Predicting the Impact of Scientific Concepts using Full Text Features

Kathleen McKeown[1], Hal Daume III[2], Snigdha Chaturvedi[2], John Paparrizos[1], Kapil Thadani[1], Pablo Barrio[1], Or Biran[1], Suvarna Bothe[1], Michael Collins[1], Kenneth R. Fleischmann[3], Luis Gravano[1], Rahul Jha[4], Ben King[4], Kevin McInerney[5], Taesun Moon[6], Arvind Neelakantan[8], Diarmuid O'Seaghdha[7], Dragomir Radev[4], Clay Templeton[3], Simone Teufel[7]

[1]Columbia University, [2]University of Maryland, [3]University of Texas at Austin, [4]University of Michigan, [5]Rutgers University, [6]IBM, [7]Cambridge University, [8]University of Massachusetts at Amherst

New scientific concepts, interpreted broadly, are continuously introduced in the literature, but relatively few concepts have a long-term impact on society. The identification of such concepts is a challenging prediction task that would help multiple parties – including researchers and the general public – focus their attention within the vast scientific literature. In this paper we present a system that predicts the future impact of a scientific concept, represented as a technical term, based on the information available from recently published research articles. We analyze the usefulness of rich features derived from the full text of the articles through a variety of approaches, including rhetorical sentence analysis, information extraction, and time-series analysis. The results from two large-scale experiments with 3.8 million full-text articles and 48 million metadata records support the conclusion that full-text features are significantly more useful for prediction than metadata-only features, and that the most accurate predictions result from combining the metadata and full text features. Surprisingly, these results hold even when the metadata features are available for a much larger number of documents than are available for the full text features.

Introduction

Over a trillion US dollars are spent annually world-wide on research and development (Grueber & Studt, 2012). Unfortunately, only a small percentage of this amount is devoted to technologies that will have a high impact on society. In order to predict which research concepts hold the most promise, a framework to forecast whether a particular new finding will be accepted in future years is needed.

As a critical building block towards this ambitious goal, in this paper we present a system that predicts the scientific impact of research concepts — represented as technical terms — based on the information available from research articles in a reference period. For example, by examining scientific articles published between 1997 and 2003 related to the term *microRNA*, our system predicts that the term gains prominence in scientific articles published in the later years that we study (2004–2007). Thus, our approach predicts that *microRNA* will have scientific impact. In contrast, by examining scientific articles related to *rewiring* in the same time period, our system predicts that this term will not be prominent in scientific articles published in 2004–2007.

Unlike much previous work on citation prediction (see the Related Work section), we use the full text available in the articles and produce an analysis that identifies concepts, relations, citation sentiment, and the rhetorical function of sentences.[1] We complement these features with measures derived from the citation and author collaboration networks, and analyze the evolution of the features over time using a variety of principled time-series analysis methods. Finally, our system combines all features using logistic regression and computes an overall prominence score for the input technical term, to predict its impact in the literature. We define impact as a function of the relative growth of term appearance over unique documents (see the Experiments section for a detailed description).

To show the relative contribution of features drawn from the articles' full text in comparison to features drawn from the metadata, we present the results of a large-scale evaluation. Our first set of experiments, using a 3.8 million document dataset drawn from Elsevier publications, show that using text features alone enables significantly more accurate prediction of scientific impact than using metadata features alone. When the system uses both text and metadata features, prediction improves further.

We also compared the predictive ability of these sets of features on a much larger dataset that combines the Elsevier full text articles with 48 million metadata records from Thompson Reuter's Web of Science (WOS). The WOS data includes abstracts for each scientific article plus metadata such as title, authors, publication venue, year of publication, and citations. Our experiments address the question of whether a very large amount of metadata enables better prediction even without the text features, making them redundant. Experiments with this combined dataset show that the accuracy of metadata features alone increases with data volume, but still does not surpass the performance with text only. Our overall conclusion is that it is well worth the effort to obtain the full text of scientific articles and to exploit the power of natural language analysis.

In the remaining sections, we first present related work. We next give an overview of our system, followed by a description of the text features and the metadata features. We then turn to a description of our experiments and results. We conclude with a discussion of the implications of our work.

Related Work

Studying science is a science in and of itself. For example, the National Science Foundation has two programs designed to fund this type of research: Science, Technology, and Society (STS), which is primarily oriented toward qualitative research, and Science of Science and Innovation Policy (SciSIP), which is primarily oriented toward quantitative research. STS as a field of study has a long history. As the name indicates, STS utilizes social science and humanities approaches to understand the relationships among science, technology, and society. There is a wide range of STS approaches. For example, laboratory ethnography (Knorr-Cetina, 1999; Traweek, 1992) involves extended fieldwork within science and technology settings; in other words, observing and interviewing scientists and engineers in their native habitats. Actor-network theory (Latour, 1988) involves tracing the relationships among human actors and non-human actants. As such, technologies are seen as having some agency, or ability to shape the world. Another approach commonly used within the

[1] We tested less sophisticated lexical features such as n-grams in early experiments, but they did not show a significant impact on results and thus, we don't report on them here.

domain of science and technology policy is an expert panel, such as the Delphi method (Bornmann & Daniel, 2008). Such qualitative approaches are useful for learning about specific labs or subfields in rich detail, however they are not typically scalable. Thus, to automatically track scientific innovation in real time, quantative approaches are far more appropriate.

Scientometrics, or the measurement of science, has long been used to understand science at the macro scale as well as to make policy recommendations (Edge, 1979; Bornmann & Daniel, 2009; Schreiber, 2013). Since ranking algorithms based on scientometric data have demonstrated real potential to influence the direction of scientific progress (Beel & Gipp, 2009), it is of utmost importance for these algorithms to take into account as much information as possible to inform the resource allocation decisions of nations, institutions, and individual researchers (Lane & Bertuzzi, 2011; Lane, 2010).

Study of scientific impact spans almost a century, during which time expanding data sets and sophisticated tools have allowed for increasingly powerful results. Following several decades of small, expensive studies conducted for journal evaluation and acquisition, major citation indexing projects enabled the application of quantitative methods to the problems of research evaluation (Narin, 1976) and scientific prestige (Cole & Cole, 1967; Bayer & Folger, 1966). Since then, metrics such as the Journal Impact Factor (Garfield, 2006) that is primarily used to evaluate the impact of a journal, and, more recently, the h-Index (Hirsch, 2005) that is primarily used to evaluate the impact of a scientist, have been used. Scientometrics builds in part on the type of qualitative research described above, such as study of the function of citation (Moravcsik & Murugesan, 1975; Chubin & Moitra, 1975; Spiegel-Rösing, 1977) or the motivations for citation, often bringing these to bear in critique of the use of citations in research evaluation (Bornmann & Daniel, 2008). For example, citation counts include not only works that build on previous work but also works that negate the previous work or cite it perfunctorily (Ziman, 1968; Bonzi, 1982). Together these research streams comprise a large part of the quantitative science of science within the social sciences. Machine learning has introduced new horizons in the study of science (Losiewicz, Oard, & Kostoff, 2000) that continue to expand with increasing computational power and the availability of full text databases (Arbesman & Christakis, 2011).

An early paper by Garfield speculated on the relationship between citation data and future author performance (Garfield & Malin, 1968), and a few recent studies have attempted to predict future citations received by an author based on features of past work. These include studies of the predictive value of the h-index, which have played a role in the debates over that metric (Hirsch, 2007; Hönekopp & Khan, 2012), as well as attempts to predict changes in an author's h-index over time (Acuna, Allesina, & Kording, 2012; Penner, Petersen, Pan, & Fortunato, 2013; Dong, Johnson, & Chawla, 2014). Zhu, Turney, Lemire, and Vellino (2015) present a variant of h-index called the hip-index (influence primed h-index) based on datasets of papers and references that were influential for a paper and use it to predict fellows of an organization. All of these studies have tended to use simple feature sets, most often including citation based indicators of past performance, although social factors (Laurance, Useche, Laurance, & Bradshaw, 2013), social network properties (McCarty, Jawitz, Hopkins, & Goldman, 2013; Sarigöl, Pfitzner, Scholtes, Garas, & Schweitzer, 2014) and structural variation models representing impact on state of the art (Chen, 2012) have also been examined. Others (Ding, Yan, Frazho, & Caverlee, 2009) have experimented with weighted Pagerank algorithms to rank authors in author co-citation networks and a HITS framework (Wang et al., 2014) for simultaneous ranking of future impact of papers and authors.

Network-based approaches, building on research in social network analysis, have proven effective in helping to understand the structure of science (Birnholtz, Guha, Yuan, Gay, & Heller, 2013; Velden, Haque, & Lagoze, 2010; Velden & Lagoze, 2013). While our research builds on these approaches, the goal of this paper is to go beyond the typical network-based analyses that focus on nodes and edges and instead consider the content of the edges via natural language processing of full text.

Previous work has applied bibliometrics at the level of entities discovered in full text (Ding et al., 2013) as well as based on productivity, collaboration and influence (Havemann & Larsen, 2014). Additionally, topic proportions from Latent Dirichlet Allocation have been used to study the history of scientific ideas (Hall, Jurafsky, & Manning, 2008). To the best of our knowledge, our work is the first to predict term frequencies as proxies for the emergence of scientific concepts, and is novel in the sophistication of the full text features we bring to bear on the problem. While previous work that has used full text in prediction has relied on bag of words (e.g., (Yogatama et al., 2011; Yan, Tang, Liu, Shan, & Li, 2011; Boyack et al., 2011)), we base some of our analysis on larger units of texts (phrases) and on more linguistically motivated features such as rhetorical analysis.

Recently, there has been more work on the analysis of scientific articles that could ultimately be helpful for prediction of scientific impact (Tsai, Kundu, & Roth, 2013; Louis & Nenkova, 2013; Tan & Lee, 2014). For example, the 2003 KDD Cup (Gehrke, Ginsparg, & Kleinberg, 2003) included a citation prediction track. Since then, approaches to prediction have matured, and despite varying research designs, several classes of predictive variables have been established. These include citation data (Manjunatha, Sivaramakrishnan, Pandey, & Murthy, 2003), journal characteristics (Callaham, Wears, & Weber, 2002; Lokker, McKibbon, McKinlay, Wilczynski, & Haynes, 2008; Kulkarni, Busse, & Shams, 2007), author characteristics (Castillo, Donato, & Gionis, 2007), n-gram features drawn from abstracts and index terms (Ibáñez, Larrañaga, & Bielza, 2009; Fu & Aliferis, 2008), download statistics (Brody, Harnad, & Carr, 2006), and social media mentions (Eysenbach, 2011). Fu and Aliferis unified much of the early work in this area, reporting evidence that author metrics improved the scores obtained by modeling journal characteristics alone, and that adding metadata features improved scores still further (Fu & Aliferis, 2008). More recent NLP research has yielded mixed results on n-gram and topic features drawn from full text (Yogatama et al., 2011; Yan et al., 2011), and the usefulness of full text in citation prediction for papers remains an open question.

Much research has focused on particular disciplines and sub-disciplines of science. For example, scientometric approaches have been applied to computer science (Guha, Steinhardt, Ahmed, & Lagoze, 2013) as well as its subfields, such as human-computer interaction (Bartneck & Hu, 2009) and computer-supported cooperative work (Horn, Finholt, Birnholtz, Motwani, & Jayaraman, 2004). Since the goal of this paper is to help predict innovation across various fields of science and engineering, we build on this earlier work, but cannot rely solely upon metrics that have proven to be effective within any one field.

System Architecture

Our system predicts the impact of a scientific concept, represented as a technical term, using features derived from the full text of scientific articles as well as more traditional features derived from the metadata of the documents. The technical terms used as input refer to specific scientific concepts and are assumed to have no synonyms.

The system is designed as a three-staged pipeline. Given an input term, our system first computes the set of documents relevant to the term by determining when there is an exact match

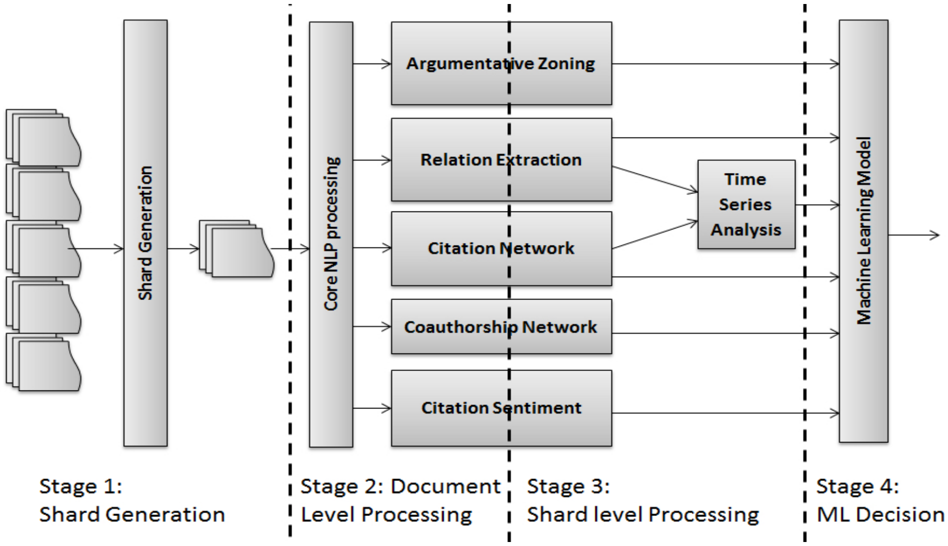

Figure 1. System Architecture.

between the term and the words of either the title or the abstract. As shown in Figure 1, this first stage, called *shard generation* produces a set of relevant documents that we call the *shard*.

In Stage 2, we process each document in the shard using our core NLP pipline, which produces annotations representing sentence segmentation, part-of-speech (POS) tagging and parsing of citation sentences. We then annotate each document with the rhetorical function of each sentence using *argumentative zones* (Teufel, 2010), entities and relations expressed in the text, and sentiment toward citations.

In Stage 3, we compute aggregate values for these annotations across the shards and build a coauthorship network and a citation network for the documents in the shard. We also generate a time series for each feature over the years in the reference period, and produce additional features from various functions applied to the time series.

Finally, in Stage 4, our machine learning modules uses the features to predict the scientific impact in the forecast period.

Metadata Features

Our system uses the metadata available for each paper in Web of Science[2] to compute some simple features and other more complex features based on networks. For the simple features, which we call *acceptance* features, we consider the number of unique papers, authors and their countries, institutions, conferences, journals, and books. Additionally, we compute the mean number of authors per paper, the number of papers with two or more authors, the number of papers with authors affiliated with multiple institutions, and the number of papers with authors from different institutions. For the network-based features, network theory (Newman, 2010) provides a number of tools to model aggregate information in relational data. Several recent papers have focused on applying network techniques to analyze bibliometric data (Batagelj & Cerinšek, 2013; Viana, Amancio, & Costa, 2013; Fu, Song, & Chiu, 2013; Pan, Kaski, & Fortunato, 2012). The use of networks to model bibliometric data such as collaborations between authors and citations between papers

[2] http://thomsonreuters.com/thomson-reuters-web-of-science/

is based on the view of science as a social process (Sun, Kaur, Milojević, Flammini, & Menczer, 2013). We derive network features from two kinds of networks, citation networks and author collaboration networks. In an author collaboration network, nodes represent authors and undirected edges represent the fact that two authors co-authored at least one paper[3]. In a citation network, nodes represent documents and directed edges record that one document cites the other (only within the shard). Citation links between papers indicate topical similarity (Small, 1973; Kessler, 1965). Many dense clusters in a citation network may represent fragmented communities of research where documents position themselves relative to papers in the same cluster and do not frequently cite other papers in the area. Similarly, a low clustering coefficient (meaning that documents don't tend to cite their cited documents' cited work) may indicate that a field tends to make large, disruptive advances (Funk & Owen-Smith, 2012), rather than incremental improvements.

In contrast, collaborations give us a more direct probe into the social dynamics of research on a given topic, e.g., dense clusters in this network represent close-knit communities that exist among the authors in a field. Similarly, an author with high betweenness centrality may act as a bridge between two different communities that do not frequently collaborate.

Given a shard, these networks can be built efficiently using our metadata database. The citation network is built by querying a database table that contains resolved citations between papers;[4] the author collaboration network is built by querying a table containing authors of each paper.

Full Text Features

Our full text features are computed based on aggregates of information extracted from the text of each article: entities and relations, argumentative zoning, and citation sentiment. Time series are then computed over aggregates of these features.

Entities and Relations

We identify two types of textual information, namely, *entities* and *relations*. The information that we extract enables a more refined analysis of crucial aspects around a given topic than would be possible using the original unannotated text. For example, we can extract the number of algorithms that have been implemented for a given input problem, and use it as evidence of the depth in which this problem has been studied. Similarly, we can gauge the interest in a research topic based on the diversity of funding agencies involved in the topic. Entities (e.g., focus, techniques, and domains (Gupta & Manning, 2010)) and relations (e.g., protein–protein interaction (Bui, Katrenko, & Sloot, 2011)) involving them have been extracted from scientific articles, although to the best of our knowledge, they have not been used in scientific prominence prediction systems.

Entities: The entity detection module produces annotations consisting of a entity *type* (e.g., algorithm, data set, gene, virus, protein, database) and a mention (e.g., CRF, an instance of algorithm; BRCA1, an instance of gene). We recognize a total of 15 entity types. Some of the entity types are general to all domains (e.g., method, problem, theory) and others are specific to the most frequently occurring family of domains in the corpus (i.e., medical, genomic, biology). We define the *primary type* as the entity type corresponding to the queried term if it matches one of our 15 entity types. Otherwise, it is the entity type with the highest document frequency in the shard. We can now measure how cohesive a shard is by using the proportion of articles containing a mention of the primary

[3] We used the author resolution results produced by (Wick, Kobren, & McCallum, 2013).

[4] Resolving a citation is the process of using the bibliographic text to locate the cited paper in the database.

Basic
number of nodes
number of edges
number of weakly-connected components
size of largest weakly-connected component
Clustering
average Watts-Strogatz clustering coefficient (Watts & Strogatz, 1998)
average Newman clustering coefficient (Newman, 2010)
Centrality
average degree
average closeness centrality (Freeman, 1978)
average betweenness centrality (Freeman, 1977)
Distances
diameter
average shortest path
Degree distribution
degree assorativity (Newman, 2003)
in-/out-/total-degree power law exponent (Newman, 2010)
in-/out-/total-degree Newman power law exponent (Newman, 2010)
in-/out-/total-degree power law R^2 (Newman, 2010)

Table 1

A list of features computed for each network.

entity type in the shard. We can also measure how diverse it is by counting the number of distinct mentions of the primary entity type in the shard.

If the term is an entity, we also compute as features the frequency and corresponding rank of the term with respect to other entities of the same type—both absolute and normalized—and the ratio between the frequency of the input term and the most frequent entity of its same type.

To annotate the entities, we use a dictionary-based tagger (Neelakantan & Collins, 2014). Dictionaries are compiled for every named entity type using large amounts of unlabeled data and a small number of labeled examples. For every named entity type, first we construct a high recall, low precision list of candidate phrases by applying simple rules on the unlabeled data collection. Using Canonical Correlation Analysis (CCA) (Hotelling, 1936), we represent each candidate phrase in a low-dimensional, real valued space. Finally, we learn a binary SVM (Joachims, 1998) in the low-dimensional space with few labeled examples to classify the candidate phrases. We filter out the noisy phrases from the high recall, low precision list of candidate phrases using the learned SVM to get a high recall, high precision dictionary.

Relations: Table 2 lists the relations that we extract. For the Funding relation, we produce

Funding ⟨grant, funding agency⟩
Novelty Claims (an article claims novelty over something, e.g., we are the first ones to apply technique X to problem Y)
Dataset Purpose (an article *proposes* a new dataset or *uses* an existing one)

Table 2

Relations extracted by the system

frequency- and average-based features indicating the number of funding agencies and the number of grants in each article. In addition, we produce Boolean features indicating whether there are multiple grants or institutions supporting the research reported by articles in the shard. For the other relations, we extract all the mentions of each type in an article and then produce numeric features indicating their frequency and average in the shard. To annotate the relations, we use two different methods. For funding information, we can, in some cases, retrieve it directly from the article metadata. However, in most cases, especially for older articles, we can only obtain this information from text, as follows. We first use string matching to locate the acknowledgment section of the article in question, where the funding information for the article usually resides. Then, we use two supervised CRF models (Lafferty, McCallum, & Pereira, 2001) to identify the funding agencies and grant numbers. Finally, we build ⟨funding agency, grant⟩ pairs by combining the agencies and grants that co-exist in a sentence in order of appearance. To annotate the remaining relations, we use a supervised sentence classification approach (Bach & Badaskar, 2007). Since only a few of the sentences in an article will likely include mention of these relations, for efficiency we only classify the sentences that mention at least one of the 10 most relevant terms according to their weight in the SVM classification model. In our experiments, we used an SVM-based classifier trained on stemmed terms along with their respective POS tags as features, from a manually annotated dataset. The accuracy of our classifiers range from 0.72 F1 measure for novelty claim relation to 0.89 for funding relation.

Argumentative Zoning

The argumentative zoning (AZ) component marks up each sentence in a scientific document according to its rhetorical function. We expect that an entity's prominence in the scientific community is reflected in the way scientists write about it, e.g., whether the entity is presented as a novel contribution (AZ category **Own Work**) or a well-established concept in the literature (AZ category **Background**). The relevance of such rhetorical categories comes from the hypothesis that the first occurrence of new ideas should be in some paper's goal statement (Myers, 1992). However, as the idea emerges and gets accepted, it is mentioned in other areas of papers referring to the original idea – thereby "travelling" through other rhetorical categories. When the new idea is competing against other existing ideas, it will occur in contrast and comparison statements (MacRoberts & MacRoberts, 1984). If it comes to be adopted by other researchers in the field, it will be mentioned as the basis for their work, indicating a different phase of acceptance (or a different status of the cited idea). If the concept becomes widely accepted, it will be found with increasing frequency in rhetorically neutral sentences and eventually even in background sections (Swales, 1990). These ideas are formalized in the "argumentative zoning" theory of Teufel (Teufel, 2010), whereby the text of an article is partitioned into zones defined by their rhetorical function.

The core functionality of the AZ component in our system is automatically labeling each sentence in an article with a category specifying the rhetorical status of that sentence. We use six categories: **Aim**, **Own Work**, **Background**, **Contrast**, **Basis** and **Other**; for more details on these categories, see (Teufel, 2010). The document-level AZ system takes a document as input and labels every sentence with one of the six categories listed above, using a Maximum Entropy Markov Model classifier suitable for sequential labeling. The features extracted for each sentence include internal information about the words, n-grams and citations it contains as well as external information about its absolute and relative position in the document, the section in which it appears, and whether a string from an extensive pattern lexicon matched. This system has been trained using a manually annotated set of documents from the computer science and chemistry domains. Using cross-validation on the chemistry subset of the data, the system's accuracy has been measured at 75%.

To produce AZ indicator values for a concept term, we aggregate over the AZ labels of all sentences that contain a mention of the term. The aggregate indicators we produce are the absolute count totals of each AZ label in the set and the relative count proportions of each AZ label in the set, i.e., 12 indicators in total.

Citation Sentiment

The Citation Sentiment component labels each sentence containing a citation as expressing positive, negative or objective sentiment towards the cited entity. It implements the hypothesis that emerging ideas will initially be cited in the context of strong opinions, whether these are negative or positive ones (Small, 2011). We also hypothesize that the more an idea is accepted in a scientific community, the more it will be presented as an "objective fact". As might be expected, most citations in scientific articles are objective in terms of sentiment (86% of sentences in the annotated corpus described below); this may be an indication that positive or negative citations are somewhat rare, and may be important.

Similarly to the AZ component, the citation sentiment module first assigns sentence-level labels and later aggregates over them to produce feature values for the entity of interest. The sentence-level classifier, based on Athar (Athar, 2011), is a Support Vector Machine that takes n-gram features and basic negation features as input and outputs one of three sentiment labels: Positive, Negative or Objective. It was trained on Athar's corpus of 8,736 hand-labeled citation sentences. The entity-level feature values are then calculated as total and proportional counts of these labels over a set of sentences that are relevant to the entity of interest. Because citation sentiment is by definition only meaningful in the presence of citations, we aggregate over all sentences that contain the term and also contain a citation. The performance of the citation sentiment component is 0.6 F-measure (macro).

Aggregation and Time Series

All components described so far produce features as aggregated statistics over the full time-window under consideration. The Time-Series Analysis (TSA) component, in contrast, computes features that can capture the temporal variation of such statistics.

For every feature given as input, TSA computes a time-series sequence that represents its aggregated values per year, instead of its aggregated value for the full time period. In order to capture how these characteristics grow and fade over time, we model time series using six growth functions: Linear, Quadratic, Logistic, Exponential, Gompertz, and Richards. For the Linear and Quadratic functions we use linear least-squares estimates; for the other functions we use non-linear least-squares estimates of their parameters. Once all functions have been fitted to a time series, we select the function with the smallest Akaike Information Criterion (AIC) (Akaike, 1974) value as the best.[5] We use the name of the best-fitted function, as well as its slope, as features for our ML component. We also use as features the coefficients of the first and second degree terms of the Linear and Quadratic functions, respectively, with which we can determine the trend and its rate of change.

In addition to these model-based features, we also consider a variety of statistical measures from the literature to capture global characteristics of time series and detect interesting patterns. In particular, we use nine such characteristics and compute them as proposed by (Wang, Smith, & Hyndman, 2006). Briefly, *Seasonality*, *Periodicity*, and *Trend* are features that attempt to detect cycles, the period of those cycles, and the strength of the long-term trend of a time series. *Skewness* measures the degree of asymmetry of data points of a time series around their mean and *Kurtosis* measures the peakness and flatness of data points, relative to a normal distribution. *Serial correlation* measures how noisy a time series is by fitting a white noise model, and is defined as the Box-Pierce Statistic (Box & Cox, 1964). *Non-linearity* measures the non-linearity structure of time series data, from which we can determine if linear or non-linear models can better forecast the data (Teräsvirta, Lin, & Granger, 1993). *Self-similarity*, which relates to the autocorrelation statistic, measures the long-range dependence of a time series; we compute this feature as the Hurst Exponent (Willinger, Paxson, & Taqqu, 1998). Finally, we use the Lyapunov Exponent that measures the chaotic behavior of a time series; it detects the degree of randomness and the possibility of accurately predicting the near future (Hilborn, 2000).

Experimental Evaluation

In this section we describe the methodology, settings, and findings of our experimental evaluation.

Dataset

Our dataset includes 3.8 million full text articles published by Elsevier as well as 48 million metadata records from Web of Science (WOS).[6] The metadata includes titles, author names and institutions, in some cases funding, citations with the IDs of cited papers, and abstracts. The full text of the Elsevier articles was parsed into a common XML representation that identifies not only metadata, but in many cases also provides structural markup for the text, e.g. identifying tables, sections, and paragraphs, and linking in-text citations to the corresponding bibliography entries.

Methodology

The system was developed as part of a government funded program to predict the scientific impact of entities such as terms in some future forecast period F given some observations in the reference period R where $R < F$. Scientific impact is quantified by the program in the form of ground truth functions (GTF) which concentrate on relative growth of term appearance in unique documents over a baseline count as opposed to absolute growth. Previous work has often looked

[5] We experimented with other measures, including BIC and Chi-square as measures to estimate the quality of each model under consideration. We did not observe significant differences in the selection of each model (i.e., for the majority of our experiments these measures were in agreement). In general, AIC penalizes less strongly the number of parameters in comparison to BIC, and previous research (Burnham & Anderson, 2002, 2004) argues that AIC has several theoretical and practical advantages over BIC.

[6] This dataset was provided by the government sponsor to all teams who were part of the funded program.

Model	R^2	τ
Linear regression	-0.025	0.322
Regression tree	0.160	0.339
Random forest	0.200	0.345
Gradient boosted dec. tree	0.235	**0.372**
Support vector regression	0.253	0.355
Logistic regression	**0.263**	**0.372**

Table 3

Performance per regression model on held-out development set using metadata features only

at absolute growth of counts such as citations (Yogatama et al., 2011). The underlying motivation of GTFs is to temper variance in count quantity across disciplines (e.g. biology tends to have more publications than pure math) and time (i.e. absolute publication counts increase from past to present). Formally, the ground truth function (GTF) for a term e is defined in terms of document counts for e for R or F:

1. $r(e)$: exponentially weighted average of counts of unique TS-documents containing e for the years leading up to and including R, where the interval used for averaging is the size of the forecast gap and where counts in recent years are weighted more heavily.

2. $f(e)$: exponentially weighted average of counts of unique TS-documents containing e for the years up to and including F, where the interval used for averaging is the size of the forecast gap and where counts in recent years are weighted more heavily.

TS-documents are documents drawn from three trusted sources, *Science*, *Nature*, and the *Proceedings of the National Academy of Sciences (PNAS)*. When $f(e) < \max(1, r(e))$ the GTF is defined to be zero, otherwise it produces values in the range from 0 to 1 and it is computed as follows:

$$\text{GTF}(e, r, f) = \left(1 - \frac{r(e)}{f(e)}\right)\left(1 - \frac{1}{f(e)}\right) \tag{1}$$

The goal of the system is to predict the GTF(e, r, f), having observed e and its derived features in the dataset up to reference period R. In sum, the goal is to predict the ground truth function of e at F (i.e., the relative increase in counts of unique TS-documents in which e appears) having observed e up to R.

Since most papers receive no citations or a very small number of citations, the distribution of GTF values for our datasets tends towards an exponential distribution as the distance between R and F increases.

The models

GTFs for terms are $\in [0, 1]$ and thus, it is possible to model the desired prediction using vanilla logistic regression.[7] Though logistic regression is typically used in the literature for classification and the output is defined in the interval $\{0, 1\}$, it can be directly applied to regression tasks where the output range is $[0, 1]$ by defining the objective function in terms of minimizing the KL divergence between the GTF and the hypothesis.

In addition to logistic regression, the following other standard regression models were considered for the task of modeling the GTF: linear regression, regression trees, random forests, gradient

[7] An alternative would be to train a model to predict the cumulative counts $r(e)$ and $f(e)$, from which the GTF can be calculated. We adopted our current approach after preliminary experiments on development data.

Term	2004	2005	2006	2007
dopamine signaling	0.250	0.062	0.208	0.249
lower ros	0.000	0.000	0.000	0.000
rewiring	0.222	0.000	0.000	0.145
wd40	0.000	0.250	0.585	0.629
microrna	0.547	0.857	0.863	0.905
cell self-renewal	0.188	0.393	0.332	0.330
plant homeodomain	0.000	0.000	0.492	0.718

Table 4

Example GTF values for four forecast periods

boosted decision trees, and support vector regression. Using the metadata features only, the models were trained on a set of documents with GTFs defined for the period $R = 2003$ and $F = 2007$. They were then evaluated on a held-out data set over the same period in terms of R^2 and Kendall's τ. The results in Table 3 show logistic regression outperforming all other models in R^2 and tying in τ. Thus, we chose logistic regression as the model for the system.

Experiments

We conducted experiments to compare systems that use only text-based features with systems that use more traditional metadata features, as well as systems that use both on a dataset of scientific documents published within 1991–2007. We ran each system to forecast scientific impact in four different scenarios varying the forecasting period from one year past the reference period (chosen as 2003) to four years past the reference period, i.e., 2004 to 2007. Finally, each of the above settings was evaluated over a dataset that contained all 48 million documents in the Web of Science metadata records as well as the subset of Elsevier-published documents for which the full text of the document was available. In total, this yields 24 experimental configurations: 3 systems to predict scientific impact on 4 forecast years for 2 datasets.

Our experiments were conducted on 5923 terms from a list provided by the evaluators for our funding agency. A term is an n-gram from one to four words; the term population is drawn from abstracts and titles of documents published within the trusted sources (*Nature*, *Science* and *PNAS*) in the time period from 1991 to 2007. Terms were filtered using a common stop word list, low frequency terms[8] and common scientific terms. Some examples are provided in Table 4 along with the GTF value defined by Equation (1) in the Methodology section. We selected terms using the following methodogy. We tallied all documents that contain the term in its title or abstract and retained terms for which at least 10% of the computed documents came from the Elsevier collection and therefore had full text. This is the same method used to compute shards and thus, we knew that the shards used for prediction would not be empty when restricted only to the Elsevier collection. We used five-fold cross validation, with 90% of the data in each fold used for training and the rest used for testing. We compare our system results to the gold standard GTF values using Pearson correlation r and Spearman rank correlation ρ.

For analysis, we categorized the features as metadata or full text features. Earlier experiments on development data showed that time series analysis over argumentative zoning, sentiment and co-

[8] Since we are evaluating impact within our corpus, if a term has low frequency, then it never emerges within the time frame of the corpus. It is possible that it emerges years later, but we will never be able to evaluate whether we can pick that up.

authorship was not helpful; given that time series analysis is computationally expensive, we did not include these features in the evaluation.

Experimental results and discussion

The charts below graphically illustrate how our predictions correlated with the ground truth over the four forecast years. We also show numeric results in Table 5 for a representative forecast year, reference year 2003 and forecast 2007.

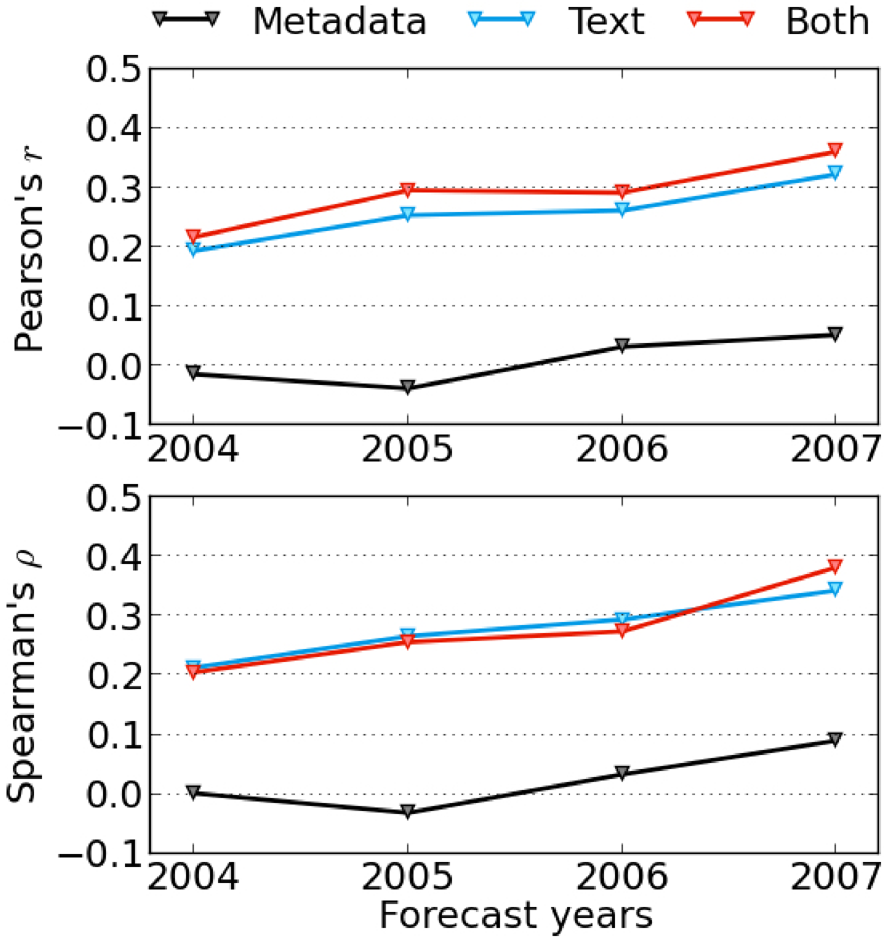

Figure 2. Evaluation on Elsevier-only data

Figure 2 shows the results for experiments carried out on full text drawn from Elsevier; the top graph shows Pearson r and the bottom one Spearman ρ. Here metadata features underperform text-based features by a substantial margin as measured by ρ and thus, the benefit of full text in comparison to metadata is clear. Adding text-based features to metadata-only features also gives substantially improved results. Our results show that the combination of full text and meta-data performs the best, outperforming the text indicators only by a slight margin, as indicated by r. These results indicate that the metadata features do add value.

Figure 3 shows the results for experiments carried out on the full dataset, including both Elsevier and WOS records. In this case, the shard, which includes all documents relevant to the term, is substantially larger. We might expect metadata features to outperform text-based features

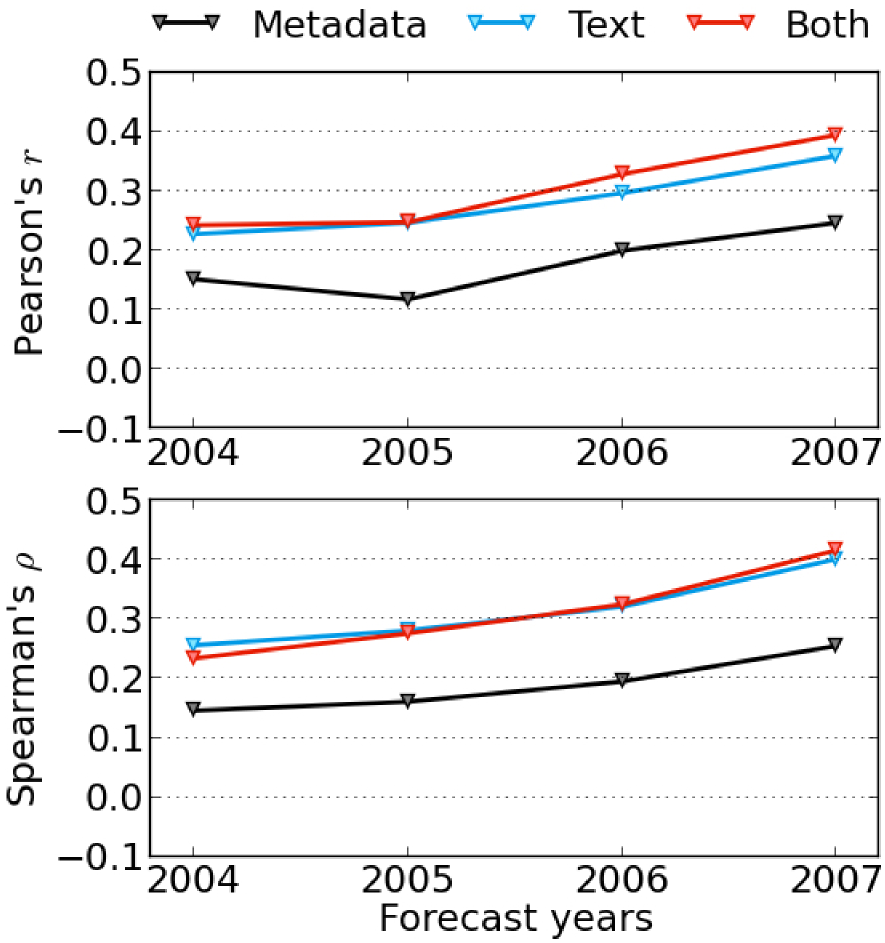

Figure 3. Evaluation on Elsevier + WOS data

since the citation and coauthorship networks that are built can be more comprehensive; more articles corresponding to the citations will be found in the dataset. Furthermore, the metadata acceptance features will be drawn from all articles, while the text features will only be drawn from a subset. We do see a substantial improvement in metadata alone, but the results still do not surpass those of the text-based and the full set of features. Under Spearman ρ, both the system based on text-based features and the system based on combined text-based and metadata perform significantly better ($p < 0.05$ using the paired permutation test) than metadata only across all forecast years except 2005. Note that the system using text-based features also improves, because the larger dataset contains abstracts and text features are extracted from these. The text-only system and the system using a combination of text and metadata indicators are similar in performance, with the combination of features usually slightly outperforming the text-only.

In Table 6 we describe an ablation study of the system for 2006, the first year where combined text and metadata features outperform text-only features according to Spearman ρ. We show the performance using individual indicators in isolation, sorted by Pearson r. Text indicators are shown in bold. The top performing indicators are times series over entities, acceptance and relations, of which only acceptance is derived from metadata. Network indicators perform at the bottom of the

Indicators	Dataset	r	ρ
All indicators	Elsevier	**0.364**	**0.392**
Text only	Elsevier	0.346	0.373
Metadata only	Elsevier	0.194	0.193
All indicators	Complete	**0.393**	**0.428**
Text only	Complete	0.365	0.407
Metadata only	Complete	0.316	0.340

Table 5

Evaluation for forecast year 2007

Indicators	r	ρ
TimeSeriesAnalysis:entities	0.301	0.317
TimeSeriesAnalysis:acceptance	0.293	0.313
TimeSeriesAnalysis:relations	0.191	0.195
Acceptance	0.19	0.214
Argumentative Zoning	0.188	0.215
Citation network	0.147	0.193
TimeSeries:networks	0.131	0.148
Coauthorship network	0.123	0.164
Citation sentiment	0.0679	0.078

Table 6

Ablation tests for forecast year 2006

times series and near the bottom of the regular indicators. Argumentative zoning, a text indicator that reflects the rhetorical structure of the article, performs near the top of individual indicators. We see two unexpected results: 1) acceptance, which is a metadata indicator, performs well, both in times series and without, and 2) citation sentiment performs poorly. Acceptance simply counts the number of venues, authors, and institutions in the shard of relevant documents, with the rationale being that the more places and authors that have published on this topic, the more impact it has had. Other than these two exceptions, the individual results support our overall results showing that text indicators tend to perform better.

We see that time-series over entities has a much greater impact than other indicators. Over time we expect the shards centered around prominent entities to be more cohesive and less diverse. We hypothesize that cohesiveness increases with the number of mentions of the prominent entity type, while diversity decreases because there are fewer comparisons to other entities of the same type. This occurs precisely because people accept that the prominent entity is important. For example, consider a gene that is in the process of being mapped. We would have discussions of other related genes early on and later on, when that gene becomes more important, it would appear more in the context of the disease it is important for and the drug that it reveals should be used, as opposed to discussion of other genes. As time goes on, we would also see more documents that mention the gene.

In addition to the indicator-level ablations, we also looked at individual feature performance using their odds ratios. While the ablations show the overall contribution of an indicator (which combines multiple features), all indicators contain important individual features. For example, although TimeSeries:networks is not among the highest performing indicators, some of its member features (e.g., the slope of the growth function best fitted to the article citation count) are among the best overall. Similarly, the total counts of the AIM and OWN categories from Argumentative Zoning, among others, are some of the most powerful features.

Conclusion

Our results show the clear benefit of text features over metadata. When prediction is performed on a dataset including only full text articles, a system that makes use of features drawn from full text performs significantly better than a system that only uses metadata features. The addition of all data in WOS does yield an improved performance of metadata features, both in the metadata only performance and in the full feature performance. Nonetheless, across all metrics, the text features are so strong that even in this scenario, where metadata features are computed over all documents relevant to a term while text features are computed over only a subset of the relevant documents, the model based on metadata alone cannot outperform text features. We conclude that the benefit of analysis of the full text of scientific articles is well worth the increased performance cost of the natural language analysis.

Acknowledgements

We thank the anonymous reviewers for their suggestions on the paper. This work was supported by the Intelligence Ad- vanced Research Projects Activity (IARPA) via Department of Interior National Business Cen- ter (DoI/NBC) contract number D11PC20153. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Future impact: Predicting scientific success. *Nature*, *489*(7415), 201–202.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Arbesman, S., & Christakis, N. A. (2011). Eurekometrics: Analyzing the nature of discovery. *PLoS Computational Biology*, *7*(6), e1002072.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the acl-11 student session.* Portland, OR.

Bach, N., & Badaskar, S. (2007). *A survey of relation extraction.* (Available at `http://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf`)

Bartneck, C., & Hu, J. (2009). Scientometric analysis of the chi proceedings. In *Proceedings of the sigchi conference on human factors in computing systems.*

Batagelj, V., & Cerinšek, M. (2013). On bibliographic networks. *Scientometrics*, *96*(3), 845–864.

Bayer, A. E., & Folger, J. (1966). Some correlates of a citation measure of productivity in science. *Sociology of Education*, *39*(4), 381–390.

Beel, J., & Gipp, B. (2009). Google Scholar's ranking algorithm: The impact of citation counts (An empirical study). In *Proceedings of the 3rd international conference on research challenges in information science (rcis-09)* (pp. 439–446). Fez, Morocco.

Birnholtz, J., Guha, S., Yuan, Y., Gay, G., & Heller, C. (2013). Cross-campus collaboration: A sceintometric and network case study of publication activity across two campuses of a single insitution. *Journal of the American Society for Information Science and Technology*, *64*(1), 162-172.

Bonzi, S. (1982). Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science*, *33*(4), 208–216.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80.

Bornmann, L., & Daniel, H.-D. (2009). The state of h index research. is the h index the ideal way to measure research performance? *EMBO reports*, *10*(1), 2.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B (Methodological)*, *26*(2), 211–252.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., … Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE*, *6*(3), e18029.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, *57*(8), 1060–1072.

Bui, Q.-C., Katrenko, S., & Sloot, P. M. (2011). A hybrid approach to extract protein–protein interactions. *Bioinformatics*, *27*(2), 259–265.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, *33*(2), 261–304.

Callaham, M., Wears, R. L., & Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *Journal of the American Medical Association*, *287*(21), 2847–2850.

Castillo, C., Donato, D., & Gionis, A. (2007). Estimating number of citations using author reputation. In *Proceedings of the 14th international symposium on string processing and information retrieval (spire-07)* (pp. 107–117). Santiago, Chile.

Chen, C. (2012, March). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, *63*(3), 431–449. Retrieved 2013-06-06, from `http://doi.wiley.com/10.1002/asi.21694` doi: 10.1002/asi.21694

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, *5*(4), 423–441.

Cole, S., & Cole, J. R. (1967). Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review*, *32*(3), 391–403.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. *PloS ONE*, *8*(8), e71416.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009, November). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, *60*(11), 2229–2243. Retrieved 2013-06-06, from `http://doi.wiley.com/10.1002/asi.21171` doi: 10.1002/asi.21171

Dong, Y., Johnson, R. A., & Chawla, N. V. (2014, December). Will This Paper Increase Your h-index? Scientific Impact Prediction. *ArXiv e-prints*.

Edge, D. (1979). Quantitative measures of communication in science: A critical review. *History of Science*, *17*(36), 102–134.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, *13*(4), e123.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, *40*(1), 35–41. Retrieved from `http://links.jstor.org/sici?sici=0038-0431\%28197703\%2940\%3A1\%3C35\%3AASOMOC\%3E2.0.CO\%3B2-H`

Freeman, L. C. (1978). Centrality in social networks: Conceptual clarification. *Social Networks*, *3*(1), 215–239.

Fu, L. D., & Aliferis, C. (2008). Models for predicting and explaining citation count of biomedical articles. In *Proceedings of the AMIA annual symposium* (pp. 222–226). Washington, DC.

Fu, T. Z. J., Song, Q., & Chiu, D. M. (2013). *The academic social network.* (`http://arxiv.org/abs/1306.4623`)

Funk, R., & Owen-Smith, J. (2012). *A dynamic network approach to breakthrough innovation.* (`http://arxiv.org/abs/1212.3559`)

Garfield, E. (2006). Citation indexes for science: A new dimension in documentation through association of ideas. *International Journal of Epidemiology*, *35*(5), 1123–1127.

Garfield, E., & Malin, M. V. (1968). Can Nobel Prize winners be predicted? In *Proceedings of the 135th meeting of the american association for the advancement of science.* Dallas, TX.

Gehrke, J., Ginsparg, P., & Kleinberg, J. (2003). Overview of the 2003 KDD Cup. *ACM SIGKDD Explorations Newsletter*, *5*(2), 149–151.

Grueber, M., & Studt, T. (2012). Global R&D funding forecast. *R&D Magazine*, *16*, 3–35.

Guha, S., Steinhardt, S., Ahmed, S., & Lagoze, C. (2013). Following bibliometric footprints: The acm digital library and the evolution of computer science. In *Proceedings of the 13th annual acm/ieee-cs joint conference on digital libraries.*

Gupta, S., & Manning, C. D. (2010). Identifying focus, techniques and domain of scientific papers. In *Proceedings of the nips-10 workshop on computational social science and the wisdom of crowds.* Whistler, Canada.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing (emnlp-08)* (pp. 363–371). Honolulu, HI.

Havemann, F., & Larsen, B. (2014, April). Bibliometric Indicators of Young Authors in Astrophysics: Can Later Stars be Predicted? *ArXiv e-prints*.

Hilborn, R. C. (2000). *Chaos and nonlinear dynamics: An introduction for scientists and engineers*. Oxford, UK: Oxford University Press.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569.

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, *104*(49), 19193–19198.

Hönekopp, J., & Khan, J. (2012). Future publication success in science is better predicted by traditional measures than by the h index. *Scientometrics*, *90*(3), 843–853.

Horn, D., Finholt, T., Birnholtz, J., Motwani, D., & Jayaraman, S. (2004). Six degrees of jonathan grudin: A social network analysis of the evolution and impact of cscw research. In *Proceedings of the 2004 acm conference on computer supported cooperative work (cscw '04).*

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3–4), 321–377.

Ibáñez, A., Larrañaga, P., & Bielza, C. (2009). Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics*, *25*(24), 3303–3309.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*. Berlin, Germany: Springer.

Knorr-Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Kulkarni, A. V., Busse, J. W., & Shams, I. (2007). Characteristics associated with citation rate of the medical literature. *PloS ONE*, *2*(5), e403.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th international conference on machine learning (icml-01)* (pp. 282–289). Williamstown, MA.

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, *464*(7288), 488–489.

Lane, J., & Bertuzzi, S. (2011). Measuring the results of science investments. *Science*, *331*(6018), 678–680.

Latour, B. (1988). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Laurance, W. F., Useche, D. C., Laurance, S. G., & Bradshaw, C. J. (2013). Predicting publication success for biologists. *BioScience*, *63*(10), 817–823.

Lokker, C., McKibbon, K., McKinlay, R. J., Wilczynski, N. L., & Haynes, R. B. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: Retrospective cohort study. *British Medical Journal*, *336*(7645), 655–657.

Losiewicz, P., Oard, D. W., & Kostoff, R. N. (2000). Textual data mining to support science and technology management. *Journal of Intelligent Information Systems*, *15*(2), 99–119.

Louis, A., & Nenkova, A. (2013). What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *Transactions of the Association for Computational Linguistics*, *1*.

MacRoberts, M. H., & MacRoberts, B. R. (1984). The negational reference: Or the art of dissembling. *Social Studies of Science*, *14*(1), 91–94.

Manjunatha, J. N., Sivaramakrishnan, K. R., Pandey, R. K., & Murthy, M. N. (2003). Citation prediction using time series approach KDD Cup 2003 (task 1). *ACM SIGKDD Explorations Newsletter*, *5*(2), 152–153.

McCarty, C., Jawitz, J. W., Hopkins, A., & Goldman, A. (2013). Predicting author h-index using characteristics of the co-author network. *Scientometrics*, *96*(2), 1–17.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*(1), 86–92.

Myers, G. (1992). In this paper we report...—speech acts and scientific facts. *Journal of Pragmatics*, *17*(4), 295–313.

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Washington, DC: National Science Foundation.

Neelakantan, A., & Collins, M. (2014). Learning dictionaries for named entity recognition using minimal supervision. In *Proceedings of the 14th conference of the european chapter of the* *association for computational linguistics (eacl-14).* Gothenburg, Sweden.

Newman, M. (2003). Mixing patterns in networks. *Physical Review E*, *67*(2), 026126.

Newman, M. (2010). *Networks: An introduction*. Oxford, UK: Oxford University Press.

Pan, R. K., Kaski, K., & Fortunato, S. (2012). World citation and collaboration networks: Uncovering the role of geography in science. *Scientific Reports*, *2*, 902.

Penner, O., Petersen, A. M., Pan, R. K., & Fortunato, S. (2013). The case for caution in predicting scientists' future impact. *Physics Today*, *66*(4), 8–9.

Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., & Schweitzer, F. (2014, February). Predicting Scientific Success Based on Coauthorship Networks. *ArXiv e-prints*.

Schreiber, M. (2013). How relevant is the predictive power of the h index? a case study of the time-dependent Hirsch index. *Journal of Informetrics*, *7*(2), 325–329.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, *87*(2), 373–388.

Spiegel-Rösing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science*, *7*(1), 97–113.

Sun, X., Kaur, J., Milojević, S., Flammini, A., & Menczer, F. (2013). Social dynamics of science. *Scientific Reports*, *3*, 1069.

Swales, J. (1990). Genre analysis: English in academic and research. In (chap. 7: Research articles in English). Cambridge, UK: Cambridge University Press.

Tan, C., & Lee, L. (2014). A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proceedings of acl.*

Teräsvirta, T., Lin, C.-F., & Granger, C. W. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis*, *14*(2), 209–220.

Teufel, S. (2010). *The structure of scientific articles: Applications to citation indexing and summarization*. Stanford, CA: CSLI Publications.

Traweek, S. (1992). *Beamtimes and lifetimes: The world of high energy physicists*. Harvard University Press.

Tsai, C.-T., Kundu, G., & Roth, D. (2013). Concept-based analysis of scientific literature. In *Proceedings of the 22nd acm international conference on conference on information &#38; knowledge management* (pp. 1733–1738). doi: 10.1145/2505515.2505613

Velden, T., Haque, A., & Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks: Mesoscopic analysis and interpretation. *Scientometrics*, *85*(1), 219-242.

Velden, T., & Lagoze, C. (2013). The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *Journal of the American Society for Information Science and Technology*, *64*(12), 2405-2427.

Viana, M. P., Amancio, D. R., & Costa, L. d. F. (2013). On time-varying collaboration networks. *Journal of Informetrics*, *7*(2), 371–378.

Wang, S., Xie, S., Zhang, X., Li, Z., Yu, P. S., & Shu, X. (2014, July). Future Influence Ranking of Scientific Literature. *ArXiv e-prints*.

Wang, X., Smith, K., & Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, *13*(3), 335–364.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 409–10.

Wick, M. L., Kobren, A., & McCallum, A. (2013, Jun). Large-scale author coreference via hierarchical entity representations. In *Proceedings of the icml workshop on peer reviewing and publishing models.* Atlanta, Georgia, USA.

Willinger, W., Paxson, V., & Taqqu, M. S. (1998). Self-similarity and heavy tails: Structural modeling of network traffic. In R. J. Adler, R. E. Feldman, & M. S. Taqqu (Eds.), *A practical guide to heavy tails: Statistical techniques and applications* (pp. 27–53). Boston, MA: Birkhäuser.

Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011). Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th acm international conference on information and knowledge management (cikm-11)* (pp. 1247–1252). Glasgow, UK.

Yogatama, D., Heilman, M., O'Connor, B., Dyer, C., Routledge, B. R., & Smith, N. A. (2011). Predicting a scientific community's response to an article. In *Proceedings of the 2011 conference on empirical methods in natural language processing (emnlp-11)* (pp. 594–604). Edinburgh, UK.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*(2), 408–427. Retrieved from `http://dx.doi.org/10.1002/asi.23179` doi: 10.1002/asi.23179

Ziman, J. M. (1968). *Public knowledge: An essay concerning the social dimension of science*. Cambridge, UK: Cambridge University Press.