Bibliometric Impact Measures Leveraging Topic Analysis

Gideon S. Mann, David Mimno, Andrew McCallum
Department of Computer Science
University of Massachusetts Amherst
Amherst MA 01003
{gmann,mimno,mccallum}@cs.umass.edu

ABSTRACT

Measurements of the impact and history of research literature provide a useful complement to scientific digital library collections. Bibliometric indicators have been extensively studied, mostly in the context of journals. However, journal-based metrics poorly capture topical distinctions in fast-moving fields, and are increasingly problematic with the rise of open-access publishing. Recent developments in latent topic models have produced promising results for automatic sub-field discovery. The fine-grained, faceted topics produced by such models provide a clearer view of the topical divisions of a body of research literature and the interactions between those divisions. We demonstrate the usefulness of topic models in measuring impact by applying a new phrase-based topic discovery model to a collection of 300,000 Computer Science publications, collected by the *Rexa* automatic citation indexing system.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; I.7.4 [**Document and Text processing**]: Electronic Publishing; I.5.3 [**Pattern Recognition**]: Clustering

1. INTRODUCTION

The potential of digital libraries is not only in making documents more accessible, but also in providing automated tools that analyze library collections in order to help readers understand unfamiliar subject areas and guide readers toward significant documents. Over the past forty years there has been increasing interest in bibliometric indicators such as citation counts and Garfield's Journal Impact Factor. These metrics help people without prior detailed content knowledge quickly evaluate the relative importance of individual papers and publication venues. Bibliometric statistics can also provide the basis for visualizations of scientific interactions and automated historiographical analyses. In this paper we demonstrate how recently developed statistical methods that leverage the availability of large digital document collections can enhance bibliometric analysis.

Discovering topical affinities between documents is an active area of research in bibliometrics and scientometrics. In particular, the use of journals as a representation of topics is problematic for a variety of reasons. Journals generally represent a variety of sub-areas and publications often combine multiple topical facets. Additionally, with the growth of open-access publishing, publication venue information is becoming increasingly dispersed and frequently simply unavailable or undefined.

There has been much work recently in machine learning on latent topic models, such as Latent Dirichlet Allocation (LDA) [2], the Author-Topic model [22], and the Author-Recipient-Topic model [19]. These models discover a prespecified number of latent facets or topics, each of which has a particular probability of "emitting" a specific word or token. Documents in the corpus can therefore be associated with a mixture of topics, while particular instances of a given word in the corpus can also be assigned to different topics depending on context. The models have generated significant interest because they create fine-grained, immediately interpretable topics that are robust against synonymy and polysemy. In addition, generating topics and document topic assignments requires virtually no human supervision.

This paper explores new possibilities for impact measures in scientometrics by leveraging topic models. Although topic models have been explored in the context of scientific publications [10], we are not aware of any studies that have explored their use in the presence of citations or in the context of bibliometric indicators. We have performed a large-scale study on a corpus of approximately 300,000 publications in computer science using a new topic model that associates phrases with topics in addition to individual word tokens [26]. The resulting topics are more clearly defined and interpretable than traditional unigram topic models, especially for scientific research texts.

The paper extends journal bibliometric to topics (Citation Count, Impact Factor, Diffusion, Half-Life) and introduces three new topic impact measures: Topical Diversity, Topical Transfer, and Topical Precedence. Topical Diversity ranks highly papers that are cited from many different subfields. Topical Transfer ranks highly papers that are highly cited outside their own subfield. Topical Precedence ranks highly papers that are among the first or help create a subfield.

Given the fine-grained topics discovered by the latent topic models, these impact measures allow a researcher to find the most important papers and the earliest papers from within their field, and also find the most useful work in related

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL'06, June 11–15, 2006, Chapel Hill, North Carolina, USA.
Copyright 2006 ACM 1-59593-354-9/06/0006 ...$5.00.

fields. Our results in Topical Diversity show that highly cited papers do not necessarily have the most broad impact. We also can clearly see that theoretical fields have more diverse impact than the more applied fields. Finally, we demonstrate the use of Topical Transfer to create a graph of the connections and influence among sub-fields.

Our proposed bibliometric analysis is important for helping individuals absorb the research literature, find old and new relevant work, and make connections to other sub-fields. These tools are also especially useful to novices and researchers who are switching fields. They will help researchers to be more efficient, and to improve the quality of their work.

2. RELATED WORK

Our work uses recently developed unsupervised clustering techniques to support topic-centered bibliometric analysis. One related area that extensively uses clustering and association techniques is conceptual visualization. In order to plot 2- or 3-dimensional representations of scientific fields, it is necessary to have an understanding of the distinct subfields, and also a way of measuring the affinity of one subfield to another. Recent work by Klavans and Boyack [18] uses a journal co-citation matrix to calculate affinities. Chen uses clustering techniques to find research fronts and "pivot points" in text collections, and highlights the difficulty of assigning labels to identified clusters [6]. The use of journals as a proxy for scientific topics is common, but generally recognized to be problematic because they may not accurately reflect fine-grained distinctions within fields [1].

Glenisson, Glänzel, and Persson study the use of word-based clustering for bibliometric analysis [16]. They used a word vector model to analyze the text of 19 articles from an issue of Scientometrics. In order to reduce the dimensionality of the vector space, they use 434 bigram features manually selected from the 900 most statistically overrepresented bigrams, such as "citation impact" and "diachronous perspective." Documents are then clustered using a cosine similarity metric. They found that this clustering produced results similar to the judgments of the human editor. Using only titles and abstracts as opposed to the full text of the articles reduces accuracy, but not greatly. Our research is similar in its aims and data sources, but at a significantly larger scale, and with a higher degree of automation.

Garfield has studied the history of topics in scientific literature through "Algorithmic Historiography" [12]. The software package HistCite traces successive layers of citations exported from ISI's Web of Science, starting with an original set of documents and proceeding to newer documents. HistCite is limited in that it requires a small set of "primordial" papers, which must be manually identified. The topic models we present in this paper identify topical areas that can be used to perform similar historiographic visualizations, but without requiring prior knowledge of the field.

Semantic ambiguity is a persistent problem in identifying subject domains: words often have more than one meaning (polysemy), and concepts can often be identified using a variety of terms (synonymy). Christopherson finds that polysemy and synonymy are the primary obstacles to using title and abstract data in finding core documents within research literatures [7]. Börner, Chen and Boyack identify Latent Semantic Analysis (LSA) as one approach to this problem [3]. LSA uses a singular value decomposition to simplify the document/word matrix. Börner uses LSA to cluster the captions of a collection of art images [4], but reports that the procedure is not capable of producing interpretable hidden factors. One of the primary advantages of the topic models used in this study is that they produce hidden topics that are robust against ambiguity, yet can be readily interpreted and used in conceptual mapping.

We use topic models to analyze interactions between scientific sub-fields. Several researchers have recently examined connections between disciplines and the diffusion of ideas across disciplinary boundaries. Small [25] creates a path through all of science by clustering documents based on co-citation. Rowlands proposes a measure of journal impact diversity called the Journal Diffusion Factor (JDF) [23]. This metric is defined as the number of unique journals that cite a particular journal per 100 source citations. A higher JDF implies that citations to the journal are more distributed through a variety of venues.

The analysis of citations over time has been studied extensively. The aging or obsolescence of documents can be studied from two perspectives. Synchronous or retrospective views start with current literature and examine the age of cited literature. Diachronous or prospective views start with articles published in the past and examine the frequency of citations to those articles over a period of time, usually 10 to 15 years [15]. Glänzel argues that the diachronous view is more natural, but that it is only applicable to publications that are old enough to have acquired a citation history. This is an important consideration when using a collection derived from web documents, where most documents are less than 10 years old. One of the earliest and simplest models of citation aging or obsolescence is citation half-life or median citation age. Egghe and Rousseau [9] contains an overview of half-life calculations. More recent work has focused on fitting observed citation distributions to probability distributions or stochastic processes. Good results are observed with Generalized Inverse Gaussian-Poisson [24], negative binomial, and Generalized Waring processes [5].

Changes in the nature of scholarly publication are affecting the applicability of current bibliometric methods and data sources. In many fields, the most current scholarly work tends to appear in conference proceedings and other venues that are not indexed by established bibliographic services. At the same time, scholars are increasingly making their work available through the Internet. One response to this situation is the creation of automated systems that gather online papers, extract and cross-reference citations, and provide an online searchable index. Two such implementations are CiteSeer [14] and Cora [20]. In 2001, Goodrum et al. [17] evaluated citation practices in Computer Science by comparing data from CiteSeer with manually entered data from ISI's SCISEARCH. They found that of the 500 most highly cited documents in CiteSeer, 15% were from conference proceedings as opposed to 3% in SCISEARCH. In addition, 10 of the 200 most cited publications in CiteSeer were unpublished technical reports. Another significant difference is the age of highly cited publications: 42% of the highly cited papers in CiteSeer were from 1990-1999, whereas only 5% of the highly cited papers in SCISEARCH were from that time period. In general, they found that although the research literature available on the web does tend to be recent (95% of the 200 most highly cited papers as of 2001 were less than ten years old), there were a surprising number of documents available that predated the web. In both CiteSeer and SCISEARCH they found some variation

in citations that could affect citation counts. They report that the CiteSeer coreference algorithm depends primarily on the accuracy with which the title is extracted.

3. CORPUS

For the experiments in this paper, we use a corpus from the *Rexa* research paper digital library. The *Rexa* collection has been gathered by spidering the web for research papers in PDF and PS formats. The initial crawl has concentrated on computer science web sites, although some math, statistics, physics, linguistics, and economics are also included.

The spidered PDF/PS files are converted to text, stored in an XML format that preserves layout and font data, and filtered to remove documents that are not research papers. The remaining documents are then segmented to locate the header (containing title, author etc), as well as each individual citation in the references section.

Then, conditional random fields—a sophisticated machine learning model for sequences—extracts up to 14 different fields from the header and each reference [21]. Extracated fields include title, author's first and last names, institution, journal, volume, date, note, etc. The method is highly accurate, for example, extracting the exactly correct title over 97% of the time. Next, these bibliographic records (whether from a citation or header) are coresolved to form the citation graph. Our method is based on efficient graph partitioning with a discriminatively-trained distance measure [27], and is also highly accurate, with precision and recall in the high 90%'s. Finally coresolved bibliographic records are normalized using headers, citations and DBLP metadata when available. Author coreference and normalization is performed also, although the author data is not used in this paper.

Altogether, the collection has 1.6 million research papers, with 200,000 authors, and more than 4 million citations.

As described below, for the experiments in this paper we select a subset of these documents centered on machine learning and some related topics, in order to concentrate our investigation on CS subfields with which we are most familiar. The subset consists of 320,000 documents, with 12 million word occurrences and 450,000 unique words. It includes 225,000 citation instances, coming from 57,000 documents, and citing 117,000 documents. To further clarify, 260,000 papers had no citations (either because references were not extracted or the full text was unavailable), and 203,000 documents weren't cited by any other papers.

There are two major consequences of the automatic collection process. First, the corpus is significantly larger than corpora used in most other bibliometric studies. Second, although the data is more noisy and incomplete than manually-gathered bibliometric data used in some other studies, our experiments demonstrate that the effects due to noise are small, and that the large size of the collection allows for useful and interesting trends to be uncovered.

4. TOPIC MODELS

Latent Semantic Analysis (LSA) [8] has been a popular method of clustering (or low-dimensional projection) for document data; however, it has been largely superceded in the machine learning community by alternative models such as Latent Dirichlet Allocation (LDA) [2], which is based on statistical foundations more appropriate to word count data, and which also tends to produce clearly interpretable output in a more robust manner.

One key strength of LDA over common document clustering methods based on simple mixtures fit by K-Means or EM is that LDA explicitly represents each document as being generated from a document-specific mixture of *multiple* topics (a multinomial over words), rather than a single topic. Thus, for example, a research paper that combines robotics and speech recognition is explicitly represented as such. This flexibility tends to lead to more clear and interpretable topics.

Several related, augmented variants of LDA have been developed in recent years—including the Author-Topic model [22] which captures topic distributions particular to documents' authors, and the Author-Recipient-Topic model [19] which captures the social network in which textual messages are exchanged. These and other methods built on the foundations of LDA have come to be known as "topic models."

In this paper we use a newly-developed topic model called *Topical N-Grams* (TNG) [26]. Rather than representing a topic in terms of individual words, this model also parameterizes per-topic *phrase* statistics. The result is that rather than viewing a topic as a list of single words, TNG can present its user with multi-word phrases, which, especially in research domains, are much more interpretable. For example, instead of trying to divine the meaning of the word list "intelligent," "student," "tutor," the user instead sees the phrase "intelligent tutoring systems." A more detailed example is discussed in the next section.

The generative model for TNG is as follows: first generate a per-document distribution, θ, over topics from a Dirichlet. Next, TNG generates the topic, t of the next word, sampled from θ. Then, conditioned on the previous word and this new topic, t, generate the unigram/bigram status of this upcoming word. If the the status indicates unigram, sample a word from t's unigram distribution, otherwise sample a word from t's bigram distribution. Multi-word phrases are modeled by repeated sampling from the bigram. This model can be fit straightforwardly and efficiently using Gibbs sampling. Full details are provided by Wang and McCallum [26].

4.1 Topic Model Experiments

In order to select a subset of the *Rexa* corpus as described above, we first applied LDA to all 1.6 million titles/abstracts of the *Rexa* corpus, producing 200 topics (in just 33 hours). A sample of the resulting topics can be seen in Table 1. In general, these topics were coarse and often spanned multiple discplines. We selected 24 of the topics as being related to machine learning, natural language processing, robotics and computer vision; papers with greater than 10% of their words assigned to one of these topic were selected. Raw statistics for this subset corpus are presented in section 3.

The Topical N-Grams model was then applied to this subset, and the resulting topics and data are used for the analysis in the remainder of the paper. As is shown in Table 1, the subset topics from TNG are considerably more fine-grained than the original topics. They identify specific subfields—often at a level of specificity tighter than a journal or conference. For example, "Information Extraction" is a subfield of computer science in which there is much interest, and which researchers would want to track; however, there is no journal or conference devoted to this topic.

Topic	Topic Unigrams
Web1 (98)	web information search digital user library users pages content libraries
Web2 (156)	web semantic ontology services world wide based ontologies hypermedia metadata
Computer Vision (5)	recognition object face tracking objects based system image video human
Game Theory (111)	decision making utility equilibrium games theory game choice preferences model
System (160)	system performance communication operating parallel implementation network applications message high

Topic	Topic Unigrams and Ngrams
Digital Libraries (102)	digital electronic library metadata access
	"digital libraries" "digital library" "electronic commerce" "dublin core" "cultural heritage"
Web Pages (129)	web site pages page www sites
	"world wide web" "web pages" "web sites" "web site" "world wide"
Semantic Web (186)	semantic ontology ontologies rdf semantics meta
	"semantic web" "description logics" "rdf schema" "description logic" "resource description framework"
Web Services (184)	web services service xml business
	"web services" "web service" "markup language" "xml documents" "xml schema"

Table 1: Above: Several unigram Topics from 1.6M collection. Below: Several unigram/ngram topics from the 300k subset collection. Topics 98 and 156 from the larger collection are automatically split into 4 fine-grained topics (102,129,186,184) in the subset collection.

Topics are often difficult to distinguish on the basis of single words, but become clear when TNG's phrases are available. Beyond making topics more identifiable and readable to a human, TNG also discovers better topics. For example, when running traditional unigram LDA on the subset corpus, the "Genetic Algorithms" topic becomes confounded with other topics, and is identified with the somewhat cryptic words {algorithms, algorithm, genetic, problems, efficient}. In contrast, TNG discovers a distinct "Genetic Algorithms" topic which is represented clearly by the words {evolution, evolutionary, population, genetic, and fitness}, as well as the phrases {genetic algorithms, genetic algorithm, evolutionary computation, evolutionary algorithms}. Since the word "algorithms" is generated by the bigram model, it does not show up on the top unigram list and allows for a more crisp unigram distribution.

Since the topic models must assign a topic label to all of the words in an abstract, some topics are not exactly research subfields, but "genre" topics; e.g. Topic 149 has top unigrams {efficient, fast, efficiency, speed} and top ngrams {times faster, exhaustive search, cpu time, lookup table, floating point}. These words and phrases are all clearly connected, but they do not correspond to a research subfield. These topics will be omitted from display in the results shown in the remainder of the paper.

Topic	Citations
Speech Recognition (120)	19063
Natural Language Parsing (16)	14764
Computer Vision (49)	12204
Neural Networks (173)	11452
Mobile Robots (22)	10642
Digital Libraries (102)	2822
PCA (135)	1378
Coding And Compression (42)	1302
Game Theory (153)	1212
Computer Security (6)	1058
Sequences (187)	1009

Table 3: The most highly cited topics are application areas, while the lowest cited topics are those most distantly related to our machine-learning-focused 300k subset collection.

5. IMPACT MEASURES

This section presents a series of methods for analyzing research literature using topic models. These methods include several widely used metrics that have been reformulated to use topic membership as the primary organizing factor rather than journal or venue. In addition, we present several additional topic-based methods that are designed to identify aspects of scientific corpora of interest to researchers. Each metric is illustrated using topics derived from the 300k document subset. Given a topic assignment to the words within a document's title and abstract, documents are assigned to all topics which cover more than 10% of the word tokens and are assigned to at least two words. On average a document was assigned to 1.4 topics.

5.1 Citation Count

Raw citation count, the number of documents citing a given document, has been extensively used to measure document impact. While examining the most highly cited documents gives a useful overview of the collection, citation count alone is insufficient for finding the most important papers in a particular sub-field. On the other hand, results from our method, in Table 2, show the highly cited documents from within several specific topics. The most highly cited papers in the "Digital Libraries" topic, for example, provide a good overview of the core publications in the field, but none of them would appear near the top of the list of most cited papers corpus-wide.

Table 3 shows the the citation counts of subfields within the entire collection. These counts are biased with regards to the particular make-up of the collection, and do not necessarily represent the true overall citation counts. For example, the field of Computer Security likely has many more papers (and citations) than what is present in our machine-learning-centered collection.

5.2 Topic Impact Factor

The Journal Impact Factor [13] is commonly used instead of raw citation count as a method for assessing the importance of a particular journal. We present an analogous definition for a topic t in a given year y, where D^y is the set of documents in year y, and D_t is the set of papers in topic t:

$$\textbf{Impact Factor}(t,y) = \frac{\#\text{citations from } D^y \text{ to } D_t^{y-1} \text{ or } D_t^{y-2}}{|D_t^{y-1}| + |D_t^{y-2}|}$$

Impact Factor has the advantage over raw citation count that it is situated in time and accounts for changes in topic

Citations	Title
Digital Libraries (102)	
61	Digital Libraries and Autonomous Citation Indexing
51	Trawling the Web for Emerging Cyber-Communities
28	Going Digital: A Look at Assumptions Underlying Digital Libraries
25	WebBase: a repository of Web pages
24	Lessons learned from the creation and deployment of a terabyte digital video library
Speech Recognition (120)	
223	Estimation of Probabilities from Sparse Data for The Language ...
152	Maximum likelihood linear regression for speaker adaptation of continuous ...
99	Perceptual linear predictive (PLP) analysis of speech
88	Comparison of parametric representations for monosyllabic word recognition in ...
73	Trainable grammars for speech recognition
Hidden Markov Models (21)	
618	A tutorial on hidden Markov models and selected applications in speech processing
173	An introduction to hidden Markov models
163	Hierarchical mixtures of experts and the EM algorithm
152	Maximum likelihood linear regression for speaker adaptation of continuous ...
127	Conditional Random Fields: Probabilistic Models for Segmenting and ...
Dimensionality Reduction (29)	
223	Estimation of Probabilities from Sparse Data for The Language ...
123	A solution to the Plato's Problem: The Latent Semantic Analysis theory of ...
119	The X-tree : An Index Structure for High-Dimensional Data
119	Automatic Subspace Clustering of High Dimensional Data for Data Mining ...
90	A Vector Space Model for Automatic Indexing

Table 2: A per-topic guide to the most highly cited papers allows a researcher to identify highly cited works within a narrow subfield without needing to formulate search terms.

Topic	Impact Factor
Search Engines (132)	1.84
PCA (135)	1.83
Databases (95)	1.72
Machine Translation (96)	1.65
Gene Expression (126)	1.65
Digital Libraries (102)	0.69
Regression (17)	0.37
Internet Routing (98)	0.35
Speech Synthesis (2)	0.33
Signal Processing (94)	0.31
Finance (115)	0.22

Table 4: Topic Impact Factors for Year 2003. Although not necessarily highly cited, the "PCA" topic is shown to be high impact.

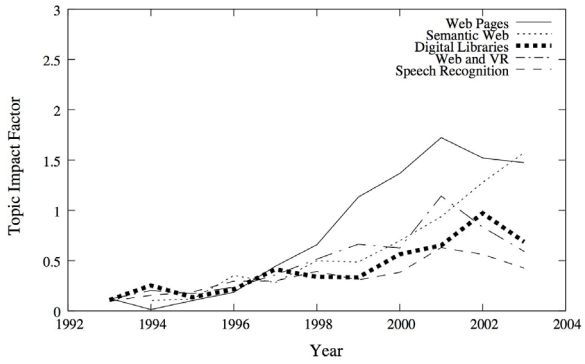


Figure 1: Topic Impact Factors over time. Because of the fine-grained topic resolution, it is possible to see that while the "Web and Virtual Reality" is becoming less important as a topic, "Semantic Web" is becoming more important.

importance over time. Table 4 gives the Topic Impact Factors for the year 2003, which select a very different set of important topics from that given in Table 3. The topic impact factor demonstrates that while there may be few papers in a given topic, those papers may have a wide effect, and that these topics may have unusually high impact with regard to the number of citations.

Figure 1 shows the change in topic impact over time. Fine-grained topics allow subtle trends to be uncovered: since the documents about the "Web" are placed into four different topics, it is possible to distinguish between the field of the "Web and Virtual Reality", which is declining, and the field of "Semantic Web" which is becoming more prominent. Such sub-fields may be at a level of granularity smaller than that of a journal or other venue, and thus would be missed by traditional bibliometric analysis.

5.3 Topical Diffusion and Diversity

One of the trends hidden by citation count and the Topic Impact Factor is whether a particular paper has broad or narrow impact. Does a highly cited paper dominate a prolific sub-field or does it have broad appeal and utility across many fields? In this section, we present topic-based impact measures that reveal more than citation count alone.

The current metric for evaluating broad-based impact is Diffusion [23], defined for a given topic t (with a document set D_t) :

$$\textbf{Diffusion}(t) = \frac{\#\text{ different topics which cite a paper in } D_t \times 100}{\#\text{ citations to } D_t}$$

However, this metric is relatively brittle, since for small topics, one topic citation instance can make a large difference in the diffusion score. In automatic topic analysis, there is a certain degree of noise associated with classifications and this noise can disrupt the true broadness or narrowness of impact. We propose Topic Diversity as a more robust measure of impact diversity. Formally, given a topic t, the citing topic distribution P_t is defined as:

$$P_t(t') = \frac{\#\text{ citations from } D_{t'} \text{ to } D_t}{\#\text{ citations to } D_t}$$

The Topic Diversity is then the entropy of this distribution:

$$\textbf{Diversity}(t) = H(P_t) = -\sum_{t'} P_t(t') log P_t(t')$$

Table 5 shows the results of using the entropy of this distribution to rank topics, and displays the five most and five least diverse topics. From inspection of diversity scores for the full range of topics, it is clear that within our corpus, relatively application-oriented sub-fields (e.g. "Speech Recognition") typically have lower impact diversity than more theoretical sub-fields (e.g. "Pattern Recognition"). While there may be fewer citations to theoretical topics, theoretical topics frequently have broader influence.

For a document d the topical diversity can also be defined similarly:

$$\textbf{Diversity}(d) = H(P_d)$$

Table 6 shows the papers with the highest topical diversity in the test collection. While these papers aren't necessarily the most highly cited, they have broad impact and thus may be more likely to be relevant to practitioners in the field in general. Other highly cited papers often have low impact diversity. For example, "Building a Large Annotated Corpus of English: The Penn Treebank" has 373 citations, but a Topical Diversity of only 3.13 , the same as "Classifier Systems and Genetic Algorithms" which has only 46 citations. In some sense, these broad impact papers may serve as more important background than a paper with a high citation count, as they are more likely to be useful in new research.

Another useful level of analysis is of papers with high impact diversity within a given topic. Table 7 shows for a selection of topics, the papers in that topic which have the highest impact diversity. Within each of these topics, the most diverse papers are those which might be the most relevant to an outsider attempting to understanding key insights of the fields. In the "Dimensionality Reduction" topic, "A Vector Space Model" is cited by 31 other topics such as: "Word Sense Disambiguation" (6 times), "IR and Queries" (17 times), and "Collaborative Filtering" (3 times). The paper "Trainable grammars", which has roughly the same citation count, in contrast is cited by only 23 topics, among them: "Natural Language Parsing" (28 times), "Grammars" (17 times), and "Hidden Markov Models" (9 times).

5.4 Topical Precedence

The institutional memory of science is shorter than it should be and often useful or relevant early work is forgotten. In this section, we define, for a topic t, a document $d \in D_t$ with a publication year of y as having precedence:

$$\textbf{Precedence}(d, t) = y \text{ if } \#\text{ citations to } d > 10$$

Table 8 shows the early documents for a sample of topics. This method successfully find some of the early work in each of the topics of relevance. "RightPages" is a very early digital library system, "Spectrographic study of vowel reduction" is a very early example of digital analysis of speech.

What appears immediately from inspection of the list, is that papers with extracted years are sparse within in the collection. This occurs for two reasons. First, there are few papers on-line from before 1995, and relatively few papers before 1995 are cited by online papers. Second, the extraction methods used to extract year information from references and documents is imperfect, and accurate time information isn't available for all documents. The fact that this metric still allows for useful analysis shows promise that when corpus coverage and extractors improve, this metric will be more useful for longitudinal studies.

5.5 Topical Longevity

To evaluate topical longevity, we use the standard half-life or median citation age metric for a collection of documents assigned to a given topic. This indicator is similar to the Journal Half-life used in ISI's Journal Citation Reports. Topical longevity provides a picture of the overall activity of a sub-field rather than a particular journal, which may shift topics along with the field as a whole or may deliberately tend towards either cutting-edge research or long-sighted review articles. The analysis of topical aging can be approached from two perspectives: synchronously, starting with a set of documents and looking at the age of the documents they cite, and diachronously, starting with a set of documents and looking at the age of the documents that cite them. We calculate synchronous half-life for topic t using Brookes' estimator, as described in [9]. The synchronous calculation depends on three parameters, i, a number of years (usually 6–8, in this case 6), k, the number of cited papers in topic t published i or more years before the current year, and ℓ, the number of cited papers in topic t published less than i years previously. These parameters determine the aging rate a.

$$a = \sqrt[i]{\frac{k}{k+\ell}}$$

The estimated half-life is then

$$\textbf{Longevity}(t) = -\frac{\log 2}{\log a}$$

A large longevity score indicates that for a given topic, the cited documents are relatively old, while a small longevity score indicates that cited documents are relatively recent. Thus, a small longevity score might suggest that there is a quick turn-over in important papers, while a large longevity score might suggest that the important papers have been around for quite a long time. The longevity of selected topics is shown for the median age of citations in publications from 2003 (Table 9).

Figure 3 shows the synchronous citation half-life for three different topics calculated each year over a period of ten years. The plot shows that in "Collaborative Filtering", a relatively new topic, the median age of cited papers has consistently been two years or less, although there has been a slight increase in the past few years. The "Maximum Entropy" topic has a longer "history", but is likewise relatively stable. In contrast, the median age of papers cited in "Neural Networks" has been increasing steadily. Although this topic is still well-cited, there are increasingly fewer citations to recent work.

5.6 Topical Transfer

Aside from high citation count, another mark of distinction for a paper comes with being involved in the movement of ideas from one research discipline to another. These pa-

Topic	Diffusion
Sequences (187)	11.15
Computer Security (6)	9.19
Coding And Compression (42)	8.65
Constraint Satisfaction (25)	8.24
Game Theory (153)	8.13
Digital Libraries (102)	3.97
Mobile Robots (22)	1.14
Neural Networks (173)	1.07
Computer Vision (49)	1.02
Natural Language Parsing (16)	0.82
Speech Recognition (120)	0.66

Topic	Impact Diversity
Simulated Annealing (52)	4.59
Pattern Recognition (125)	4.57
Probabilistic Modeling (3)	4.55
Finite Automata (66)	4.55
Probability (89)	4.5
Digital Libraries (102)	3.77
Machine Translation (96)	3.32
Mobile Robots (22)	3.31
Graphics (9)	3.21
Speech Recognition (120)	3.09
Computer Vision (49)	2.95

Table 5: Impact Diffusion and Impact Diversity. Both models give similar results for narrow impact topics (in particular application areas), but for topics with broader impact, Impact Diffusion is essentially identical to the topics with the lowest citation counts. High Impact Diversity yields more theoretical topics, and appears to be more appropriate for a measure of broad or narrow appeal.

Topical Diversity	Citations	Title
4.00	618	A tutorial on hidden Markov models and selected applications in speech processing
3.80	138	The self-organizing map
3.77	163	Hierarchical mixtures of experts and the EM algorithm
3.74	65	Quantifying Inductive Bias: AI Learning Algorithms and ...
3.74	144	Knowledge Acquisition via Incremental Conceptual Clustering
3.73	155	A Tutorial on Learning With Bayesian Networks
3.72	244	Term-Weighting Approaches in Automatic Text Retrieval
3.71	294	Finding Structure in Time
3.7	173	An introduction to hidden Markov models
3.7	132	Nearest neighbor pattern classification

Table 6: The papers with highest diversity, are often different than the most cited papers, suggesting that sometimes the most highly cited papers affect a narrow set of literature.

Impact Diversity	Citations	Title
Digital Libraries (102)		
3.03	61	Digital Libraries and Autonomous Citation Indexing
2.68	20	Copy Detection Mechanisms for Digital Documents
2.6	15	Repository of Machine Learning Databases: Machine Readable Data Repository
2.58	25	WebBase: a repository of Web pages
2.51	28	Going Digital: A Look at Assumptions Underlying Digital Libraries
Speech Recognition (120)		
3.33	223	Estimation of Probabilities from Sparse Data for The Language Model ...
3.01	63	Self-Organized Language Modeling for Speech Recognition
2.89	73	Trainable grammars for speech recognition
2.84	41	The festival speech synthesis system: system documentation.
2.81	26	The "Neural" Phonetic Typewriter
Dimensionality Reduction (29)		
3.54	90	A Vector Space Model for Automatic Indexing
3.52	123	A solution to the Plato's Problem: The Latent Semantic Analysis ...
3.33	223	Estimation of Probabilities from Sparse Data for The Language Model ...
3.28	46	Probabilistic Latent Semantic Analysis
3.28	77	Probabilistic Latent Semantic Indexing

Table 7: While high citation count within a topic may be most useful for those interested in doing work within that field, documents with high Impact Diversity may be more useful for those outside the field.

Year	Citations	Title
Digital Libraries (102)		
1987	13	Pictures of relevance: a geometric analysis of similarity measures
1992	11	The RightPages Image-Based Electronic Library for Alerting and Browsing
1992	15	Repository of Machine Learning Databases: Machine Readable Data Repository
1995	14	SCAM: A Copy Detection Mechanism for Digital Documents
1995	28	Going Digital: A Look at Assumptions Underlying Digital Libraries
Dimensionality Reduction (29)		
1964	24	A relationship between arbitrary positive matrices and doubly stochastic matrices
1971	18	Direct linear transformation from comparator coordinates into object space ...
1975	90	A Vector Space Model for Automatic Indexing
1983	15	Extending the Boolean and Vector Space Models of Information Retrieval ...
1985	11	A vector quantization approach to speaker recognition
Speech Recognition (120)		
1953	13	Some experiments on the recognition of speech, with one and two ears
1963	14	Spectrographic study of vowel reduction
1965	23	Automatic Lipreading to enhance speech recognition,
1974	11	Effectivness of linear prediction characteristics of the speech wave for ...
1976	12	Automatic Recognition of Speakers from Their Voices,

Table 8: Precedent (early) papers for a selected sample of topics. Although the papers aren't necessarily highly cited, they constitute the past history of the topic and may help in understanding the growth of the field.

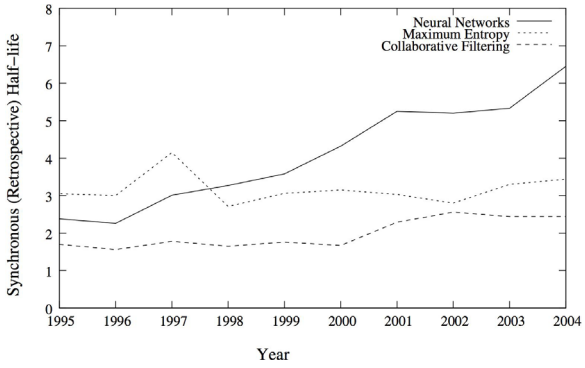


Figure 2: Synchronous (retrospective) Citation Half-Life. The median age of citations to "Collaborative Filtering" papers, a young, fast-moving topic, is consistently less than two years, while "Neural Network" literature is aging rapidly.

Half-life	Topic
6.72	Speech Synthesis
6.70	Cognitive Science (191)
6.59	KL Divergence (151)
6.43	Maximum Likelihood Estimators (40)
6.35	Markov Chain Monte Carlo
6.33	Neural Networks (173)
2.11	Digital Libraries (102)
1.54	Question Answering (111)
1.42	Search Engines (132)
1.32	Web Pages (129)
1.29	Semantic Web (186)
1.00	Web Services (184)

Table 9: Synchronous view of topic citation half-life for references in papers published in 2003

pers often help to explain the past growth of a field and thus useful for understanding where the field is going. For documents, we define the Topical Transfer from a paper $d_{t^1} \in D_{t^1}$ to Topic t^2 (with documents D_{t^2}) as:

$$\textbf{Transfer}(d_{t^1}, t^2) = \#\text{ citations from } D_{t^2} \text{ to } d_{t^1}$$

Table 10 shows the papers with the highest topic transfer out of selected topics, and also shows for the "Digital Library" topic (120), the papers with the highest topic transfer into that topic. For learning about a particular field, this set of documents along with the prior sets discussed in the previous sections provides another relevant view of the document collection.

The Topical Transfer between two topics can also be defined, given a Topic t^1 (with document set D_{t^1}) and Topic t^2 (with document set D_{t^2}) as:

$$\textbf{Transfer}(t^1, t^2) = \#\text{ citations from } D_{t^2} \text{ to } D_{t^1}$$

Using this notion of Topic-to-Topic transfer a birds-eye-view of the relationships between topics in the collection was created. For the graph in Figure 3 an arrow was placed between t^1 and t^2 if the number of citations between the two topics constituted more than 5% of the citations outgoing or incoming from either topic. The size of the arrowhead indicates the number of citations (larger arrowhead indicates more citations in that direction).

The figure displays the connections between fields, where the "Information Extraction" topic and the "Digital Libraries" topic were given as start points, and the graph was drawn to show topics connected up to two edges distant. From this graph, it is possible to observe relationships between research areas. Speech recognition, for example, relies heavily on hidden markov models. Information extraction provides a bridge between various natural language topics and the web.

Though other views of science have been created before, Figure 3 is remarkable in its degree of automation. First, the entire collection along with citations was automatically downloaded from the web. Second, the topic clusters were automatically formed, without using venue information. Finally, the topic names required only inspection of the top words and phrases in each cluster, not an extensive review of the documents within the cluster.

6. DISCUSSION

The use of latent topic models has the potential to solve many of the existing problems with bibliometic methods. At

Topic	Citations	Title
Transfer out of Digital Libraries (102)		
Web Pages (129)	31	Trawling the Web for Emerging Cyber-Communities
Computer Vision (49)	14	On being 'Undigital' with digital cameras: extending dynamic range ...
Video (74)	12	Lessons learned from the creation and deployment of a terabyte digital video library
Graphs (144)	12	Trawling the Web for Emerging Cyber-Communities
Web Pages (129)	11	WebBase: a repository of Web pages
Transfer into Digital Libraries (102)		
Web Pages (129)	8	The Anatomy of a Large-Scale Hypertextual Web Search Engine
Web Pages (129)	7	The WebBook and the Web Forager: An Information Workspace for ...
Video (74)	7	Query by Image and Video Content: The QBIC System
Semantic Web (186)	7	Resource Description Framework (RDF)
Information Retrieval (45)	7	The Harvest Information Discovery and Access System

Table 10: Documents with highest topic transfer from Topic1 to any Topic2.

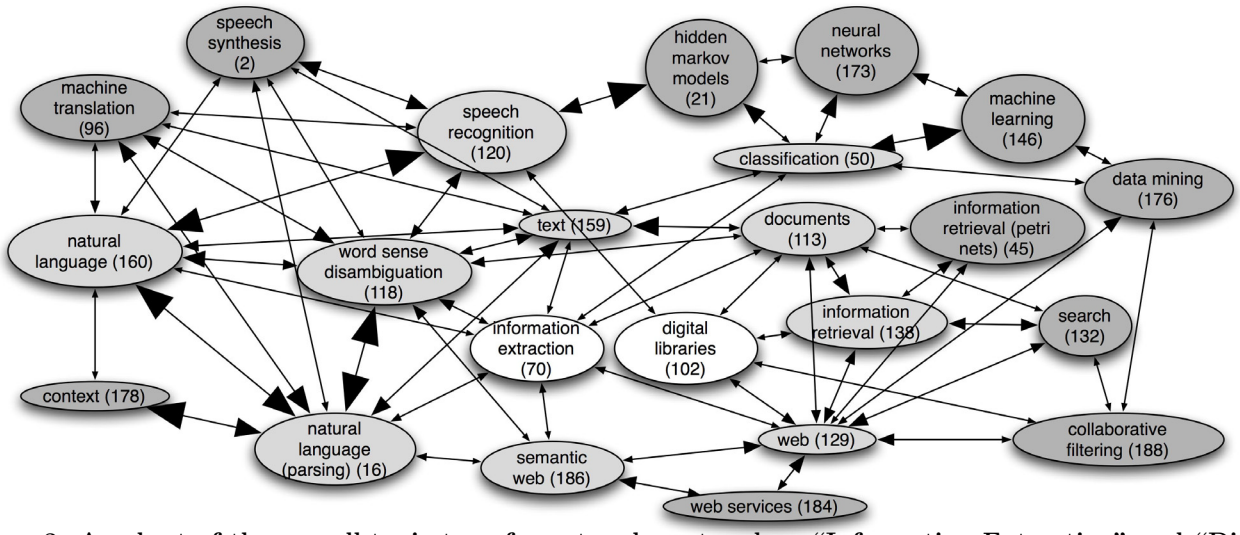


Figure 3: A subset of the overall topic transfer network centered on "Information Extraction" and "Digital Libraries". The size of the arrowheads is proportional to the number of citations from one topic to another.

the same time, topic models can also provide better tools to analyze the dynamics of rapidly changing scientific fields.

There are many advantages to using topic models in analyzing scientific publications. They are robust against semantic ambiguities such as polysemy and synonymy, thus drastically reducing the dimensionality of document representations without losing semantic distinctions. They are capable of assigning the semantic components or facets of documents to different topics, allowing the separation of important topical words from methodological language, a problem identified in [16]. Finally, latent topic models, and in particular the new n-gram topic model used in this study, are capable of producing immediately interpretable topics. Prior work in dimensionality reduction for clustering research papers has either not been able to handle semantic ambiguity or generate a meaningful feature space [3].

An awareness of sub-field distinctions can make bibliometric comparisons more meaningful. It is widely acknowledged that different scientific areas have different levels of citation activity: a high citation count in mathematics might be merely average in molecular biology [11]. At a smaller scale, comparing documents within one small subfield such as collaborative filtering can provide more meaningful information about the relative impact of a paper than measurements within a broader field such as artificial intelligence.

The primary advantage of topic modeling in the context of the study of research literature, however, is its unprecedented automation, from the initial automated discovery of research publications to the creation of high-level visualizations. Automatic analysis will become increasingly important in the context of the current move towards open-access publishing. More and more documents are becoming available in full-text form on author websites. Scholarly documents are appearing in a wider variety of venues, many of which have never been indexed by standard bibliographic databases. Simply making text available, however, is only part of the promise of open-access publishing. A potentially larger impact will come from the ability to detect trends in scientific literature based on large collections of spidered web-accessible documents. Many of the standard "journalometric" measurements will become less applicable as venue information becomes more dispersed or is simply not available. At the same time, the availability of full text and the ability to analyze it may shift bibliometric indicators away from the analysis of journals to the analysis of more specialized sub-fields.

More automated tools for scientific historiography will especially benefit researchers. As scientific literature continues to expand, it becomes increasingly difficult for individuals to follow work in their own fields, much less related fields. Using bibliometric analysis based on topic models and freely accessible digital documents, a researcher could identify subfields within a broader research field, determine which subfields are rapidly advancing and which are more established,

understand the resources, tools, and applications of a sub-field, and view the bibliographic history of the topic from its earliest beginnings. All of this can occur within one click of the source documents themselves.

Given access to sufficient quantities of documents and the citation relationships between them, the methods used in this study are easily reproducible. Several packages for performing topic modeling are freely available on the Internet, including ours. All of the metrics discussed in this paper are relatively simple to implement given topic and citation data and moderate programming experience.

7. CONCLUSIONS

We have demonstrated the applicability of topic model-based sub-field discovery to bibliometric evaluation in scientific publications. We have shown how topic models facilitate several new ways of measuring the impact of individual documents and sub-fields. In addition, there are several journal-based bibliometric indicators that can easily be modified to use topics instead of publication venues, resulting in more scientifically meaningful results. The automated and text-centered nature of topic modeling makes it ideally suited to take advantage of the explosion of open-access publishing. Topic-based visualization and historiography have the potential to significantly increase the accessibility of open-access scientific document collections.

This paper has only begun to address the potential uses of latent topic modeling in bibliometric and scientometric analysis. In particular, applications in visualization and mapping are particularly promising. In addition, more work is necessary in evaluating and improving the quality of automatic topic models.

8. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, in part by the Defense Advanced Research Projects Agency (DARPA), and the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor. We would also like to thank Xuerui Wang, Charles Sutton, and Andrew Tolopko for their advice and assistance.

9. REFERENCES

[1] D. W. Aksnes, T. B. Olsen, and P. O. Seglen. Validation of bibliometric indicators in the field of microbiology: A norwegian case study. *Scientometrics*, 49(1):7–22, 2000.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] K. Börner, C. Chen, and K. W. Boyack. Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37, 2003.

[4] K. Börner, A. Dillon, and M. Dolinsky. LVis – digital library visualizer. In *Information Visualization 2000, symposium on Digital Libraries*, pages 77–81, 2000.

[5] Q. L. Burrell. The use of the generalized Waring process in modelling informetric data. *Scientometrics*, 64(3):247–270, 2005.

[6] C. Chen. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST*, 2006.

[7] M. Christopherson. Identifying core documents with a multiple evidence relevance filter. *Scientometrics*, 61(3):385–394, 2004.

[8] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[9] L. Egghe and R. Rousseau. *Introduction to Informetrics: quantitative methods in library, documentation, and information science*. 1990.

[10] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *PNAS*, 101(Suppl. 1):5220–5227, 2004.

[11] E. Garfield. Expected citation rates, half-life, and impact ratios: comparing apples to apples in evaluation research. *Current Contents*, 1994.

[12] E. Garfield. Historiographic mapping of knowledge domains literature. *Journal of Information Science*, 30(2):119–145, 2004.

[13] E. Garfield. The history and meaning of the journal impact factor. *Journal of the American Medical Association*, 293:90–93, January 2006.

[14] C. Giles, K. Bollacker, and S. Lawrence. Citeseer: An automatic citation indexing system. In *DL'98 Digital Libraries, 3rd ACM Conference on Digital Libraries*, pages 89–98, 1998.

[15] W. Glänzel. Towards a model for diachonous and synchronous citation analyses. *Scientometrics*, 60(3):511–522, 2004.

[16] P. Glenisson, W. Glänzel, and O. Persson. Combining full-text analysis and bibliometric indicators. a pilot study. *Scientometrics*, 63(1):163–180, 2005.

[17] A. Goodrum, K. W. McCain, S. Lawrence, and C. L. Giles. Scholarly publishing in the internet age: a citation analysis of computer science literature. *Information Processing and Management*, 37(5):661–675, 2001.

[18] R. Klavans and K. W. Boyack. Identifying a better measure of relatedness for mapping science. *JASIST*, 57(2):251–263, 2006.

[19] A. McCallum, A. Corrada-Emanuel, and X. Wang. Topic and role discovery in social networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.

[20] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3:127, 2000.

[21] F. Peng and A. McCallum. Accurate information extraction from research papers using conditional random fields. In *HLT-NAACL*, 2004.

[22] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.

[23] I. Rowlands. Journal diffusion factors: a new approach to measuring research influence. *Journal of Documentation*, 54:77–84, 2002.

[24] H. S. Sichel. A bibliometric distribution which really works. *JASIS*, 36(5):314–321, 1985.

[25] H. Small. A passage through science: crossing disciplinary boundaries. *Library Trends*, 48(1):72–108, 1999.

[26] X. Wang and A. McCallum. A note on topical n-grams. Technical Report UM-CS-2005-071, University of Massachusetts, Amherst, December 2005.

[27] B. Wellner, A. McCallum, F. Peng, and M. Hay. An integrated, conditional model of information extraction and coreference with application to citation matching. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.