVel1y	
	CitVel2y	
	CitVel5y	
	Bestquartile	Journal's best quartile from the 2013 Journal Citation Report
	DocumentTypeID	Describes publication type (article, review, etc.)
	Npayoffs	Total number of payoff events in Web of Science since January 2013 • A payoff event is when a WoS user downloaded the full-text article, added EndNote library, or saved for future use • Robot data filtered using multiple algorithms

Author-level indicators, person disambiguation

Some of the indicators in the study at the publication-level require a time lag after publication so we sought to increase the number of indicators that could identify potential breakthroughs immediately upon publication. Currently, these additional indicators are based on author publication history characteristics (Table 3). A critical aspect of author-based indicators is ensuring that each author's characteristics are correctly attributed. Therefore, we used a proprietary semi-automated algorithm to disambiguate authors and assign publications to each unique author.

Author-level indicators were assigned to each publication and computed in one of two ways: by averaging the indicator for all authors on a publication or by averaging the indicator for the top three authors on the paper as ranked by the indicator values.

Table 3. Author-level Indicators Considered for Inclusion in Random Forest Models.

Indicator level	Variable	Description
author	AvgNCoAuth	Number of distinct co-authors on all publications in the journal subject categories of oncology or general and internal medicine from 2008-2014
	AvgNCoAuth_Top3	
	AvgHindex	H-index based on all publications in the journal subject categories of oncology or general and internal medicine from 2008-2014
	AvgHindex_Top3	
	AvgPubHist	Total number of publications in the journal subject categories of oncology or general and internal medicine from 2008-2014 divided by six years
	AvgPubHist_Top3	
	NHighCitPubs	Highly cited publications defined by top 10% of publications in a particular year and journal subject category
	AvgNHighCitPubs	
	AvgNHighCitPubs_Top3	

Random Forest™ Model

We used the Random Forest™ machine learning algorithm (Brieman, 2001) as implemented by Liaw and Wiener (Liaw & Wiener, 2002) to assess the relative importance of each of the indicators listed above for differentiating breakthroughs from our comparison group. As Random Forest™ cannot handle null values; we were required to exclude all publications without citations and all publications where authors could not be disambiguated. This resulted in a final dataset of 223 breakthrough publications and 1,170 comparison publications.

The Random Forest™ algorithm is an example of a bagged decision tree algorithm (Breiman, 1996) that combines the classification results of some number N of individual decision trees. This set of N trees comprises the forest and is one of two input parameters that can be specified by the user. The other input parameter is an integer m which specifies the number of variables to consider when deciding how many variables to use for each node in the tree. Details on implementing this algorithm can be found in Liaw 2002 and references therein. As the random forest is built, a random subset of 2/3 of the data is used in the construction of each tree. The remaining 1/3 of the data is referred to as 'out-of-bag' (oob). For the analyses shown, the values $N = 500$ and $m = 4$ were found to minimize the out-of-bag error rate, which is a measure of the misclassification of the oob data by the random forest.

Results

We first examined the correlation among our publication indicators and removed the following indicators that were highly correlated: CitVel6m; CitVel2y; CitVel5y; TimesCitedTotal; TimesCited2y; AvgHindex_Top3; NHighCitedPubs_Top3. With the remaining set of indicators, we then ran the first Random Forest models using both the Mean

Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) to determine the relative importance of the indicators, as shown in Figure 1. The indicators with the highest relative importance are time-dependent (left of the dotted line). However, in order to best inform program management, it would be preferable to predict breakthroughs soon after publication, requiring indicators that can be calculated at, or near, the time of publication.

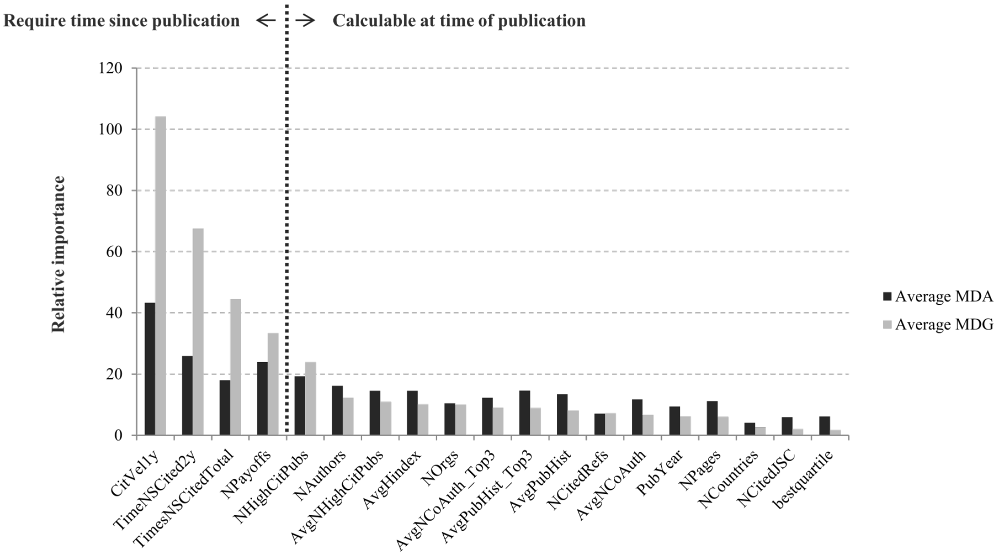

Figure 1. Relative Importance Ranking of Time-Independent and –Dependent Indicators based on Random Forest models (MDG and MDA). Out-of-bag error rate is 4.67%.

Because this work focuses on identification of publications with strong breakthrough potential near time of publication, we then considered only the time-independent indicators and produced new Random Forest models using these data. The relative importance ranking of the time-independent indicators are shown in Figure 2.

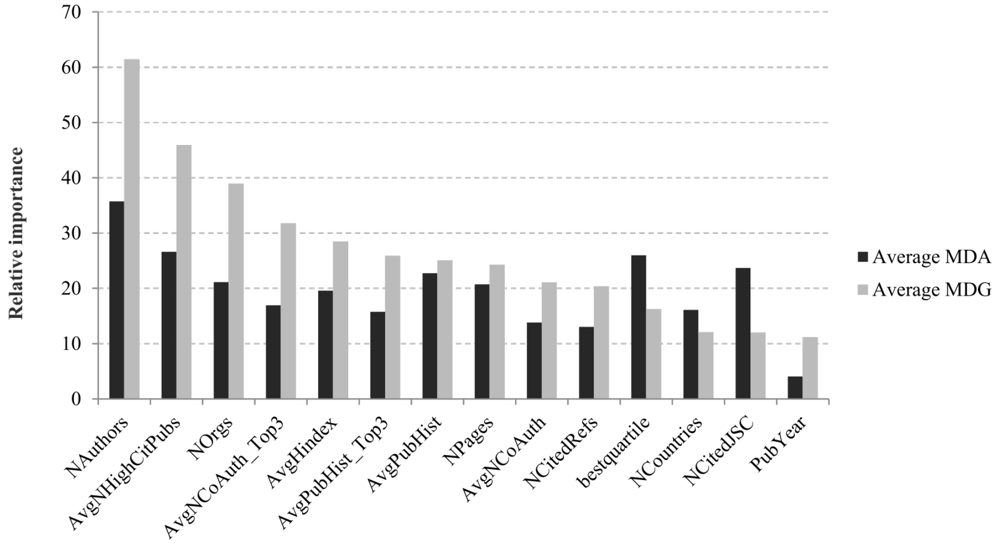

Figure 2. Relative Importance Ranking of Time-Independent Indicators based on Random Forest models (MDG and MDA). Out of bag error rate is 9.48%.

The highest ranked time-independent indicators, sorted by Average MDG, were: NAuthors, AvgNHighCitPubs, NOrgs, AvgNCoAuth_Top3, and AvgHindex. Sorting by Average MDA gives a slightly different set of top five variables: NAuthors, AvgNHighCitPubs, bestquartile, NCited Journal Subject Category (JSC), and AvgPubHist. While the first two variables are the same for either type of ranking, it would be interesting to explore the divergence of the other variables between the two rankings. The relative importance of these time-independent indicators is consistent with breakthrough work being associated with teams and researchers with a history of strong performance.

Conclusions and Next Steps

We have identified and ranked a set of time-dependent and -independent indicators for their importance in differentiating a set of breakthrough publications from a comparison group. Our results are early steps in developing tools for potentially identify promising emerging research in a timely manner. Our next steps include using a subset of these indicators to establish a multivariate model where the outcome is the estimated probability of being a breakthrough paper based on the existing training set. Using this model, we will prospectively identify candidate breakthroughs and share the results with program officers within NCI to assess the practical value of the model. Future work could include efforts to determine which indicators gain or lose predictive value over time through iterative evaluation of the relative strength and importance of each indicator.

Acknowledgments

This study was improved by contributions from Danielle Daee (NCI); Di Cross, Leo DiJoseph and Joshua Schnell (Thomson Reuters); and extends work by Ilya Ponomarev (formerly Thomson Reuters) and Charles Hackett (National Institutes of Allergy and Infectious Diseases). This work was supported in part by NIH contract #HHS263201000058B.

References

Boyack, K.W., & Börner, K. (2003). Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers, *Journal of the American Society for Information Science and Technology*, *54*, 447-461.

Breiman, L. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International Group.

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, *Journal of the American Society for information Science and Technology*, *57*, 359-377.

Compañó, R., & Hullmann, A. (2002). Forecasting the development of nanotechnology with the help of science and technology indicators, *Nanotechnology*, *13*, 243.

Dunne, C., Shneiderman, B., Gove, R., Klavans, J., & Dorr, B. (2012). Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization, *JASIST*, *63*, 2351-2369.

Fujita, K., Kajikawa, Y., Mori, J., & I. Sakata. (2012). Detecting Research Fronts Using Different Types of Combinational Citation, *Detecting Research Fronts Using Different Types of Combinational Citation*.

Huang, Y.H., Hsu, C.N., & Lerman, K. (2013). Identifying Transformative Scientific Research, *IEEE 13th International Conference on Data Mining* (ICDM), (pp. 291-300).

Klavans, R., Boyack, K.W., & Small, H. (2012). Indicators and precursors of "hot science", *17th International Conference on Science and Technology Indicators*, (pp. 475-487).

Klavans, R., Boyack, K.W., & Small, H. (2013). Identifying Emergent Opportunities in Science. Retrieved June 2, 2015 from: http://www.mapofscience.com/pdfs/EAGER_Final_v1.pdf

Liaw, A. & Wiener, M. (2002). Classification and Regression by Random Forest. *R News*, 2/3, (pp. 18-22).

National Science Board. (2007). Enhancing Support of Transformative Research at the National Science Foundation, *National Science Foundation*, (p. 14).

ODNI, (2011). IARPA Launches New Program to Enable the Rapid Discovery of Emerging Technical Capabilities.

Ponomarev, I.V., Lawton, B.K., Williams, D.E., & Schnell, J.D. (2014). Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction?, *Scientometrics*, *100*, 755-765.

Reardon, S. (2014). Text-mining offers clues to success: *Nature*, *509*, 410.