This thesis should be cited as:
Savov, P., 2021. Measuring the Novelty of Scientific Papers. Ph.D. Thesis. Polish-Japanese Academy of Information Technology.

Measuring the Novelty of Scientific Papers

by

Pavel Savov

Supervisor

Dr hab. Jerzy Paweł Nowacki, prof. PJAIT

Auxiliary Supervisor

Dr Radosław Nielek

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy (Computer Science)

at the
Polish-Japanese Academy of Information Technology

2021

Declaration of Authorship

I, Pavel Savov, declare that this thesis titled, 'Measuring the Novelty of Scientific Papers' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Streszczenie

Z powodu szybkiego wzrostu liczby artykułów naukowych publikowanych co roku coraz trudniej jest nadążać za rozwojem nawet tylko swojej dziedziny nauki. Badacze, a także np. urzędnicy decydujący o przydziale środków na badania naukowe polegają na tradycyjnych indeksach scjentometrycznych w celu wyszukiwania obiecujących lub potencjalnie przełomowych projektów badawczych. To podejście jest jednak obarczone pewnymi wadami.

Celem niniejszego projektu badawczego jest poszukiwanie rozwiązań niwelujących te wady i zaproponowanie automatycznej metody pomiaru innowacyjności publikacji naukowych poprzez predykcję ich wieku na podstawie analizy tekstu. Na niniejszą pracę składają się trzy recenzowane artykuły opublikowane w znaczących międzynarodowych źródłach opisujące postęp prac nad predykcją dat publikacji przy użyciu modelowania tematycznego, nadzorowanych modeli predykcyjnych i wreszcie aktualnych modeli osadzania słów (BERT) trenowanych na diachronicznych korpusach artykułów naukowych obejmujących wieloletnie okresy. Na bazie tych predykcji zaproponowano liczbową miarę odzwierciedlającą podobieństwo zawartości ocenianych artykułów do zawartości artykułów publikowanych w przyszłości lub przeszłości, a zatem ich prawdopodobną innowacyjność.

Proponowaną metodę zastosowano na trzech korpusach obejmujących publikacje ze źródeł wiodących w swoich dziedzinach. Pokazano, jak wartości proponowanej miary innowacyjności korelują z liczbą cytowań. Pokazano też na przykładzie dwóch korpusów obejmujących ponaddwudziestoletni okres, jak przy użyciu modeli BERT obniżyć średni błąd bezwzględny dla predykcji wieku publikacji odpowiednio z 3,56 i 2,56 roku do 0,68 i 0,64 roku.

POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY

Abstract

Faculty of Information Technology

Doctor of Philosophy

by Pavel Savov

As the number of scholarly papers published each year keeps growing, it is becoming increasingly difficult to follow all research, even in one's own area. Researchers as well as decision makers at funding bodies have relied on traditional scientometric measures to identify promising research or potential breakthroughs. This approach, however, has several flaws.

The aim of this project is to address those shortcomings and propose an automated method of identifying novelty utilizing paper age prediction based on full text content analysis. The following dissertation is comprised of three peer-reviewed papers published at highly-ranked publication venues, describing incremental research on predicting publication dates using latent topic models, supervised prediction models, and finally state-of-the-art word embeddings (BERT), trained on diachronic corpora of papers published over multi-year periods. Based on the results of these predictions, a real-number metric has been proposed, reflecting how similar the papers' content is to that of past or future papers, and thus, how innovative it likely is.

The method has been applied to three corpora of papers from leading venues in their respective areas. It has been shown how the proposed innovation scores correlate to citation counts. Finally, it has been shown how, using BERT models, the mean age prediction error on two test corpora spanning over 20 years may be reduced from 2.56 and 3.56 years to 0.68 and 0.64 years respectively, compared to the original topic model-based approach.

Acknowledgments

I wish to thank Professor Jerzy Paweł Nowacki for agreeing to be my supervisor.

I would like to extend my sincere gratitude to my auxiliary supervisor Dr Radosław Nielek for his great mentorship, patience, tremendous support during my PhD study, and all his invaluable comments and remarks. I also greatly appreciate how he shielded me from the less fun parts of Academia.

I am profoundly grateful to Dr Adam Jatowt, without whom this project could not have been conceived, for his brilliant ideas, countless inspiring discussions and all the research we have done together. Working with him on this project has been a great pleasure and honor.

I must also thank Professor Adam Wierzbicki for his insightful comments and constructive criticism which greatly helped me improve this dissertation.

Finally, I wish to thank my family for their support and putting up with me staring at my computer all the time.

Contents

List of Figures

List of Tables

Abbreviations

BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
CTM	**C**orrelated **T**opic **M**odels
LDA	**L**atent **D**irichlet **A**llocation
SVM	**S**upport **V**ector **M**achines
TF-IDF	**T**erm **F**requency - **I**nverse **D**ocument **F**requency

Chapter 1

Introduction

In this work we propose an innovative automated method of assessing the novelty of scientific papers based on the analysis of the papers' textual content only. Its aim is to complement traditional scientometrics in identifying potentially pioneering or breakthrough publications by finding papers covering topics popular in the future (with respect to the papers' publication dates). As outlined in Section 1.4, traditional approaches suffer from various biases. Therefore, this method may be useful as a tool facilitating the understanding of the development of fields of research. Its practical applications also include e.g. helping funding bodies identify promising research in the process of selecting grant beneficiaries.

The following dissertation is comprised of three peer-reviewed papers published at highly-ranked publication venues, and is structured as follows: Chapter 1 describes the problem, provides an outline of the proposed solution and discusses its limitations and possible directions for future research. Chapter 2 provides an overview of related literature. Chapters 3, 4, and 5 contain the aforementioned papers verbatim:

- P. Savov, A. Jatowt, R. Nielek. "Identifying Breakthrough Scientific Papers." *Information Processing & Management* 57.2 (2020): 102168. – full-length article in a leading journal (impact factor: 4.787) (Chapter 3),
- P. Savov, A. Jatowt, R. Nielek. "Innovativeness Analysis of Scholarly Publications by Age Prediction using Ordinal Regression" *International Conference on Computational Science*, pp. 646-660. Springer, Cham, 2020 – paper in the proceedings of an A-ranked (excellent) conference according to the CORE conference ranking[1] (Chapter 4),

[1] `http://portal.core.edu.au/conf-ranks/`

- P. Savov, A. Jatowt, R. Nielek. "Predicting the Age of Scientific Papers", *International Conference on Computational Science*, 2021 – paper at the International Conference on Computational Science (A-ranked according to CORE 2020) (Chapter 5).

Section 1.5 describes the relationship between these papers, how they stem from one another, and how they constitute a research project.

The source code in Python implementing the proposed method, used during its development, is available publicly at `https://github.com/pavelsavov/paper-scores.git`. The implementation requires a plain text corpus divided into time slices. All documents are preprocessed – converted to lowercase, punctuation, stopwords and numbers are removed, and finally lemmatization is performed. Multiple topic models are built, the optimal model is selected and used as input for training the prediction model, which is then used for predicting the ages of all documents. Innovation scores are finally calculated for all documents in the input corpus, based on the prediction results. The details of the algorithm are described in Chapters 3 and 4.

1.1 Research Problem and Goals

Scientometrics – measuring science – in its current shape and form, dates back to the mid 20^{th} century and, in particular, to works such as "*Citation Indexes for Science*" by Eugene Garfield [1], or "*Little Science, Big Science*" by Derek John de Solla Price [2]. The main focus of scientometrics is on measuring the innovation and impact of scientific publications, their authors and publication venues. The most important academic journal in the field is *Scientometrics* founded in 1978 by Tibor Braun. Aside from providing better understanding of the evolution of particular fields of study, the identification of innovative or potential breakthrough publications also serves a practical purpose – it helps research funding bodies select the most promising projects to invest in. Recent research in this area includes works such as: Schneider and Costas [3, 4], Ponomarev et al. [5], or Wolcott et al. [6].

Traditionally, citation analysis has been used to identify pioneering scientific papers. This approach, however, suffers from various biases. Works by well-known authors and/or ones published at well-established publication venues (similar to the rich-get-richer effect) tend to receive more attention than others. Papers could then achieve increased visibility through early citations [7]. Widely cited papers also tend to attract even more citations, while some innovative ideas may be appearing in papers well before their popularity time, hence, receiving little recognition.

This dissertation demonstrates a novel machine learning-based method of analyzing corpora of scholarly papers published over a period of time. The aim is to find the answers to the following questions:

- Which are the most pioneering papers, i.e. papers being ahead of a trend, or early papers covering topics that would have become popular in the future?
- Which are the *breakthrough* years, i.e. years when trends that were later researched for several years were started?
- Can we offer a supplementary approach to traditional citation analysis for assessing the merit of publications, without requiring expert knowledge and/or manual labor?

1.2 Novelty vs. Impact

Traditional scientometrics focus on measuring the impact of publications and researchers, i.e. their influence on their field, but not necessarily their novelty. Also, impact may only be measured retrospectively. For the purpose of this research we define novelty as early occurrences of new topics or ideas, not researched before. The proposed method measures the novelty of publications regardless of their impact. In the case of recent papers, novelty may or may not lead to impact, but we assume that novel publications have higher potential than others.

1.3 Why Identify Novelty?

As the number of papers published each year keeps growing, it is becoming increasingly difficult to follow all published research, even in one's own field. The ability to find innovative publications automatically would be useful to researchers for finding inspiration, identifying promising research directions, or creating a bibliography. It would help investors or decision-makers in funding bodies select the most promising and potentially groundbreaking projects, to allocate finite resources in the best possible way. Also, non-expert readers of technical documents lacking timestamps (e.g. web pages) may wish to know how novel or outdated they (or their parts) are. Finally, the methods proposed in this dissertation may be used to give credit to less prominent authors for their work, which might otherwise remain unknown. Due to the biases of citation analysis outlined in Section 1.4, truly innovative works by less well-known authors may remain relatively

unnoticed, while publications covering established topics, but authored by more influential researchers, may receive more citations, and thus be regarded as more impactful. This may also benefit the scientific community and science in general by facilitating the dissemination of novel ideas regardless of the prominence of their authors.

1.4 Shortcomings of Traditional Scientometrics

Relying solely on citation counts as a measure of scientific output has been criticized for a number of reasons. Problems with citation analysis include:

- Citing prominent publications, following the crowd [7]
- Matthew Effect – term inspired by the biblical Gospel of Matthew; according to Merton [8], who first described this phenomenon in 1968, publications by more eminent researchers will receive disproportionately more recognition than similar works by less-well known authors.
- Increased visibility through early citations: Singh et al. [9] have shown that papers cited within two years of publication tend to attract more citations. They have observed, however, that early citations by influential authors negatively impact the cited paper's long term scientific impact by way of *attention stealing* where the subsequently published citing paper authored by the more influential and well-known researcher collects further citations instead of the original work.
- Google Scholar Effect: Serenko and Dumay [10] observed that old citation classics keep getting cited because they appear among the top results in Google Scholar, and are automatically assumed as credible. Some authors also assume that reviewers expect to see those classics referenced in the submitted paper regardless of their relevance to the work being submitted.
- Self-citations: Increased citation count does not reflect the work's impact on the field of study.
- Ignoring the purpose of citations (support vs criticism)
- Erroneous citations
- Slowness: It may take several years to acquire the first citations [11].

1.5 Proposed Approach

Our approach is based on comparing the predicted publication years of scientific papers with their actual publication years. Predictions are made by a supervised model trained on the textual content of the analyzed papers. The main novelty is the usage of the prediction model's error as input for assessing the innovativeness of the analyzed papers. The idea behind it is that *the more a paper's vocabulary or topic distribution resembles that of papers published in the future (and the less it resembles the ones from the past), the more innovative the paper is.*

The first version of the method, described in detail in Chapter 3, uses multiclass Support Vector Machine (SVM) classifiers with Latent Dirichlet Allocation (LDA) topic distributions as feature vectors. A real-number paper innovation score is proposed, based on the results of the classifier's predictions. The iteration described in Chapter 4 replaces SVM classifiers with an ordinal regression model and LDA topic models with Correlated Topic Models (CTM). Chapter 5 describes how the prediction accuracy can be improved by using BERT – state-of-the-art embedding models fine-tuned for text classification to predict the age of each sentence and aggregating the results to predict the age of the entire paper.

The proposed method assesses the novelty of scientific papers solely by analyzing their text content. None of its steps – topic model training and selection, prediction model training and score calculation – require expert knowledge or manual intervention.

1.6 Method Validation

We show the results of applying our method to the corpora of papers published at three well-known and influential publication venues in distinct yet overlapping fields, each with a history of over two decades, and covering a broad range of topics: The Web Conference, formerly known as The International World Wide Web Conference (WWW), the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), and the Journal of Artificial Societies and Social Simulation (JASSS). Both conferences are top-tier venues in their respective fields – among the top 4% according to the CORE conference ranking. JASSS is a leading journal in the field of social simulation and computer simulation in social sciences with an impact factor of 2.222 as of 2019.

The International World Wide Web Conference (WWW) was first held in May 1994 at CERN in Geneva, Switzerland by Robert Cailliau – one of the founders of the World

Wide Web. It was held once more in October 1994, twice in the spring and autumn of 1995 and once a year ever since by the International World Wide Web Conference Committee (IW3C2)[2] founded by Robert Cailliau and Joseph Hardin. In 2018 the conference has been renamed to The Web Conference. It has served as an important publication venue influencing many researches that center around diverse aspects of the Web.

The SIGIR conference has been held annually by the Association for Computing Machinery's Special Interest Group on Information Retrieval since 1978. It is considered the most important conference in the field of information retrieval. Research areas covered by SIGIR include: document representation, content analysis, query analysis, content recommendation, social media analysis, etc. In addition to the usual Best Paper awards, the *Test of Time Award* has been awarded since 2014 to papers published 10-12 years before that have had "long-lasting influence, including impact on a subarea of information retrieval research, across subareas of information retrieval research, and outside of the information retrieval research community"[3]. Every three years, a researcher is also awarded the *Gerard Salton Award* for "...significant, sustained and continuing contributions to research in information retrieval"[4].

The Journal of Artificial Societies and Social Simulation (JASSS) is an open-access quarterly *interdisciplinary journal for the exploration and understanding of social processes by means of computer simulation*[5]. It has been published by the European Social Simulation Association since 1998 and is publicly available at `http://jasss.soc.surrey.ac.uk/`.

In this work we propose a novel method for analyzing the popularity of research topics over time in fields of study, for which there exists a significant body of work that spans multiple years. Based on this topic analysis, we predict the publication years of research papers in the analyzed fields of study, and assign innovation scores measuring how far ahead of (or behind) their time are the topics appearing in the paper in question. By aggregating the innovation scores for all papers published in every year, we then calculate importance scores for each year and identify breakthrough years, in which topics popular in the future were first researched. We also include lists of topics extracted from both analyzed corpora with data on their popularity over time, and lists of papers identified as the most innovative. The scores produced by our method were compared to citation counts and correlation coefficients were calculated (see Sections 3.5.4 and 4.5).

[2] `http://www.iw3c2.org/`
[3] `http://sigir.org/awards/test-of-time-awards/`
[4] `http://sigir.org/awards/gerard-salton-awards/`
[5] `https://jasss.org/admin/about.html`

1.7 Contribution to Computer Science

This dissertation makes the following contributions:

- Algorithm for dating scientific papers in a given domain, using latent topic models and Support Vector Machine classifiers (see Chapters 3 and 4),
- Paper Innovation Score – a real-number measure of scientific paper novelty based on age prediction result (see Chapters 3 and 4),
- Improved method of dating scientific papers using state-of-the-art word embeddings and ordinal regression (see Chapter 5).

Chapter 2

Literature Review

2.1 Citation Analysis

Traditional scientometrics revolve around citations. As outlined in Section 2.2, the importance of journals, conferences, papers, books, authors, etc. is measured by the number of times they have been cited. In general, the more citations the greater the impact, however as described in Section 1.4, this approach is not perfect.

Citation analysis is not merely about counting citations. Price [12] described how citation networks, i.e. directed graphs, where vertices represent publications and the edges point from citing papers to cited papers, may be used to study citation patterns in collections of documents. This is then useful to study the development of research fields, or identify important topics and researchers.

2.2 Scientometric Measures

This section outlines various measures of scientific impact – measuring the importance of individual articles, entire publication venues, as well as the output of specific authors.

2.2.1 Journal-level

Some of the most widely recognized academic impact metrics rank publication venues according to their impact on their respective fields. Usually the higher the rank, the more prestigious the venue, and the more it is regarded as influential. Some examples of journal-level metrics include:

- Institutional Lists: An example is the ranking of journals and conference proceedings by the Polish Ministry of Science and Higher Education (Ministerstwo Nauki i Szkolnictwa Wyższego)[1], assigning scores from 20 (worst) to 200 (best) to academic journals and conferences, and used to measure the output of scholars and researchers. Such lists are usually decided by a committee and/or experts. This approach has been criticized for its potential for bias [13].

- Impact Factor: Calculated annually from 1975, the two-year impact factor for the year y is the mean number of times articles published in the previous two years were cited in the year y. The impact factor was proposed by Garfield [14] as a means of comparing the quality of journals in a given field, in order to help libraries choose journals to subscribe to. The journal impact factor is commonly (mis)used to evaluate individual articles or the output of specific authors, however this approach has been criticized by Garfield himself [15] due to the high variation of citations between papers in a single journal. The use of the mean instead of the median for non-normal distributed data has also been criticized [16].

- Eigenfactor: Proposed by Bergstrom et al. [17], based on the same concept as the PageRank algorithm [18], the Eigenfactor measures the total number of citations, giving more weight to citations from journals with higher Eigenfactor values. Unlike the impact factor, the Eigenfactor is a weighted sum rather than the mean, it is therefore highly dependent of the size of the journal. The scores may be viewed at `http://eigenfactor.org/`.

- SCImago Journal Rank (SJR): Inspired by PageRank and using a similar iterative algorithm, the SJR, like the Eigenfactor, gives more weight to citations from more highly-scored journals [19].

- Source Normalized Impact per Paper (SNIP): Introduced in 2012 by the publisher Elsevier[2] and available in Scopus[3], calculated as the three-year Impact Factor normalized to account for the citing practices in different fields: "Essentially, the longer the reference list of a citing publication, the lower the value of a citation originating from that publication."[4] [20].

- Conference rankings: In some fields more than others – computer science is a notable example – conferences play an important role. One of the most well-known and universally trusted conference rankings is the CORE Ranking[5] published and

[1] `http://www.bip.nauka.gov.pl/`
[2] `https://www.elsevier.com/`
[3] `http://www.scopus.com/`
[4] `https://lib.guides.umd.edu/bibliometrics/SNIP`
[5] `https://www.core.edu.au/conference-portal`

periodically updated since 2006 by The Computing Research and Education Association of Australasia. The CORE Ranking provides an assessment of major conferences in the broad field of computer science and assigns them to eight categories, the most important of which are:

- A*: Flagship
- A: Excellent
- B: Good
- C: Other ranked conference venues that meet minimum standards

Conferences are ranked based on citation rates, submission and acceptance rates (the more difficult it is to have a paper accepted, the better), and the track records of the key people hosting the conferences.

2.2.2 Article-level

The Journal Impact factor has been used as the standard indicator of research quality, but using journal-level metrics as a criterion for selecting individual papers for reading has been criticized due to the high variation of article impact within journals. Rossner et al. [16] pointed out that, following the Pareto principle, 20% of papers published in the Nature journal account for 80% of the journal's impact. Citation counts as the measure of papers' quality has also been criticized for its time-delay. Reputation systems based on comments similar to StackOverflow[6] have been proposed, but never gained traction in academia [21]. A major factor is the comments' sparseness combined with the relatively small size of the scientific community.

Various article-level metrics have been proposed, including download and view counts, as well as the number of times the article was bookmarked in tools like Zotero or Mendeley (see Section 2.3). Around the year 2010 article-level metrics started being provided by many publishers and libraries including the Public Library of Science[7], Elsevier, ACM's Digital Library[8] etc.

2.2.3 Author-level

The most prestigious acknowledgment of an individual's scientific output is the Nobel prize. Being employed at a top university is also used as an indicator of quality of one's research.

[6] `https://stackoverflow.com/` – a Q&A site for programmers, where users answering questions earn reputation points and thus gain credibility

[7] `http://plos.org/`

[8] `https://dl.acm.org/`

Various metrics have been proposed for quantifying the output of individual authors. Perhaps the most well-known and widely used is the h-index proposed in 2005 by Hirsch [22]: An author's h-index is the maximum number h such that h of the author's papers have at least h citations each. Due to differences in citation practices between fields, the h-index is only suitable for comparing the output of authors in the same field of research. Batista et al. [23] and Kaur et al. [24] addressed this shortcoming of the h-index by introducing scaling factors based on the numbers of authors in each discipline. The h-index normalized by these factors may thus be compared across different fields. Numerous other variations of the h-index have been developed, e.g. taking into account the order of authors in publications having multiple authors [25], or a time window-based version for studying the careers of researchers over time [26].

Another example of a metric based on citation counts is the $i10$-index introduced in 2011 by Google Scholar. It is defined as the number of publications having at least 10 citations.

West et al. [27] have proposed the author-level Eigenfactor metric based on the journal-level Eigenfactor. Like its journal-level counterpart, it gives more weight to citations by authors with higher Eigenfactor values. The Eigenfactor score of an author is their eigenvector centrality [28] in the citation network.

Besides measuring citation impact, studying scientific collaboration networks and the ranks of authors in those networks also gives an insight into the productivity and importance of researchers [29, 30]. One of the most central figures in science was the mathematician Paul Erdős, who published more papers than any other mathematician [29]. The distance from Paul Erdős in the co-authorship graph is known as the *Erdős number*. It has been shown that many leading mathematicians have low Erdős numbers [31]. It could be argued, therefore, that the Erdős number or distances from other central researchers in the co-authorship graph measures the importance of a researcher.

Social networking sites such as ResearchGate[9] have made feasible complex metrics taking into account various indicators. The RG Score calculated for each member of the network is based on both published and unpublished research, various types of feedback from others, such as citations, recommendations, following and being followed, as well as activity in the community such as asking and answering questions. It also depends on the scores of the interacting members.

[9] `https://www.researchgate.net/` – networking site for researchers where research may be shared, recommended etc.

2.3 Altmetrics

Altmetrics, proposed by Priem et al. [11] as an alternative to traditional scientometrics, aim to address some of the problems outlined in their 2010 manifesto: Slowness of peer review and citation counting, inability of the Journal Impact Factor (JIF) to measure the impact of individual publications, and the relative ease of gaming the JIF. Aside from being slow (it may take an article several years to gather its first citations), citation counting is criticized for its inability to take into account the context and reason for citing, as well as ignoring the papers' impact outside academia.

There is no strict definition of altmetrics, but they generally measure the papers' various aspects of impact on the Web and social media, such as mentions on Twitter, recommendations e.g. on ResearchGate. They fall into the following categories according to Lin and Fenner [32]:

- Viewed: The number of times the article has been accessed online, e.g. in a digital library,
- Saved: Saves in online bibliography management tools such as Zotero[10] or Mendeley[11],
- Discussed: Mentions in tweets, blog posts, Wikipedia pages, etc.,
- Recommended: Recommendations on platforms such as ResearchGate,
- Cited: Citations in peer-reviewed scientific journals or conference proceedings.

Several nonprofit organizations such as Our Research[12], and companies (Altmetric[13], Plum Analytics[14]) collect altmetrics. Several publishers and digital libraries, e.g. ACM's DL provide altmetrics on their articles.

Costas et al. [33] made a comparison of altmetrics from Altmetric.com with traditional citation counts. They compared the altmetrics' coverage in various fields as well as their correlation with citations. The coverage has been rising steadily from around 15% in 2011 to over 20% in 2015. The correlation with citations was found to be only moderate. The final conclusion was that in order for altmetrics to be able to replace citation scores, substantially more than 30% of papers should have them. Furthermore, the moderate

[10] `https://www.zotero.org/`
[11] `https://www.mendeley.com/`
[12] `https://our-research.org/`
[13] `https://www.altmetric.com/`
[14] `https://plumanalytics.com/`

correlation with citation counts suggests that altmetrics and citations measure different kinds of impact.

A recent study of altmetrics' accumulation patterns and velocity by Fang and Costas [34] showed that altmetrics are not necessarily "fast", even though "speed" has been advertised as one of the key characteristics of altmetrics and citations have been criticized for their "slowness". However, some sources of altmetrics (Reddit, Twitter, News, Facebook) were found to be faster than others (Policy documents, Q&A, Peer review, Wikipedia). Variations between research fields and individual topics within those fields were also found.

2.4 Development of Research Areas

The development of research areas and the evolution of topics in academic conferences and journals over time have been investigated by numerous researchers. An early example is the paper by Lounsbury et al. [35] analyzing the content of all 604 articles published in the *Community Mental Health Journal* between the years 1965 and 1977. The authors manually identified 61 topics and studied topic trends over time.

The turn of the 21st century saw a rapid increase in the number of papers published yearly at many publication venues, see Figure 3.1. This growth of the number of available publications made manual analysis by experts impractical and created the need for unsupervised automated methods, although the manual approach is still being taken, such as the systematic review of the ACM Conference on Computer Supported Cooperative Work (CSCW) by Wallace et al. [36]. The authors reviewed over 1,200 papers published between the years 1990 and 2015, and analyzed data such as publication year, type of empirical research, type of empirical evaluations used, and the systems/technologies involved, to study the changes in research practice over time.

Research topics are usually modeled as semantic word clusters or probability distributions over words inferred from the corpora of documents being analyzed by topic modeling algorithms such as Probabilistic Latent Semantic Analysis (PLSA) [37], Latent Dirichlet Allocation (LDA) [38] or one of their extensions.

2.4.1 Co-citation Analysis

Co-citation frequency measures how frequently two papers are cited together by other papers. The analysis of this measure was proposed by Small [39, 40] as a means of finding clusters of similar documents and researcher networks. This is based on the idea that

frequently co-cited documents contain similar ideas and cover similar topics. Clusters of frequently co-cited papers have been used to create maps of research fields and study their evolution over time [41, 42].

Examples of citation and co-citation analysis include research by Meyer et al. [43] and Hauke et al. [44] who study the Journal of Artificial Societies and Social Simulation (JASSS). They identify the most influential works and authors and show the multidisciplinary nature of the field of social simulation.
Gipp and Beel [45] extended co-citation analysis by introducing the Citation Proximity Index (CPI). This is based on the idea that the closer to each other citations appear in the citing paper, the more similar the cited papers are, e.g. papers cited in the same sentence are more likely to be similar than papers cited in the same paragraph, chapter, etc.

2.4.2 Topic Models

Latent Semantic Analysis (LSA) [46] is an early approach to semantic clustering of text documents by means of a linear projection of the high-dimensional term space onto a low-dimensional vector space by singular value decomposition (SVD) of the document-term matrix. It has been used in the field of information retrieval for automatic document categorization, improving query results (Latent Semantic Indexing) by clustering semantically related documents and in other areas as a tool for dimensionality reduction or noise reduction.

Probabilistic Latent Semantic Analysis (PLSA) is an approach to LSA based on the likelihood principle and defining a generative model for the documents in the training corpus (but not for new documents). Model parameters are learned based on the observed term frequencies by maximizing the log-likelihood function using the Expectation Maximization (EM) algorithm [47]. Aside from it being a proper generative model with a statistical foundation, PLSA deals with polysemous words, i.e. words with different meanings, better than LSA by SVD. However, the parameters of a k-topic model are k multinomial distributions over the vocabulary and one k-topic mixture explicitly tied with each document in the training set. This means that the model cannot generate new documents and cannot be used for prediction on unseen documents.

This problem was addressed by Latent Dirichlet Allocation (LDA) introduced by Blei et al. [38], Blei [48]. Unlike in PLSA, the topic mixture weights are treated as a k-parameter hidden random variable with a Dirichlet distribution. The two main advantages of LDA over PLSA are:

- Ability to generalize to unseen documents,
- The number of parameters does not depend on the corpus size, thus LDA is less prone to overfitting.

Correlated Topic Models (CTM) proposed by Blei and Lafferty [49] are based on LDA, but the Dirichlet distribution of topic proportions has been replaced by the logistic normal distribution. Unlike LDA, which assumes topic independence, CTM directly models correlation between topics. As the authors have shown on a corpus of articles from the journal *Science*, this allows for a better fit than LDA. This approach stems from the observation that in practice some topics are more likely to co-occur in a document than others: "an article about genetics may be likely to also be about health and disease, but unlikely to also be about x-ray astronomy" [49]. Another topic model that models topic correlation is the Pachinko Allocation Model (PAM) by Li and McCallum [50], where correlations between topics are modeled using a directed acyclic graph. The topics are distributions not only over words, but over other topics as well.

Another topic model derived from LDA are Dynamic Topic Models (DTM) conceived by Blei and Lafferty [51] for modeling evolving topics in diachronic corpora. The corpus is divided into time slices and a k-topic model is inferred for each slice, where the topics are derived from the topics for the previous slice. One of DTM's main assumptions is the fixed number of topics (k) present in all time periods. Large changes in topics from one time slice to the next are penalized and topics may not appear or disappear over time.

Wang et al. [52] extended DTM to model the evolution of topics in a chronological sequence of documents using Brownian motion. Continuous Time Dynamic Topic Models (cDTM) do not require the corpus to be divided into discrete time slices.

Wang and McCallum [53] extended the generative model of LDA to also generate continuous-time timestamps. Their Topics Over Time (TOT) model associates a continuous distribution over timestamps with each topic, thereby capturing how the occurrence of static topics changes over time. On-Line LDA by AlSumait et al. [54] is an extension of LDA modeling topics evolving over time in a stream of documents, i.e. an unbounded, potentially infinite sequence. They used Gibbs sampling to discover topics, like Griffiths and Steyvers [55]. The vocabulary does not need to be known upfront but words are never removed. This makes On-line LDA prone to the curse of dimensionality for sufficiently long document streams. The extension of PLSA proposed by Gohr et al. [56] models evolving topics as well as changing vocabulary. A sliding window is applied to a sequence of documents and a PLSA model is built for each window. The model at each

window position is derived from the previous one by removing words and documents no longer present in the current timeframe and adding new ones.

An earlier approach based on PLSA by Mei and Zhai [57] models topics on finite document sequences, assumes that the vocabulary is static and known in advance, and utilizes a model built on the entire corpus in addition to the PLSA models for each timepoint.

An important decision which needs to be made in advance when training topic models is the number of topics, similarly to the number of clusters for clustering models. One common approach is manual selection [58, 59]. Besides the need for expert knowledge of the studied field, its most obvious drawback is the arbitrary choice not based on any objective criteria. Even to an expert the number of topics may not always be a straightforward choice. Also, as is often the case, experts in the same area may have different views and may not always agree even on matters such as the number of research topics which need to be distinguished in their field of study. Reviewers of research papers based on this approach may dispute the authors' choice and may not be easily convinced unless that choice had been made based on an objective and measurable criterion.

Another common approach is choosing the number of topics based on the model's perplexity, or predictive likelihood on a held-out sample [60]. However, as shown by Chang et al. [61], this approach leads to models which are less interpretable by humans. Model quality may also be measured by topic coherence [62, 63]. Röder et al. [64] discuss various measures of coherence and show how they correlate with human interpretability. C_V topic coherence was found to give the best approximation of model understandability. This coherence-based approach to model selection was chosen by Chen et al. [65].

This problem is solved by the Hierarchical Dirichlet Process (HDP) [66] – a nonparametric generalization of LDA able to learn the number of topics from the data. A nonparametric Bayesian prior based on HDP was also proposed for the Pachinko Allocation Model [67]. Nonparametric PAM is able to learn both the number of topics and how they are correlated from the data.

Topic models may be trained on full paper texts [58], titles and abstracts [56, 65], or abstracts only [57]. Using full texts gives the ability to capture all topics mentioned in the papers, but may introduce "noise", especially if the data have not been cleaned sufficiently. This is particularly problematic for old OCR'ed documents, where e.g. page headers and footers are mixed with the documents' contents. On the other hand, using only titles and abstracts ensures that the essence of the papers will be captured, however some finer details such as secondary topics may be lost.

One of the most influential purely latent topic-based approaches is the exploration of trends in the field of computational linguistics by Hall et al. [58], who used an LDA model

trained on ∼12,500 papers from the ACL Anthology published between the years 1965 and 2008. The corpus contains both journal and conference papers from ACL, COLING, and other venues, published with various frequencies. The authors first ran LDA with 100 topics, hand-picked 36 they deemed most relevant and seeded additional 10 to improve coverage of the field. They finally trained a new 100-topic LDA model with using the resulting 46 topics as priors. The final topic model was used to show trends over time, i.e. topics increasing and declining in popularity. Topic popularity was expressed as the probability mass, or the sum of observed probabilities of a given topic for all documents. The authors also answered the following questions about the topics covered by three conferences on commputational linuistics (ACL, COLING and EMNLP): Are the topics of the conferences converging? Is the breadth of topics covered becoming more similar? The former question was answered by comparing the topic distributions for the three conferences using Jensen-Shannon divergence – a measure of similarity of distributions. To answer the latter question, the authors defined a measure of *topic entropy* based on the topic distribution.

More recently Chen et al. [65] studied the evolution of topics in the field of information retrieval (IR). They trained a 5-topic LDA model on a corpus of around 20,000 paper titles and abstracts from *Web of Science*. The number of topics was chosen to maximize topic coherence out of a range of 5-10 topics pre-selected by domain experts. They also divided the corpus into 5-year time intervals and trained additional "local" LDA models on each of those intervals. The authors calculated cosine similarities between the "local" and "global" topics to show topic merging and splitting, knowledge transfer between topics and the state of each topic: from developing to fully mature.

Sun and Yin [59] have used a 50-topic LDA model trained on a corpus of over 17,000 abstracts of research papers on transportation published over a 25-year period to identify research trends by studying the variation of topic distributions over time. They also studied how similar are the topics covered by 22 top tier journals in the field of transportation research.

Mimno et al. [68] trained a Topical N-Grams model [69] (topic model using n-grams instead of words) on 320,000 papers on machine learning and related topics from a corpus collected by the *Rexa* citation indexing system. The authors proposed a number of impact metrics inspired by widely used bibliometric impact measures, but redefined with topic membership instead of publication venue: Topic impact factor, topical diffusion and diversity, topical transfer, etc. Using the *Rexa* corpus they demonstrated how the topic-based metrics may be applied to obtain scientifically meaningful results.

Another interesting example is the paper by Hu et al. [70] where Google's Word2Vec model is used to enhance topic keywords with more complete semantic information,

and topic evolution is analyzed using spatial correlation measures in a semantic space modeled as an urban geographic space.

2.4.3 Topic and Citation Models

A number of topic models have been developed, which exploit the citation structure among papers. Important information about the topical clusters in corpora of papers is contained in the citation link structure, as citing documents are likely to cover the same or similar topics as the cited documents. To capture this information, the models proposed by Cohn and Hofmann [71] and Erosheva et al. [72] extend the generative processes of PLSA and LDA respectively by adding a step generating citation links from a multinomial distribution, similarly to generating words from the vocabulary. The main drawback of these models is the independence of citations and text similarities. This was addressed by Dietz et al. [73] in two models based on LDA, whose generative processes for the citing paper draw words from the topic mixtures of the cited papers. The Inheritance Topic Model proposed by He et al. [74] explicitly divides the citing document into the inherited part – generated from the cited documents – and the autonomous part generated independently. This explicitly models reusing existing ideas in the inherited part and contributing new material in the autonomous part. Nallapati et al. [75] proposed two latent topic models utilizing LDA and PLSA, and inspired by modeling protein-protein interactions, in which the presence or absence of a citation link between each pair of documents is generated from a Bernoulli distribution whose parameter depends on the latent topics in those documents. The main advantage of this approach over the previous models is its ability to predict links for previously unseen papers.

2.4.4 Other Text Mining Approaches

The first 25 years of the SIGIR conference on information retrieval were studied in 2002 by Smeaton et al. [76]. They represented each of the 853 papers as Bag-of-Words vectors derived from the titles, author names and abstracts and calculated a 853×853 document similarity matrix. They then performed hierarchical clustering on the set of papers into 29 clusters which they labeled with descriptions of the common topics in the majority of papers in each cluster. The authors used the resulting 29 "topics" in an analysis to show how they gained and lost popularity over the first 25 years of the conference. They also constructed a co-authorship graph and identified the 5 most central authors, i.e. those with the shortest average path lengths to all other authors in the graph.

Saft and Nissen [77] analyzed papers published in the Journal of Artificial Societies and Social Simulation (JASSS) between the years 1998 and 2013, using a text mining

approach linking documents into thematic clusters in a manner inspired by co-citation analysis. Similarly to a co-citation matrix they built a term co-occurrence matrix reflecting how often terms co-occur in a paper and visualized the discovered term clusters as topics. Like CCA, their method also allows for finding researcher networks – groups of authors working on similar topics.

Pohl and Mottelson [78] analyzed trends in the writing style in papers from the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) published over a 36-year period. They defined text-based features such as readability, title length, novelty or name-dropping, study their changes over time and demonstrated how they correlate with citation count. These measures, however, are defined using simple rules, based on word occurrences, e.g. a paper is marked as novel, if it contains words such as "novel" or "new". It should also be noted, that this study is only about writing style, not research topics covered by the analyzed papers.

2.5 Trend Detection

Trend detection is the task of discovering and predicting positive or negative trends in time series data such as population numbers [79], air pollution [80], stock prices [81] etc. Gray [82] provides a comprehensive description of various approaches to trend detection. They generally fall into one of the following categories:

- Regression methods such as linear regression which fit a line to the data points. The slope of this line is the estimate of the trend (positive or negative),
- Non-parametric rank models, such as Kendall's τ, where the trend is estimated by the rank correlation coefficient,
- Smoothing methods, where a smooth curve is fitted.

Similar methods have also been used on textual data. Färber and Jatowt [83] study emerging trends using the Mann-Kendall test and a linear regression model trained on 76M noun phrases extracted from 90,000 computer science papers from arXiv.org. Their method has been implemented in ScholarSight – a system for visualizing temporal trends [84].

Mane and Börner [85] define topics as clusters of co-occurring words and detect topics bursts using Kleinberg's burst detection algorithm [86]. Small et al. [87] proposed an approach to identifying emerging research topics based on citation networks and co-citation clusters.

Kontostathis et al. [88] – survey of emerging trend detection systems – divides systems into two categories: Fully automatic (corpus as input, list of emerging topics as output), semi-automatic (require user input, produce visualisations etc.).

Jiang et al. [89] propose a novel embedded trend detection framework where key phrases and authors are extracted into a multigraph, phrase vectors are constructed based on individual word embeddings, clustered using the k-means algorithm and fed into a recurrent neural network (RNN) to infer trending topics. Sohrabi and Khalilijafarabad [90] build a similarity graph of document bag-of-words vectors with TF-IDF weighting and identify scientific sub-disciplines by solving the community detection or graph partitioning problem. They utilize the resulting structure to identify trends by calculating the Jaccard similarity index between communities. Their method is able to detect thread birth, growth, decline, merging, splitting and death.

Other approaches involve Latent Dirichlet Topic distributions over time [91] or correlation between the occurrence of terms and citation networks [92].

2.6 Scientific Impact Prediction

Several works have employed machine learning-based approaches to predict citation counts and the long-term scientific impact (LTSI) of research papers, e.g., Yan et al. [93, 94] or Singh et al. [9].

2.7 Identifying Breakthroughs

Traditional scientometrics focuses on citation data to identify influential or "breakthrough" publications and authors. According to a number of studies in the 1970s [95], 1980s [96], and 1990s [97], where several most-cited rankings were reviewed, citation impact is a good indicator of scientific excellence and predictor of Nobel prize awards. The problem with most citation-based approaches, however, is that they are only able to identify breakthroughs retrospectively [4]. Several approaches to early identification of potential breakthroughs have been researched in the 21st century. A distinction also needs to be made between journal impact, researcher impact and individual paper impact.

2.7.1 Citation-based

Schneider and Costas [3, 4] proposed three citation-based approaches, two of which use the Characteristics Scores and Scales (CSS) method by Schubert et al. [98] to divide papers into four classes with respect to citation counts: low, moderate, high, and outstanding. In contrast to traditional percentile approach, CSS iteratively bisects the set of papers into subsets having a number of citation below and above the mean. This process is repeated three times, each subsequent time dividing the subset above the current mean. The four resulting classes are bounded by the three calculated means.

Schneider and Costas [3] define "breakthrough paper" as "a highly cited paper, with an important spread over its own field(s) and also other fields of science, and it must be a paper that is not a mere follower of other highly cited publication(s) but that it has a genuine relevance on its own". They argue that merely having many citations is not sufficient for a paper to be considered a breakthrough, as the high citation count may result e.g. from following a previous breakthrough.

To filter out followers they analyze the citations of the candidate papers citing other breakthrough publications. For each cited paper, the citing paper is considered its follower if the number of times it is cited on its own, i.e. not co-cited with the previous breakthrough paper, does not exceed a certain arbitrarily chosen threshold.

The three proposed approaches are the following:

- Divide the corpus into "micro fields" and perform the follower test on the most cited paper in each micro field,
- Identify candidate papers using CSS and filter out followers,
- Identify candidate papers using CSS, filter out followers, and select the papers with higher-than-average impact on other fields.

Ponomarev et al. [5, 99] proposed a method for early detection of potential breakthrough papers utilizing early citation dynamics. They analyzed a corpus of over 375,000 papers on biochemistry and molecular biology to find typical time-dependent patterns of highly-cited publications. Based on the assumption, that 5 years is a sufficiently long time for seminal papers to be discovered, they pre-selected candidate top-cited papers at 5 years from publication in each subject category and studied their monthly citation patterns during the first 5 years. Three types of patterns were identified. Using citation data for the first 6, 12, or 24 months after publication, curves were fitted and the number of citations at the five year mark was then predicted by extrapolation. The downside of

this approach is the inability to predict "sleeping beauties" – papers which receive high citation numbers later.

Wolcott et al. [6] trained a Random Forest classifier on a corpus of papers indexed by the *Web of Science* using a number of publication-level bibliographic, citation and altmetric features (citation numbers and velocity, number of authors, number of countries associated with the authors, number of pages, etc.) as well as author-level features such as H-index, number of publications etc. Breakthrough papers in the training set were selected manually. The classifier was then used to determine the most important features. The most important features were found to be time-dependent (citation counts and velocity). Because the goal is the early identification of potential breakthroughs, another Random Forest model was trained using time-independent features only.

Winnink and Tijssen [100] showed using the paper on graphene by Novoselov et al. [101] as an example, that bibliographic data contain information enabling the identification of potential breakthroughs as early as two years after publication.

2.7.2 Analogy Mining

Analogy in science as a driver for innovation has been studied for over half a century [102]. A notable example of advancement in one domain inspired by another, distant one, is the simulated annealing optimization algorithm inspired by the annealing process commonly used in metallurgy [103]. Large sources of potential analogies are widely available – corpora of scientific publications, patent databases, etc., however finding analogies in such large unstructured datasets is a difficult task. Traditional information retrieval approaches like Latent Semantic Indexing [46] or Latent Dirichlet Allocation [38] address this problem, but they are only able to identify "surface" similarity at the level of topics. Conversely, deeper "conceptual" analogies useful in this context, may often be missed [104]. Approaches based on rich semantic structures [105], on the other hand, are prohibitively expensive on such datasets, since they require considerable human effort.

Chan et al. [106] trained an analogy detecting deep learning model on a dataset obtained by crowdsourcing. Workers on the crowdsourcing platform Amazon Mechanical Turk[15] were given the task of finding analogies for sample product descriptions from the invention platform Quirky[16]. The identified positive cases and negative cases – either explicitly rejected or implicitly ignored query results – were used as a training set for a Convolutional Neural Network. The queries used to find the positive examples were also

[15] `https://www.mturk.com/`
[16] `https://quirky.com/`

recorded and incorporated in the model as the semantic links, or concepts linking the identified pairs of analogous documents.

The same group of researchers approached the same problem differently in their subsequent paper [107]. Instead of tasking crowdsourcing workers with explicitly finding analogies between products from Quirky, they only obtained annotations for *purpose* (i.e. what problem the product solves) and *mechanism* (i.e. how it does it) through crowdsourcing. Using word embeddings [108] and Recurrent Neural Networks they constructed purpose and mechanism vectors comparable by vector space distance metrics such as cosine similarity. This representation was then used to find pairs of products serving the same purpose in different ways, or using a similar mechanism to solve different problems. This research led to the SOLVENT system for finding analogies between research papers across domains [104].

Identifying interdisciplinary ideas as a driver for innovation was also studied by Thorleuchter and Van den Poel [109].

2.8 Detecting *Sleeping Beauties*

The term *Sleeping Beauty* coined by Van Raan [110] is commonly used to describe papers so "ahead of their time", that they go unnoticed, often for many years, before attracting significant attention. Notable examples are the 1865 paper on plant genetics by Gregor Mendel, or the Einstein-Podolsky-Rosen "paradox" paper from 1935. Van Raan [110] defined bibliometric measures characterizing Sleeping Beauties, such as depth of sleep, sleeping time, and awakening intensity. They studied a set of 20 million papers, and gave an example of a Sleeping Beauty and her prince – a paper which "awakens" her, or first cites the SB after a long sleeping period, after which more citations follow.

Ke et al. [111] introduced an objective metric – the *Beauty Coefficient* – measuring to what extent a paper is a Sleeping Beauty. Its value increases as the length of sleep and awakening intensity increase. It also penalizes early citations. Unlike previously proposed measures, the Beauty Coefficient does not use arbitrary parameters or thresholds.

2.9 Document Dating

Another related field of research is document dating (timestamping), or predicting document creation dates based on their textual content. Typical approaches relevant to this dissertation are based on changes in word usage and on language change over time.

Statistical language models – widely used in speech recognition, text classification, word prediction and other NLP tasks – model natural language by assigning a probability to each word (unigram) or word sequence (n-gram) based on their observed frequencies in a language corpus. A temporal language model is a time series of statistical language models reflecting changes in word or phrase usage over time. De Jong et al. [112] show how such models may be used for document dating: A temporal language model is built based on a reference corpus of texts from the same domain as the texts to be dated, published over a period of time. Document creation dates are then predicted by comparing the language model of the document to the models of the time partitions and selecting the best fit. Statistical language models are compared by various metrics used for comparing probability distributions, such as the Kullback-Leibler divergence [113]. Another example of a similar approach may be found in [114]. Kanhabua and Nørvåg [115] improved upon this approach by adding a semantic preprocessing step and implemented a document timestamping tool with a web-based user interface [116].

Jatowt and Campos [117] have implemented an online visual and interactive system based on n-gram frequency analysis. N-gram frequencies in the analyzed documents are compared to the distributions in the Google Books NGram[17] corpus. The tool facilitates the interpretation of prediction results by generating age probability plots and providing evidence such as top contributing n-grams.

Classification models have been utilized in document dating by a number of researchers. Garcia-Fernandez et al. [118] have used Support Vector Machine (SVM) classifiers on feature vectors of word and n-gram frequencies from the Google Books NGrams corpus and named entity occurrences. Other classification models such as Random Forest have been used [119], as well as diverse feature sets. Salaberri et al. [120] use diachronic word frequencies from the Google Books NGrams corpus (1-grams), and also features capturing changes over time in orthography, semantics, lexicon, syntax and morphology. They also used named entity occurrences and detected year entities as features. Niculae et al. [121] also use SVM classifiers, however their approach is different to the multiclass classification approach in earlier works. Because publication dates are ordinal values rather than categorical ones, they expressed the problem of predicting document creation dates as an *Ordinal Regression* or *Learning-to-Rank* task, where publication dates are viewed as ranks. A series of *before-after* binary classifiers are trained and the most likely publication date is selected as the model's final prediction. Popescu and Strapparava [122] is another example of an ordinal regression approach.

[17] `https://books.google.com/ngrams`

Examples of research articles based on heuristic methods include: Garcia-Fernandez et al. [118], Kumar et al. [123], Kotsakos et al. [124] or [122]. These approaches, however are less relevant to the topic of this work.

Another approach to temporal language modeling are neural language models based on word embeddings such as Word2Vec [108]. Kim et al. [125] study the shift in word semantics over time by training a model for each time interval and then plotting the words' cosine similarities to their reference points.

2.10 Paper Recommendation

Research paper recommender systems date back to 1998 and the search engine *CiteSeer*[18] [126, 127]. This work was motivated by the problems posed to researchers by the rapid growth of the amount of published literature and its poor organization on the World Wide Web, where scientific papers are often available only in non-text formats such as *PostScript* or *PDF*. *CiteSeer* originally recommended papers similar to the search results by using a similarity measure based on TF-IDF, the LikeIt string distance [128] and common citations extracted from the parsed papers retrieved from the Web and stored in the database.

Beel et al. [129] identified the following classes of research paper recommender systems in their metaanalysis:

- Stereotyping: Recommendations are made based on certain characteristics of the user, assuming that users exhibiting similar traits are likely to be interested in similar content. Beel et al. [130], Beel [131] use mind maps as a user modeling tool.
- Content-based Filtering: Recommending papers similar to other papers searched or viewed by the user. Various feature sets have been used to determine paper similarity, e.g. words or n-grams [132], citations [127], latent topics obtained through Latent Dirichlet Allocation (LDA) [133], or topics as word combinations [134].
- Collaborative Filtering: Pioneered by Resnick et al. [135], this approach is „based on the heuristic that people who agreed in the past will probably agree again". Systems of this class recommend items highly rated by other people who have rated other items similarly. Examples of research paper recommender systems based on Collaborative Filtering include: McNee et al. [136], Pennock et al. [137], or Vellino [138].

[18] Replaced in 2008 by CiteSeer^{X}: `http://citeseerx.ist.psu.edu/`

- Co-occurrence recommendations: Similarly to recommendations of items frequently bought together on e-commerce sites (frequent itemsets), in this approach recommendations are made based on the co-occurrence of papers in certain contexts, e.g. co-citation [45], co-viewing [139, 140], or co-download [140].
- Graph based approaches, where a graph is built with papers at its vertices and various connections as edges, e.g. citations [141–143], authors [144–146], or venue [141, 145, 146]. Popular papers are then found in the graph using e.g. random walks [145].
- Global Relevance: Ranking papers based on measures not pertaining to a specific user, such as PageRank [133], Katz metric [147], citation count [133, 147], h-index [133], and others.
- Hybrid: Systems combining several of the approaches described above, e.g. TechLens [148].

Examples of recent research on paper recommender systems include Kanakia et al. [149] or Maake et al. [150]. The former paper describes the large scale recommender system used by Microsoft Academic[19]. It is a hybrid of the content-based filtering and co-citation approaches. The content-based component uses paper embeddings constructed as linear combinations of Word2Vec [108] embeddings of the words from the title, keywords and abstract. The measure of paper similarity used is the cosine distance.

The latter shows how to find relevant papers from diverse domains by finding latent relationships between those domains through common terms and concepts.

The innovation score proposed in this work may be used alongside other measures such as citation count to refine the results returned by existing paper recommender systems.

2.11 Keyword Extraction

Keyword or keyphrase extraction is the task of identifying words or phrases summarizing and characterizing the contents of a document. Typical approaches may be divided into two broad categories: Supervised, where an external source of knowledge such as a training corpus, dictionary, or thesaurus is needed to build a model, and unsupervised which can be applied to individual documents without the need to train a model.

Examples of supervised approaches include Turney [151], Witten et al. [152] utilizing Naïve Bayes, or Zhang et al. [153] based on Support Vector Machines. Mihalcea and

[19] `https://academic.microsoft.com/`

Tarau [154], Wan and Xiao [155], Rose et al. [156], or Boudin [157] are unsupervised approaches utilizing graph-based ranking algorithms on word co-occurrence graphs. Campos et al. [158, 159] developed an unsupervised text feature and heuristic-based multilingual keyword extraction system. Their method combines several features to calculate a score for each keyword candidate. These features include: Word frequency, the number of times a word starts with an uppercase letter, how far from the beginning of the document a word occurs, or the number of different terms appearing in proximity to all instances of the candidate word.

Latard et al. [160] have constructed topics from semantic clusters of extracted keywords, which they used to categorize scientific articles. These topics may be used as an alternative to LDA or CTM. However, the method which is the subject of this dissertation uses CTM since its topics are more general than topics defined by semantically clustered keywords. This also makes it better suited for analyzing the evolution of ideas, as keywords may have different meanings in different contexts and conversely, different keywords may describe the same concept.

Pavel Savov, Adam Jatowt, and Radoslaw Nielek. Identifying breakthrough scientific papers. *Information Processing & Management*, 57(2):102168, 2020.

Chapter 3

Identifying Breakthrough Scientific Papers

Citation analysis does not tell the whole story about the innovativeness of scientific papers. Works by prominent authors tend to receive disproportionately many citations, while publications by less well-known researchers covering the same topics may not attract as much attention. In this paper we address the shortcomings of traditional scientometric approaches by proposing a novel method that utilizes a classifier for predicting publication years based on latent topic distributions. We then calculate real-number innovation scores used to identify potential breakthrough papers and turnaround years. The proposed approach can complement existing citation-based measures of article importance and author contribution analysis; it opens as well novel research direction for time-based, innovation-centered research scientific output evaluation. In our experiments, we focus on two corpora of research papers published over several decades at two well-established conferences: The World Wide Web Conference (WWW) and the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), containing around 3,500 documents in total. We indicate significant years and demonstrate examples of highly-ranked papers, thus providing a novel insight on the evolution of the two conferences. Finally, we compare our results to citation analysis and discuss how our approach may complement traditional scientometrics.

3.1 Introduction

Scientometrics – measuring science – in its current shape and form, dates back to the mid 20^{th} century and, in particular, to works such as "*Citation Indexes for Science*" by Eugene Garfield [1], or "*Little Science, Big Science*" by Derek John de Solla Price [2].

The main focus of scientometrics is on measuring the innovation and impact of scientific publications, their authors and publication venues. The most important academic journal in the field is *Scientometrics* founded in 1978 by Tibor Braun. Aside from providing better understanding of the evolution of particular study fields, the identification of innovative or potential breakthrough publications also serves a practical purpose – it helps research funding bodies select the most promising projects to invest in. Recent research in this area includes works such as: Schneider and Costas [3, 4], Ponomarev et al. [5], or Wolcott et al. [6].

Traditionally, citation analysis has been used to identify pioneering scientific papers. This approach, however, suffers from various biases. Works by well-known authors and/or ones published at well-established publication venues (similar to the rich-get-richer effect) tend to receive more attention than others. Papers could then achieve increased visibility through early citations [7]. Widely cited papers also tend to attract even more citations, while some innovative ideas may be appearing in papers well before their popularity time, hence, receiving little recognition.

In this paper we demonstrate a simple, and yet novel, machine learning-based method of analyzing corpora of scholarly papers published over a period of time. The aim is to find the answers to the following questions:

- Which are the most innovative papers, i.e. papers covering topics that would have been researched in the future?
- Which are the *breakthrough* years, i.e. years when trends that were later researched for several years were started?
- Can we offer a supplementary approach to traditional citation analysis for assessing the merit of publications?

As described in detail in Section 3.4, our approach is based on predicting publication years by means of a Support Vector Machine classifier using LDA topics as features. The main idea behind it is that *the more a paper's topic distribution resembles that of papers published in the future (and the less it resembles the past ones), the more innovative the paper is.* Therefore, the novel innovation score we propose is based on how much the predicted publication year is ahead of or behind the actual publication year, which reflects whether the paper covers more topics researched by papers published in the past or more of its topics are covered by future papers.

We show the results of applying our method to the corpora of papers published until the year 2017 at two well-known and influential conferences: The International World Wide

Web Conference (WWW) and the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Both conferences are top-tier venues in their respective fields – among the top 4% according to the CORE conference ranking[1].

The International World Wide Web Conference (WWW) was first held in May 1994 at CERN in Geneva, Switzerland by Robert Cailliau – one of the founders of the World Wide Web. It was held once more in October 1994, twice in the spring and autumn of 1995 and once a year ever since by the International World Wide Web Conference Committee (IW3C2)[2] founded by Robert Cailliau and Joseph Hardin. The conference has served as an important publication venue influencing many researches that center around diverse aspects of the Web.

The SIGIR conference has been held annually by the Association for Computing Machinery's Special Interest Group on Information Retrieval since 1978. It is considered the most important conference in the field of information retrieval. Research areas covered by SIGIR include: document representation, content analysis, query analysis, content recommendation, social media analysis, etc. In addition to the usual Best Paper awards, the *Test of Time Award* has been awarded since 2014 to papers published 10-12 years before that have had "long-lasting influence, including impact on a subarea of information retrieval research, across subareas of information retrieval research, and outside of the information retrieval research community"[3]. Every three years, a researcher is also awarded the *Gerard Salton Award* for "...significant, sustained and continuing contributions to research in information retrieval"[4].

In this work we make the following contributions: We propose a novel method for analyzing the popularity of research topics over time in fields of study, for which exists a significant body of work that spans multiple years. Based on this topic analysis, we predict the publication years of research papers in the analyzed fields of study, and assign innovation scores measuring how far ahead of (or behind) their time are the topics appearing in the paper in question. By aggregating the innovation scores for all papers published in every year, we then calculate importance scores for each year and identify breakthrough years, in which topics popular in the future were first researched. We also include lists of topics extracted from both analyzed corpora with data on their popularity over time, and lists of papers identified as the most innovative.

The remainder of this paper is structured as follows. In Section 3.2 we outline the shortcomings of citation analysis as the sole measure of innovation. We also reference other research focusing on analyzing the evolution of a field of study. In Section 3.3 we

[1] `http://portal.core.edu.au/conf-ranks/`
[2] `http://www.iw3c2.org/`
[3] `http://sigir.org/awards/test-of-time-awards/`
[4] `http://sigir.org/awards/gerard-salton-awards/`

describe the datasets we have used in this work and their preprocessing. Section 3.4 covers the details of our approach to measuring innovation. We explain the construction and evaluation of the topic model and how we use classifiers with latent topics as features to predict publication years. We also define a simple measure of paper innovativeness, and how it can be adjusted to make papers published in different years comparable. Finally, in Sections 3.5 and 3.6 we present the outcomes of applying our method to corpora of research papers from two influential conferences, we discuss its strengths and weaknesses, and show how it can be used to complement citation analysis in identifying potential breakthrough papers. The last section concludes the paper.

3.2 Related Work

The development of research areas and the evolution of topics in academic conferences and journals over time have been investigated by numerous researchers. For example, Meyer et al. [43] study the Journal of Artificial Societies and Social Simulation (JASSS) by means of citation and co-citation analysis. They identify the most influential works and authors and show the multidisciplinary nature of the field. Saft and Nissen [77] also analyze JASSS, but they use a text mining approach linking documents into thematic clusters in a manner inspired by co-citation analysis. Like CCA, their method also allows for finding researcher networks – groups of authors working on similar topics. Wallace et al. [36] study trends in the ACM Conference on Computer Supported Cooperative Work (CSCW). They took over 1,200 papers published between the years 1990 and 2015, and they analyzed data such as publication year, type of empirical research, type of empirical evaluations used, and the systems/technologies involved. Pohl and Mottelson [78] analyze trends in the writing style in papers from the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) published over a 36-year period. They define measures such as readability, novelty or name-dropping, study their changes over time and demonstrate how they correlate with citation count. These measures, however, are defined using simple rules, based on word occurrences, e.g., a paper is marked as novel, if it contains words such as “novel” or “new”.

A different approach to identify novelty was proposed by Chan et al. [104]. They developed a system for finding analogies between research papers, based on the premise that “scientific discoveries are often driven by finding analogies in distant domains”. One of the examples given is the simulated annealing optimization algorithm inspired by the annealing process commonly used in metallurgy. Identifying interdisciplinary ideas as a driver for innovation was also studied by Thorleuchter and Van den Poel [109].

Several works have employed machine learning-based approaches to predict citation counts and the long-term scientific impact (LTSI) of research papers, e.g., Yan et al. [93, 94] or Singh et al. [9].

Examples of topic-based approaches include Hall et al. [58]. They trained an LDA model on the ACL Anthology, and showed trends over time like topics increasing and declining in popularity. Unlike our approach, they hand-picked topics from the generated model and manually seeded 10 more topics to improve field coverage. More recently Chen et al. [65] studied the evolution of topics in the field of information retrieval (IR). They trained a 5-topic LDA model on a corpus of around 20,000 papers from *Web of Science*. Sun and Yin [59] have used a 50-topic LDA model trained on a corpus of over 17,000 abstracts of research papers on transportation published over a 25-year period to identify research trends by studying the variation of topic distributions over time. Another interesting example is the paper by Hu et al. [70] where Google's Word2Vec model is used to enhance topic keywords with more complete semantic information, and topic evolution is analyzed using spatial correlation measures in a semantic space modeled as an urban geographic space.

Färber and Jatowt [83] study emerging trends using the Mann-Kendall test and a linear regression model trained on 76M noun phrases extracted from 90,000 computer science papers from arXiv.org. Jiang et al. [89] propose a novel embedded trend detection framework where key phrase and authors are extracted into a multigraph, phrase vectors are constructed based on individual word embeddings, clustered using the k-means algorithm and fed into a recurrent neural network (RNN) to infer trending topics. Sohrabi and Khalilijafarabad [90] build a similarity graph of document bag-of-words vectors with TF-IDF weighting and identify scientific sub-disciplines by solving the community detection or graph partitioning problem. They utilize the resulting structure to identify trends by calculating the Jaccard similarity index between communities. Their method is able to detect thread birth, growth, decline, merging, splitting and death.

To the best of our knowledge, no prior works have employed a classifier-based approach similar to ours. We are also not aware of any analytical research on the evolution of topics of the WWW conference. The first 25 years of the SIGIR conference were studied in 2002 by Smeaton et al. [76]. They performed hierarchical clustering on the set of papers into 29 clusters they then labeled with descriptions of the common topics in the majority of papers in each cluster. The authors then analyzed how the topics gained and lost popularity over the first 25 years of SIGIR. They also identified the 5 most central authors, i.e. those with shortest average path length to all other authors in the co-authorship graph. We note that all of those authors (Chris Buckley, Gerard Salton, James Allan, Clement Yu and Amit Singhal) were found by our method to have

co-authored some of the most innovative papers in 1985, 1991, 1995, 1996, 1997 and 1998.

One of our main motivations is providing a method complementing citation analysis in identifying pioneering papers. Problems with citation analysis include:

- Citing prominent publications, following the crowd [7]
- Matthew Effect – term inspired by the biblical Gospel of Matthew; according to Merton [8], who first described this phenomenon in 1968, publications by more eminent researchers will receive disproportionately more recognition than similar works by less-well known authors.
- Increased visibility through early citations: Singh et al. [9] have shown that papers cited within two years of publication tend to attract more citations. They have observed, however, that early citations by influential authors negatively impact the cited paper's long term scientific impact by way of *attention stealing* where the subsequently published citing paper authored by the more influential and well-known researcher collects further citations instead of the original work.
- Google Scholar Effect: Serenko and Dumay [10] observed that old citation classics keep getting cited because they appear among the top results in Google Scholar, and are automatically assumed as credible. Some authors also assume that reviewers expect to see those classics referenced in the submitted paper regardless of their relevance to the work being submitted.
- Self-citations: Increased citation count does not reflect the work's impact on the field of study.
- Ignoring the purpose of citations (support vs criticism)
- Erroneous citations

In our previous work, which appeared as a poster at the WWW 2017 conference [161], we have introduced the preliminary idea behind the method covered in the current paper. It used a Latent Dirichlet Allocation (LDA) model [38] and Support Vector Machine (SVM) classifier to predict publication years based on the latent topic distribution. We then used prediction errors to measure how innovative papers are and calculated year importance scores based on the mean prediction error for papers published in each year. The idea behind that approach was that the more papers published in year y were predicted as published in the future, especially in the distant future, the higher y's score is, and thus, the more we consider y to be important. We used that method to analyze over 3,000

papers published in the proceedings of the WWW conference between the years 1986 and 2016 and show some research trends spanning several years. In this paper we take a similar approach but focus on measuring the innovativeness of individual publications and propose a real-number innovation score based on the classifier's prediction. We also use our method to analyze corpora of papers from two well-established conferences with long histories: World Wide Web and SIGIR.

Research on document dating (timestamping) is related to our work, too. Typical approaches to document dating are based on changes in word usage and on language change over time, and they use features derived from temporal language models [112, 115, 116], diachronic word frequencies [119, 120], or occurrences of named entities. Examples of research articles based on heuristic methods include: Garcia-Fernandez et al. [118], Kotsakos et al. [124] or Kumar et al. [123]. Jatowt and Campos [117] have implemented the visual and interactive system based on n-gram frequency analysis. In our work we rely on predicting publication dates to determine paper innovativeness. Classifiers trained on topic vectors are a variation of temporal language models and reflect vocabulary change over time. Aside from providing means for timestamping, they also allow for studying how new ideas emerge, gain and lose popularity.

Another field of research related to ours is paper recommendation. Recently Beel et al. [129] conducted a comprehensive literature review on the subject of research paper recommender systems. Examples of different approaches to paper recommendation include Zhao et al. [162] or Raamkumar et al. [163]. The innovation score we propose in this work may be used alongside citation counts to refine the results returned by existing paper recommender systems.

Finally, worth mentioning is research on keyword extraction. Latard et al. [160] have constructed topics from semantic clusters of extracted keywords, which they used to categorize scientific articles. These topics may be used as an alternative to LDA. However, we chose LDA since its topics are more general than topics defined by semantically clustered keywords. This also makes it better suited for analyzing the evolution of ideas, as keywords may have different meanings in different contexts and conversely, different keywords may describe the same concept.

3.3 Datasets

Both datasets for the two conferences that we study contain full papers, short papers and poster abstracts published until the year 2017. The WWW corpus contains in total 3,056 papers published between the years 1994 and 2017. Papers from the 3rd International

World Wide Web Conference held in April 1995 in Darmstadt, Germany are missing from the corpus due to unavailability of the proceedings. The SIGIR corpus contains 3,434 papers published between the years 1978 and 2017.

Most papers from both the conferences were published as PDF documents. They have been converted to plain text using the *pdftotext* tool[5]. Older, pre-2000 texts may contain more errors than newer ones due to OCR inaccuracy. We note that proceedings from the World Wide Web Conference until the year 2000 were published as HTML pages.

The following preprocessing steps have been performed on all papers prior to training the LDA models and classifiers:

- Convert to lower case
- Remove punctuation and numbers, including ones spelled out, e.g. "one", "two", "first" etc.
- Remove stopwords using the standard English stopword set in NLTK
- Detect Part-of-Speech tags using the Penn Treebank POS tagger implemented in NLTK and lemmatize using the WordNet Lemmatizer in NLTK [164]

WordNet Lemmatizer is a lemmatizer based on WordNet – a lexical database of English [165]. It maps different inflected forms of a word to its *lemma* or base form taking into account its POS tag, e.g. $(\text{“}is\text{”}, V), (\text{“}are\text{”}, V), (\text{“}was\text{”}, V) \rightarrow (\text{“}be\text{”}, V)$.

The POS tags were only used by the lemmatizer and were discarded afterwards.

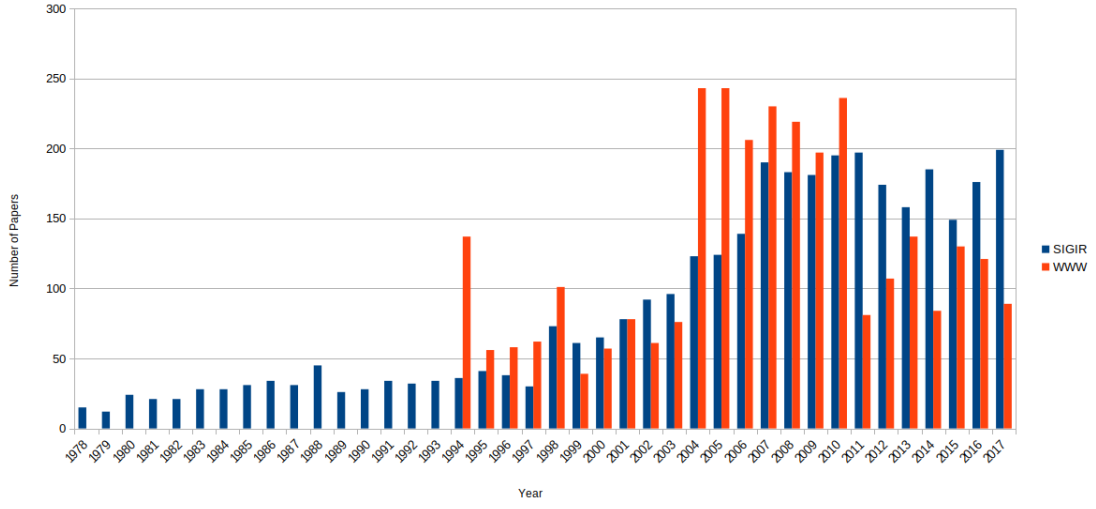

FIGURE 3.1: Number of papers per year

[5] `https://www.xpdfreader.com/pdftotext-man.html`

3.4 Methodology

The following section contains the detailed description of our approach. In Section 3.4.1 we describe the training of our topic model and how we chose the best model for our use case. We also explain why we chose LDA over other topic modeling algorithms. In Section 3.4.2 we provide an overview of the topic quality measure we have used for model selection - C_V Coherence and the rationale behind its use. Section 3.4.3 covers the training of a machine learning model for predicting publication years. Finally, in Sections 3.4.4 and 3.4.5 we define measures of paper innovativeness and year importance.

3.4.1 Topic Model

Using a topic model lets us achieve two goals: A vast dimensionality reduction of the publication year prediction problem, and – more importantly – a means of understanding how research areas evolve over time.

Dynamic Topic Models [51] are often used to model topics evolving over time. However, one of DTM's main assumptions is the fixed number of topics present in all time periods. Large changes in topics from year to year are penalized. Because we need the ability to model topics which may appear and disappear over time, we chose to use Latent Dirichlet Allocation [38] instead. LDA is the suitable topic model for our use case as it makes no assumptions about time. DTM was also considered but ultimately rejected by Hall et al. [58].

Selecting the number of topics k in an LDA model is a decision that needs to be made upfront, and there exists no universally agreed formula. Hagen [166] suggests a two-step approach: Pre-selecting a small number of models from a wider range of values of k based on perplexity and having each of them evaluated by experts. Manual evaluation is labor-intensive, we have therefore decided to use C_V topic coherence – a measure introduced by Röder et al. [64] as the best approximation of human topic interpretability. To select the optimal topic model, we have trained k-topic LDA models for each k between 10 and 70. We then chose the model with the highest mean value of topic coherence C_V – 43 topics. Because topic quality is more important than outright prediction accuracy for the purpose of understanding the evolution of the conferences, we chose the best topic models by maximizing C_V coherence rather than by minimizing prediction error or model perplexity. According to Chang et al. [61] model selection based on topic coherence produces models more understandable to humans than traditional likelihood-based approaches. In fact they have shown predictive likelihood to be

negatively correlated to topic understandability and thus real-world usability. A similar coherence-based approach to topic number selection was taken by Chen et al. [65].

In order to measure how much the choice of the number of topics will affect the prediction of publication years and consequently the *paper innovation scores* defined later in Section 3.4.4 (Equation 3.2), we have calculated scores for all papers in our corpora using each k-topic LDA model for all k between 10 and 70. We then calculated Spearman's rank correlation coefficients between paper scores for each pair of LDA models. For the WWW corpus the mean correlation coefficient was 0.75 with a standard deviation of 0.04. For the SIGIR corpus the mean correlation coefficient was 0.65 and the standard deviation: 0.05. The scores calculated using each LDA model are, therefore, strongly correlated, which supports our reasoning that choosing the right LDA model is more important for understanding the outcome than for predicting publication years and paper innovation scores.

3.4.2 C_V Topic Coherence

C_V Coherence has been described in detail by Röder et al. [64]. It is based on the *boolean sliding window* approach and *normalized pointwise mutual information* (NPMI), and it uses the inverse cosine confirmation measure.

Its calculation starts with constructing *context vectors* $\vec{v_i}$ for each word w_i in the top n words of each topic. The j-th element of vector $\vec{v_i}$ is given by the equation:

$$v_{ij} = NPMI(w_i, w_j) = \frac{log\frac{P(w_i,w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-log(P(w_i, w_j) + \epsilon)}$$

where $P(w)$ and $P(w_i, w_j)$ are the observed probabilities of word w's occurrence and of words' w_i and w_j co-occurrence in a *virtual document*, respectively. Each of these virtual documents is defined by a step of the sliding window moving through all documents in the corpus one token at a time. Intuitively speaking, the sliding window captures proximity between top topic-defining words – how often and how close to each other they co-occur in the corpus. ϵ is an arbitrarily chosen small constant added to the term $P(w_i, w_j)$ to avoid the logarithm of zero in case w_i and w_j do not co-occur within the sliding window anywhere in the corpus.

The sliding window size and n – the number of top topic words are parameters of the algorithm. We chose the values 110 and 10, respectively. As shown by Röder et al. [64], the correlation to human ratings is the highest for these values.

The next step is the calculation of *cosine similarity* for each pair of context vectors $(\vec{u}, \vec{v})$:

$$s_{cos}(\vec{u}, \vec{v}) = \frac{\sum_{i=1}^{|W|} u_i \cdot v_i}{||\vec{u}||_2 \cdot ||\vec{v}||_2}$$

The main advantage of the indirect cosine confirmation measure is its ability to capture semantic similarity between words that do not necessarily co-occur often but make sense in the same context, e.g. synonyms.

Finally C_V Coherence is calculated as the arithmetic mean of all cosine similarities of words belonging to the given topic:

$$C_V = \mu(\{s_{cos}(\vec{u}, \vec{w}) \mid \vec{u}, \vec{w} \in W\})$$

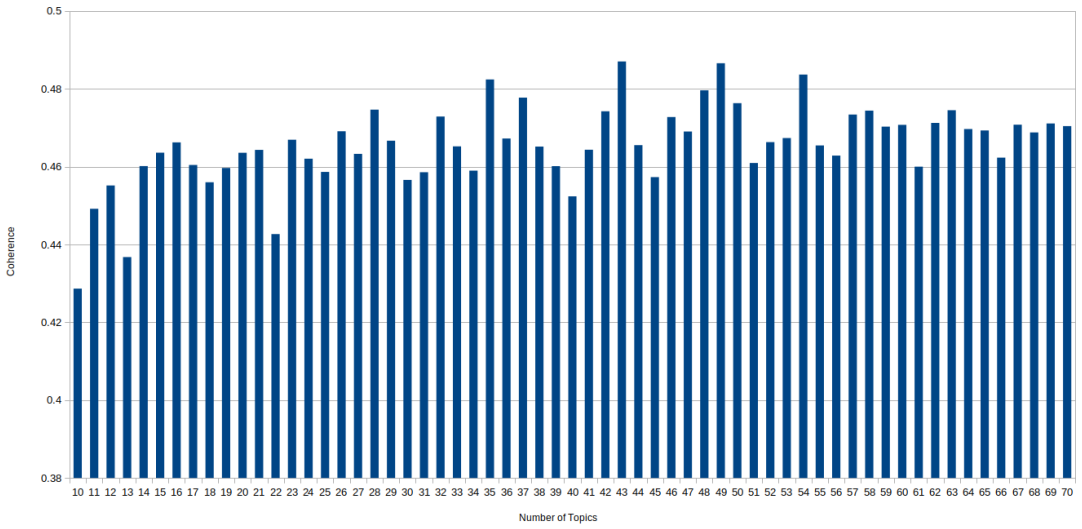

FIGURE 3.2: C_V topic coherence by number of topics

3.4.3 Predicting Publication Years

To predict publication years we used multiclass linear SVC classifiers from the *scikit-learn* Python machine learning library [167] where each class corresponds to a year from 1978 to 2017 (SIGIR) or from 1994 to 2017 (WWW). Multiclass functionality is achieved by training $n(n-1)/2$ *one-vs-one* classifiers where n is the number of classes. Topic probability distributions were used as feature vectors representing the papers. The maximum number of iterations was set to -1 (unlimited) and the class weight was set to "balanced", since the number of samples in each class (i.e. papers published in each year) varies widely, as shown in Figure 3.1. Due to the relatively small sizes of the corpora, instead of dividing the corpora into training and test sets, each paper was scored by a separate classifier trained on all remaining papers.

Figure 3.6 shows the resulting confusion matrices for SIGIR and WWW, respectively as heat maps where darker shades of red represent greater numbers, paler shades of yellow represent smaller numbers and white denotes zero.

3.4.4 Breakthrough Papers

We consider a paper to be innovative, if it covers topics that will be popular in the future, counting from the publication date of the paper, especially, in the distant future, but which have not been popular in the past, especially in the distant past.

Let us define the innovation score of paper p as:

$$S_P(p) = \frac{\sum_y conf(p,y) \cdot (y - Y_p)}{\sum_y conf(p,y)} = \sum_y conf(p,y) \cdot (y - Y_p) \tag{3.1}$$

where Y_p is the year paper p was published and $conf(p,y)$ is the classifier confidence for paper p and year y expressed as class membership probability calculated using the multiclass extension of Platt scaling proposed by Wu et al. [168], i.e. the estimated likelihood that paper p was published in year y. The sum of all $conf(p,y)$ for a given p is, therefore, equal to 1 and we may omit the denominator.

Platt scaling [169] is a logistic transformation of SVM scores whose purpose is to make the classifier scores interpretable as class membership probabilities. For a binary classification problem where inputs x are labeled 1 and -1 it is defined by the equation

$$P(y = 1 \mid x) = \frac{1}{1 + exp(Af(x) + B)}$$

where y is the predicted label, $f(x)$ is the SVM output, and A and B are parameters learned by the algorithm.

The innovation score in Equation 3.1 is, thus, defined as the weighted mean classification error where classifier confidences are the weights and their sum is 1. A positive value of $S_P(p)$ means paper p covers more topics covered by papers published in the future than in the past and negative values mean the opposite – that paper p covers more topics popular in the past. It should be noted, however, that only the innovation scores of papers published in the same year may be compared directly, since the classifier's error is more likely to be positive for papers published in earlier years and negative for papers published in later years. Later in this section we will show how they can be adjusted to account for this.

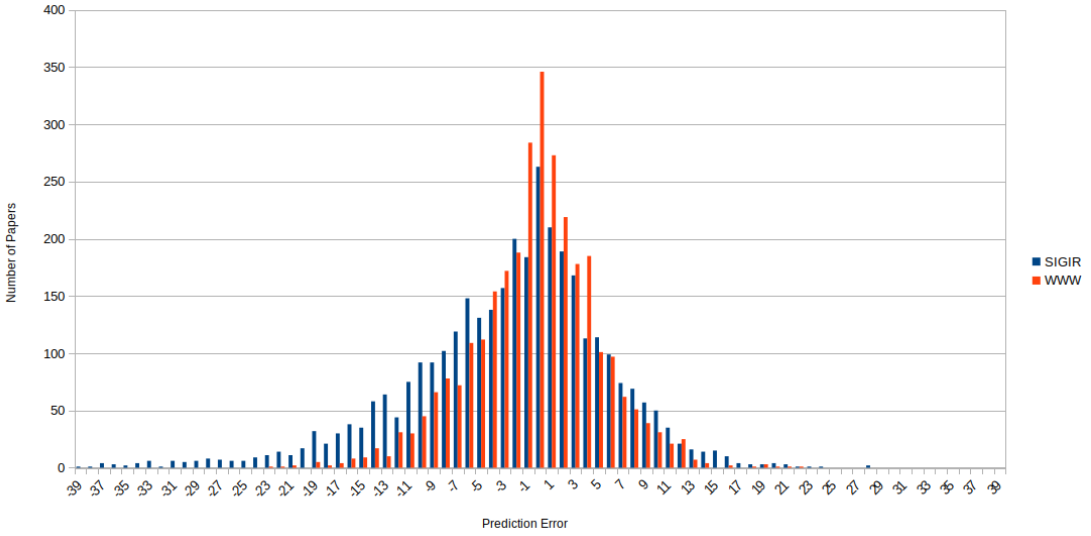

FIGURE 3.3: Prediction error distributions

Figure 3.3 shows the distributions of publication year prediction errors for SIGIR and WWW. Based on these distributions, let us define the prediction error for papers published in year Y as a discrete random variable $Err_Y : \{Y_b, \ldots, Y_e\} \rightarrow \{Y_b - Y, \ldots, Y_e - Y\}$ with a probability density function defined as:

$$Pr(Err_Y = n) = \begin{cases} \frac{|\{p \in P | \bar{Y_p} - Y_p = n\}|}{\sum_{n=Y_b-Y}^{Y_e-Y} |\{p \in P | \bar{Y_p} - Y_p = n\}|} & \text{if } Y_b - Y \leqslant n \leqslant Y_e - Y \\ 0 & \text{if } n < Y_b - Y \text{or } n > Y_e - Y \end{cases}$$

where Y_b and Y_e are the first and last years of the conference, respectively, Y_p and $\bar{Y_p}$ are the actual and predicted publication years of paper p, respectively, and P is the set of all papers from the conference.

Its expected value is then:

$$E(Err_Y) = \sum_{n=Y_b-Y}^{Y_e-Y} n \cdot Pr(Err_Y = n)$$

Note that for papers published in year Y the minimum prediction error is $Y_b - Y$, and the maximum prediction error is $Y_e - Y$. We, therefore, truncate the distribution in Figure 3.3 to $\{Y_b - Y, \ldots, Y_e - Y\}$ for each year Y.

For example, let us consider papers from the WWW Conference published in the year 2000. The minimum prediction error is $Y_b - Y = 1994 - 2000 = -6$ and the maximum prediction error is $Y_e - Y = 2017 - 2000 = 17$. To calculate the probability density

function of Err_{2000}, we truncate the distribution in Figure 3.3 to $\{-6, \ldots, 17\}$ and define $Pr(Err_{2000} = n)$ for each n between -6 and 17 as the number of papers in the WWW corpus for which the prediction error is equal to n divided by the number of papers for which the prediction error is between -6 and 17. $Pr(Err_{2000} = -6)$ is then equal to $\frac{98}{2670}$, $Pr(Err_{2000} = -6) = \frac{110}{2670}$, etc. The expected value of Err_{2000} is:

$$E(Err_{2000}) = \sum_{n=-6}^{17} n \cdot Pr(Err_{2000} = n) = -6 \cdot \frac{98}{2670} - 5 \cdot \frac{110}{2670} + \ldots + 17 \cdot \frac{2}{2670} \approx 1.073$$

Let us now define the innovation score of paper p adjusted for its publication year as:

$$S'_P(p) = S_P(p) - E(Err_{Y_p}) \tag{3.2}$$

$S'_P(p)$ allows us to compare innovation scores of papers published in different years. Its value is zero for papers whose predicted publication year is equal to the expected publication year, positive if it is later than expected and negative if it is earlier than expected.

3.4.5 Breakthrough Years

We consider a year important, if many highly innovative papers were published in that year. Based on the definition of the adjusted paper innovation score – S'_P (Equation 3.2) we define the innovation score of year y as the mean innovation score of all papers published in y:

$$S_Y(y) = \mu_{\{S'_P(p)|p \in P_y\}} = \frac{\sum_{p \in P_y} S'_P(p)}{|P_y|} \tag{3.3}$$

where Y_p is the publication year of paper p and P_y is the set of all papers published in year y, i.e. $\{p|Y_p = y\}$.

As illustrated in Figure 3.7, despite adjusting the paper scores for their publication years, the mean paper scores per year are in a decreasing trend. This can be explained by the fact that as the year increases, the maximum and minimum prediction errors decrease. We therefore consider important years as the local maxima of S_Y.

3.5 Results

3.5.1 Topics

3.8 shows the latent topics found for both WWW and SIGIR corpora. The "*Most relevant terms*" column lists words having the highest relevance for each topic calculated using the following formula as implemented by the topic model visualization tool pyLDAvis[6]:

$$R(w|t) = \lambda \cdot P(w|t) + (1 - \lambda)\frac{P(w|t)}{P(w)}$$

where $P(w|t)$ is the probability of word w given topic t and $P(w)$ is the observed probability of w's occurrence in the entire corpus. The second term of this equation is the so-called *lift* – a measure of how common words are in a given topic relative to their overall frequency. Its value is high for words that are highly probable in topic t, but rare in general. $\lambda \in [0, 1]$ is a parameter determining the weight of the words' topic-specific probability vs. their lift in calculating their relevance to each topic. This is a variant of the word relevance measure proposed by Sievert and Shirley [170]. The idea behind this approach is that simply listing the most common terms in each topic may not give a good enough idea to understand what those topics are really about, since the most frequent words overall will also most likely be among the most frequent words in each topic. Some of the most common words in many of the topics found for our corpora are *query*, *document*, *user* and *page*. For example, the word clouds for topics 14, 19, 34 and 40 look similar (Figure 3.4) but as it can be seen in 3.8, those topics are quite different.

Sievert and Shirley conducted a user study to find the best value of λ. For most topics they found the optimal value to be around 0.6. We used this as the default value and adjusted it upwards as necessary to avoid emphasizing rare "noise" words introduced by OCR errors etc.

The popularity of each topic over time is illustrated in Figure 3.5. The color of the cell in row t and column y represents the percentage of papers published in year y in which topic t occurs. Dark red means 100% and white means 0%.

3.5.2 Predicting Publication Years

The mean classification error achieved by our classifier on the WWW corpus was 3.9414 and on the SIGIR corpus: 7.8474. The confusion matrices are shown in Figure 3.6.

[6] `https://github.com/bmabey/pyLDAvis`

FIGURE 3.4: Word clouds showing the most frequent words in four different topics. Clockwise from top-left: #14, #19, #34 and #40. After ranking topic terms by relevance with $\lambda = 0.6$, we identified the topics as: "Query Expansion/Reformulation/Intent Prediction", "Speech Retrieval/Voice Queries", "Eye Tracking" and "Medical Search Engines".

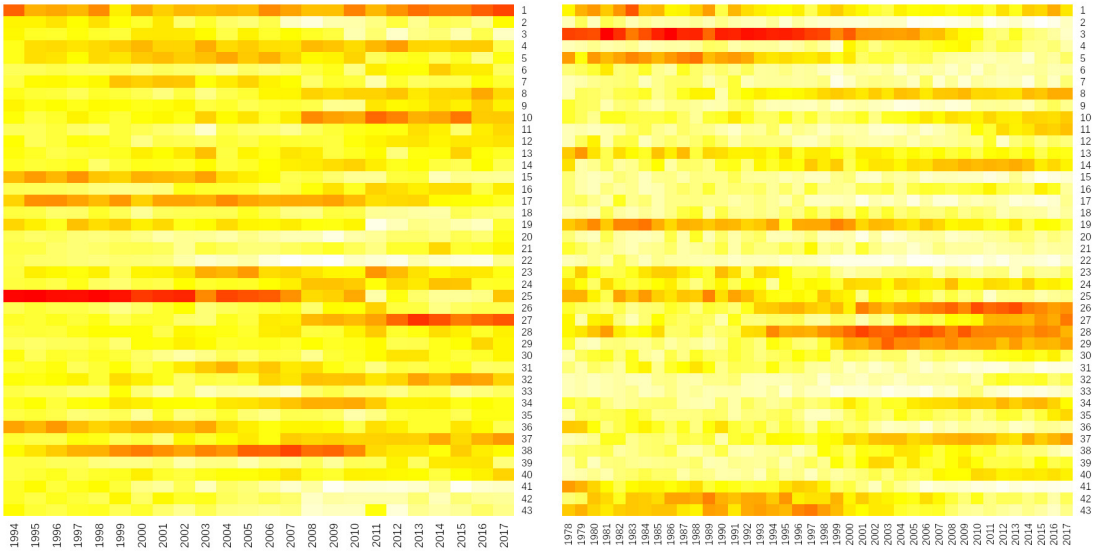

FIGURE 3.5: Topic popularity over time – WWW on the left, SIGIR on the right. Examples of topics gaining popularity in recent years are: #8 (Translation/Sentiment Analysis/Opinion Mining) and #27 (Recommender Systems). An example of a topic losing popularity is #25 (Web Applications).

After calculating paper innovation scores for all papers as described in Section 3.4.4 and all years as per Section 3.4.5, we have identified the following years as the most innovative:

SIGIR: 1983, 1989, 1994, 2001 and 2017.

WWW: 1997, 2000, 2002, 2011.

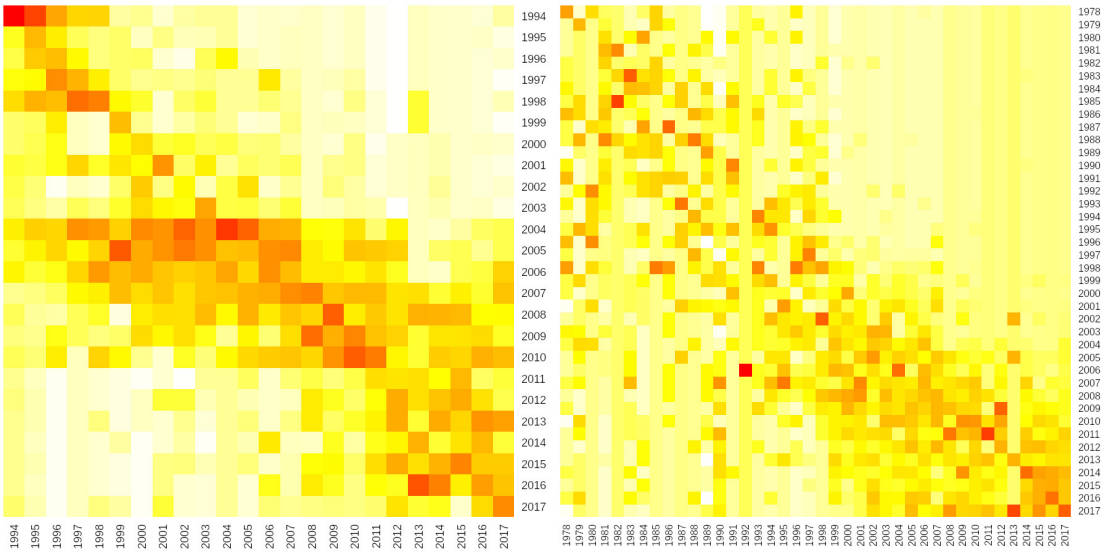

FIGURE 3.6: Confusion matrices as heatmaps – WWW on the left, SIGIR on the right. Actual publication years are in rows and predicted publication years are in columns. Brighter shades of red represent larger numbers, paler shades of yellow represent smaller numbers.

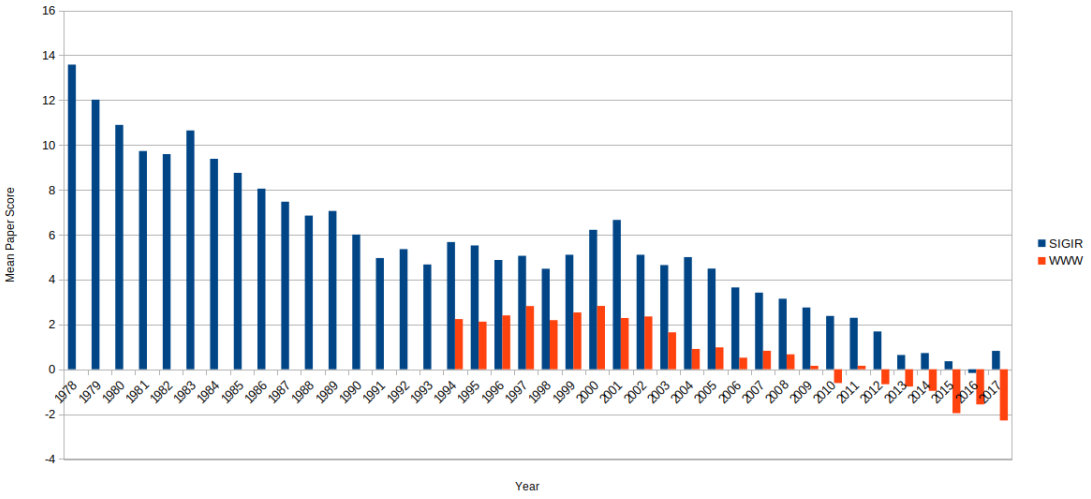

FIGURE 3.7: Year Importance Scores

3.5.3 Breakthrough Papers

3.9 and 3.10 list the top 10 papers for WWW and SIGIR respectively with the highest innovation scores computed as per Section 3.4.4. The *Topics* column lists topics found in each paper sorted by probability in decreasing order. The number of citations according to Google Scholar as of July 2018 is listed in the *Citations* column.

Papers from the SIGIR conference authored or co-authored by 3 out of 12 honorees of the Gerard Salton Award (see Section 3.1) have been found by our method to be among the top 3 papers with the highest innovation scores in their years:

- Nicholas J. Belkin: 1988, 2003
- Susan Dumais: 1992
- Gerard Salton: 1995

The authors who have published the most top-3 papers are:

WWW

- Xing Xie (2005, 2015, 2016, 2017)
- Rakesh Agrawal (2001, 2002, 2003)
- Ramakrishnan Srikant (2001, 2002, 2003)
- Fuzheng Zhang (2015, 2016, 2017)
- Qing Li (2005, 2007, 2008)

SIGIR

- Jun Wang (2006, 2011, 2013, 2014, 2015, 2017)
- Yi Zhang (2010, 2011, 2013, 2014)
- David A. Hull (1993, 1995, 1996)
- Hao Ma (2007, 2009, 2013)
- Neal Lathia (2009 – 2 papers, 2010)
- Chris Buckley (1995, 1997)
- Clement T. Yu (1985, 1991)
- James Allan (1995, 1996)
- Jan Pedersen (1995, 1996)
- Vijay V. Raghavan (1979, 1982)

3.5.4 Comparison with Citation Analysis

We have compared the results of our method with citation analysis by calculating Spearman's rank correlation coefficients between paper innovation scores S_P (Equation 3.1) and citation counts for papers published in each year separately (Figure 3.8) as well as on the entire span of years. The citation counts were obtained from ACM's Digital Library[7]. Data for the WWW conference prior to 2001 are unavailable.

In the calculations we have used $log(citation\ count + 1)$ rather than raw citation counts, based on Price's Cumulative Advantage principle. As shown by Price [171], the number of citations is expected to grow exponentially. The value of this expression is always non-negative and it is zero when citation count is zero. This also somewhat solves the problem of recent papers having disproportionately few citations compared to older ones. We have found the innovation scores and citation counts to be moderately correlated for earlier years and weakly correlated for later years. A possible explanation of this phenomenon would be the fact that more recent papers, even the "innovative" ones, have not yet accumulated enough citations. The results for SIGIR before 1995 are inconclusive for two reasons: Few papers and low quality data due to many OCR errors.

The overall correlation coefficients are: 0.2944 (p-value: $6.46 \cdot 10^{-52}$) for WWW and 0.2416 for SIGIR with a p-value of $1.1 \cdot 10^{-46}$. These correlation values do not indicate a strong linear relationship between innovation scores and citation counts, but as illustrated by the scatter plots in Figure 3.9, papers with many citations tend to also have high S_P scores. The inverse statement does not hold, i.e. a high S_P score does not necessarily equate to a high number of citations. This is partly caused by the fact that recent papers in general have fewer citations than older ones, but it also indicates that a paper's citation count does not always reflect its innovativeness. Some truly innovative papers may have been relatively unnoticed. Our method may be used to support identifying such potential "hidden gems".

3.5.5 Some Examples

One of the most noteworthy examples is a paper published at the WWW Conference in 2001: *Item-Based Collaborative Filtering Recommendation Algorithms* by B. Sarwar, G. Karypis, J. Konstan and J. Riedl. With a S'_P score of 10.82 it is the fifth highest-ranked WWW paper. Among those papers for which citation data are available in ACM DL, it is the highest ranked and the most cited with 1,507 citations.

[7] `http://dl.acm.org/`

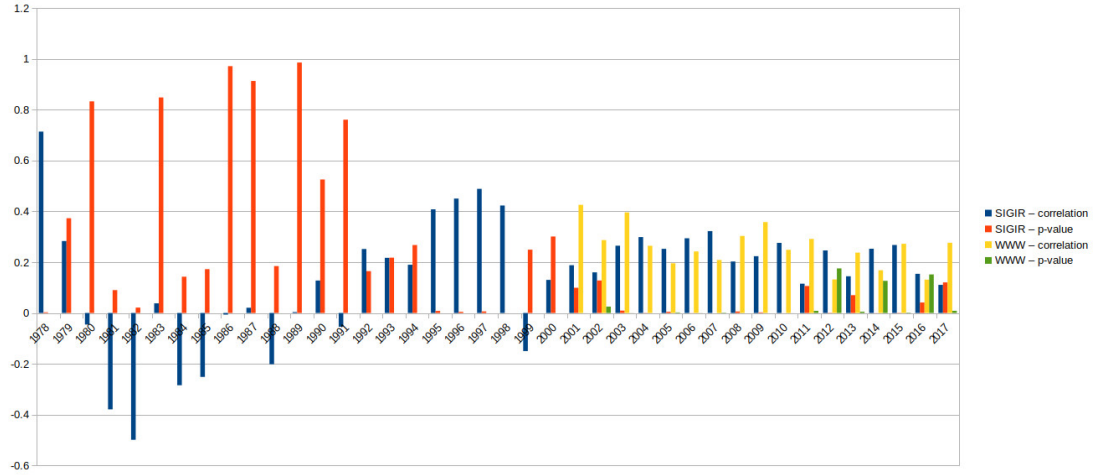

FIGURE 3.8: Spearman's rank correlation coefficients between paper innovation scores and citation counts for each year. *Correlation* is the correlation coefficient, or *Spearman's ρ*. *P-value* is the probability that a random dataset has a greater or equal correlation coefficient.

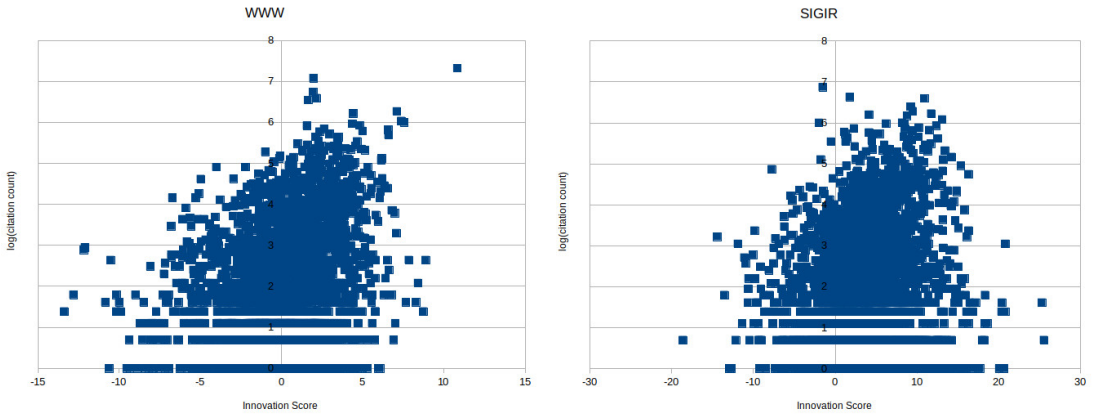

FIGURE 3.9: Scatter plots of Innovation Scores (S_P) vs. citation counts. High citation counts usually imply high S_P values but not vice versa.

Another example is the third highest ranked SIGIR paper in 1994: *A Sequential Algorithm for Training Text Classifiers* by D. D. Lewis and W. A. Gale. It is also highly cited with 437 citations.

Examples of highly-scored papers with few citations include:

- M. Dumas, L. Aldred, G. Governatori, A. ter Hofstede, N. Russell, *A Probabilistic Approach to Automated Bidding in Alternative Auctions* – WWW, 2002 This paper researches topic #6 – Online Auctions – whose popularity peaks in 2014. It is also one of the first papers to combine topics #6 and #16: Online Auctions and Algorithms

- G. Hulten, J. Goodman, R. Rounthwaite, *Filtering Spam E-mail on a Global Scale* – WWW, 2004 One of the first papers to combine topics #1, #9 and #32: E-mail, E-commerce, Geo/Location

- K. Yamamoto, D. Katagami, K. Nitta, A. Aiba, H. Kuwata, *The Credibility of the Posted Information in a Recommendation System Based on a Map* – WWW, 2006 This paper researches topic #27 – Recommender Systems, which has been highly popular in the last decade. It is also the first to combine topics #27, #30 and #39: Recommender Systems, Credibility and Signal Processing

- L. C. Smith, *'Memex' as an image of potentiality in information retrieval research and development* – SIGIR, 1980 This paper is one of the first occurrences of topic #21 – Citation Networks – researched for many years until 2017. It also one of the first co-occurrences of topics #3 and #21 – Document Retrieval and Citation Networks

3.6 Discussion

As shown above, the main strength of our method is its ability to automate the analysis of large corpora, while avoiding some of the problems of traditional scientometric approaches. It also provides interesting insights into fields of research such as topic popularity or co-occurrence. It is important to note, however, that our approach cannot measure publication importance or impact in their traditional understanding as it focuses on paper innovativeness. The former can be measured by citation analysis, which has its own shortcomings as outlined in Section 3.2.

The main limitations of our method are the need for a relatively large corpus spanning many years or even decades, and sensitivity to input data quality. This is especially challenging, since a significant portion of older documents digitized using OCR contain many misrecognized characters. This may lead to poor quality topic models if not corrected either automatically or manually.

The results for papers published in recent years may not be particularly conclusive and informative due to the lack of future data. Only time will tell whether or not the topics covered in those papers will become popular. However, the same may be said about citation analysis. Important publications may only be identified retrospectively. It may also be argued that our method could potentially be used for early identification of scientific breakthroughs, as the innovation score proposed in this paper penalizes papers covering topics researched in the past and rewards new topics.

An argument may be made that our innovation score should only take into account prediction errors "into the future" and ignore errors "into the past". Just because a paper covers topics popular in the past does not mean it is not innovative, as long as it also covers topics researched in the future. Also, papers rediscovering long forgotten topics may be viewed as more innovative than papers repeating topics popular in recent years. We will explore this further in our future work.

3.7 Conclusion and Future Work

We have demonstrated a simple, yet novel classification-based method of measuring innovation, which may be used to complement citation analysis in identifying potential breakthrough publications in bodies of research spanning multiple years. We proposed a real-number measure of innovativeness based on the prediction error of the publication year and the classifier's confidence. We also showed how to adjust this measure to allow for comparing scores of papers published in different years. Finally, we applied our method to all publications from the World Wide Web and SIGIR conferences and compared our results to citation analysis.

The most important contribution and the main advantage of our method is its ability to fully automate the analysis of the corpus. None of its steps – topic model training and selection, classifier training and score calculation – require expert knowledge or manual intervention.

In the future we plan on experimenting with using selected keywords as features alongside LDA topics. The choice of keywords is essential to avoid e.g. problems with named entities occurring rarely.

We will also consider Correlated Topic Models [49], which replace the Dirichlet distribution in the generative process with the logistic normal. This removes the topic independence assumption in LDA and models topic correlation with the covariance matrix of the logistic normal distribution. Blei and Lafferty [49] evaluate CTM and compare them to LDA on a corpus of over 16,000 articles published in *Science* between the years 1990 and 1999. They show that CTM gives a better fit to the data and has greater interpretability. Using the common held-out log-likelihood maximization approach they have found the optimal number of topics in each model for the *Science* corpus: 90 for CTM vs 30 for LDA. As they state, therefore, CTM supports more topics than LDA. This means that our classification problem could have higher dimensionality, which makes overfitting more likely due to the relatively small sizes of our corpora.

3.8 LDA topics found in the WWW and SIGIR corpora

#	Description	Most relevant terms
1	Social Studies, Surveys	user email people information participant study student privacy group use social activity system research survey community personal online data design message communication post find reply question experience time work make
2	Games	game player agent badge pda thin fat client contest wireless web pthinc student screen cheater advisee reward incentive ower attrition redland pdas advisor rdp pc leaderboard device hire wifi
3	Document Retrieval	term document retrieval weight query boolean collection index probability frequency model vector probabilistic thesaurus set value ir use concept function retrieve give information relevant fuzzy system relevance number indexing formula
4	Page Rank, Web Crawling	page link pagerank web crawl authority hub hit crawler surfer spam walk hyperlink graph random importance download hubs freshness www distribution kleinberg trustrank host urls algorithm seed changes core degree
5	Knowledge Representation	rdf object data relational database triple xml knowledge relation structure semantic system query language predicate attribute property sql frame element information model type example statement operator sparql representation join tuple
6	Online Auctions	auction bid bidder equilibrium revenue utility optimal mechanism price lemma agent payoff theorem valuation allocation proof nash gsp vertical reserve value bidding welfare surplus player vcg winner budget strategy bundle
7	XML	attribute schema xml element apps type app field label xquery name entity schemas dbpedia instance metadata xsl value xsd xpath expression wikidata xslt namespace inex dtds json class maturity fsdm
8	Translation, Sentiment Analysis, Opinion Mining	entity translation word review sentiment english wikipedia opinion name feature language corpus chinese product mention use extraction sentence disambiguation phrase clir noun lexicon extract dictionary candidate base translate bilingual monolingual
9	Crowdsourcing, E-commerce	worker price customer market seller crowdsourcing transaction payment buyer reputation purchase trust sale sell negotiation pricing commerce provider buy pay requester account job merchant marketplace scrip bonus abandonment consumer demand
10	Graphs	image graph edge vertex node hash algorithm label visual neighbor set subgraph distance walk similarity path random subgraphs lsh matrix network space problem function sample data large degree vi method
11	User Behavior Modeling	user click model action aspect session time metric comment sequence behavior position risk state qt transition interaction reward dwell dbn topic satisfaction attraction card perplexity probability propose video post urisk
12		event blog temporal theme influence sprea ddiffusion cascade spatiotemporal network time blogger poi sensor information social weblogs district video interestingness earthquake backlight life tribeflow evolution owlr trajectory causality flow investor
13	Document Storing, Indexing, Compression	index compression list block bit inverted invert compress posting query prune size time memory store processing cost document sort efficiency encode use space term doc id disk integer docids intersection post
14	Query Expansion, Reformulation, Intent Prediction, Suggestion	query suggestion log search predictor use intent result clarity term method keyword similarity reformulation shard patent expansion reformulations retrieval score original qc prediction phrase feature q1 top frequency list substitution
15	Networking	client resource proxy server packet tcp connection network http protocol request p2p header traffic transfer bandwidth multicast ip content peer user web mobile port send payload use trace dns object
16	Algorithms	class algorithm problem set label instance solution distance function time rank aggregation method objective data optimization give value constraint greedy use number xi define approach chemical category diversification probability temporal
17	Security	service web application security server request client policy qos code browser attack certificate use http execution javascript script password protocol composition soap vulnerability provider program secure process workflow invocation message
18		segment segmentation concept domain text oer dom label block target keyphrase vip form webpage key hearsay keyphrases oers source use information gna fullmatch transfer apprentice knowledge domains method web scl
19	Speech Retrieval, Voice Queries	search subject user system interface query term task searcher information voice speech retrieval use display dialogue study cognitive interactive transcript participant find feedback image asr relevant ask need result questionnaire
20		hashtags hashtag tweet trend day coupon burst elg seasonal spike periodic period stream activity time query hour bf peak energy msu periodicity guids push group competition klsh volume halo keyword
21	Citation Networks	community citation spam network author cite modularity paper grader member citeseer student grade group hole detection guild set conductance authorship membership mooc graph method ham kog dblp xql use citesight
22	Autocompletion	qac completion prefix suggestion keystroke mpc endorsement auto uiml cyworld detachment drinker child aesthetic sparqs volvo autocompletion key mostpopularcompletion vci appliance messidor group character xsquirrel endorse aspects tdcm mrr sogou beer
23	Pattern Matching, Automata	node tree message peer path match rule expression pattern root state child xpath leaf conversation automaton operation structure set choreography subtree join branch parent variable constraint ancestor order figure operator
24	Clustering, Document Tagging	cluster tag clustering similarity tags method photo algorithm use matrix label document flickr result data user base centroid set measure tdt folksonomy number different feature agglomerative dataset hierarchy icio mean
25	Web Applications, Hypertext	html file interface object support application link format provide presentation web tool document browser hypertext server use www system information access program allow form java element user hypermedia display create

26	Relevance Feedback	rank document query relevance ndcg relevant use retrieval model score ranking expansion feedback term learn method set performance result weight base baseline approach top train ranker function improvement fusion metric
27	Recommender Systems	user item recommendation rating social model prediction recommender preference friend collaborative network matrix latent recommend movie profile factor cf factorization base predict data influence filtering product interest method dataset use
28	Text Retrieval	collection document trec topic judgment relevant system test assessor relevance measure run evaluation judge precision pool assessment sample score average correlation use difference retrieval ap rank effectiveness track set estimate
29	Topic Models	topic word model document term lda tf proximity dirichlet vector idf latent lsi parameter text mixture propose language representation use bm25 distribution probability embeddings method wi corpus semantic collection weight
30	Twitter, News, Credibility	tweet news article twitter story facet user faceted follower retweet topic system retweets credibility timeline post shortlist retweeted headline information day book celebrity time silk list ir road sjasm digg
31	Ontology, Semantics	ontology owl semantic semantics concept dl axiom daml rdf rule logic rdfs description property assertion oil reasoner subsumption kb sparql datalog role team ontological knowledge class expressive definition lite interpretation
32	Geo	location website visit site attack malicious cooky user geographic attacker gps browser activity geographical pin restaurant geo url cookie country city phishing content mobile google place tv honeypot popular domain
33	Ads	ad advertiser advertising ctr impression campaign click advertisement bid conversion sponsor advertise landing revenue publisher target attack rider auctioneer page wc iolaus keyword pay brand cpc fulld banner uber driver
34	Eye Tracking	search result engine click query user vertical snippet serp behavior eye study serps rank page position gaze cursor task examine show navigational trail fixation time session information return examination interaction
35	Question Answering	answer question qac qa expertise asker passage candidate answerer nugget lstm vote expert thread post factoid reply quality score yahoo forum definitional best correct use mrr system base faq sentence
36	Network Performance	cache server request load time workload latency policy access hit proxy client rate replication response invalidation prefetching lru data update performance size system throughput disk delay distribute replica bandwidth memory
37	Classification	classifier feature classification category svm training train learn music video class categorization accuracy classify emotion label f1 set use data positive ensemble text kernel example performance unlabeled audio test vector
38		page website anchor content url duplicate text data wrapper urls shingle crawl link extraction crawler extract algorithm search information record html sit pattern fingerprint find result use tree show
39	Signal Processing	model distribution smooth music parameter song prior passage language data likelihood retrieval mixture estimate fit probability use signal performance posterior jelinek bayesian gaussian patent mercer lm rumor variational generative method
40	Medical Search Engines	task search session query user annotation feature trail web medical use log symptom page dwell data context switch intent recipe searcher personalization predict usefulness study information click behavior result history
41		document profile hierarchy term visualization pgp search panel cci ingrid search shrec elsi transaction xxx searchpad iqe spl user row number sal arc fish vtml feedback trf cluster dissemination referent information
42	Text Summarization	sentence summary summarization phrase signature document rouge text word paragraph use noun syntactic duc article structure information method rhetorical relation summarizer extract contain system set unit fragment figure lpi lexical
43	Morphology, Stemming	word character dictionary stem stemmer index text gram morphological string sense letter ocr retrieval use query number sens term collection match frequency inquery arabic compound occurrence suffix code morphology selfie

3.9 Top 10 Papers – WWW

Year	Paper	Topics	Citations
1994	N. Arnett, *The Internet and the Anti-net. Two public internetworks are better than one*	24, 11, 23, 9	
1995	S. Glassman, M. Manasse, M. Abadi, P. Gauthier, P. Sobalvarro, *The Millicent Protocol for Inexpensive Electronic Commerce*	1, 17, 38, 36, 25, 27	
1998	T. Fenech, *Using perceived ease of use and perceived usefulness to predict acceptance of the World Wide Web*	1, 12, 39, 2, 19	
2001	B. Sarwar, G. Karypis, J. Konstan, J. Riedl, *Item-Based Collaborative Filtering Recommendation Algorithms*	27, 13, 24, 36, 28, 9	1507
1994	E. Fischer, *Graffiti on the Web: A Cultural Interchange. A Lighthearted Romp with an Artist in Webland that Stops Being Lighthearted at the End*	2, 1, 25	
1998	M. Allen, *Are we yet cyborgs? University students and the practical consequences of human–machine subjectivity*	1, 21, 25	
1996	A. Richmond, *Enticing Online Shoppers to Buy – A Human Behavior Study*	6, 14, 23	116
1994	P. Tsang, J. Henri, S. Tse, *Internet Growth in Australia and Asia's Four Dragons*	12, 1, 25, 9, 15	
1997	H. Sakagami, T. Kamba, *Learning Personal Preferences on Online Newspaper Articles from User Behaviors*	30, 27, 19, 4, 34, 25	
1995	J. E. Pitkow, C. M. Kehoe, *Results from the Third WWW User Survey*	7, 6, 11	

3.10 Top 10 Papers – SIGIR

Year	Paper	Topics	Citations
1983	A. E. Wessel *PROGRESS REPORT ON PROJECT INFORMATION BRIDGE*	10, 3, 43, 19, 41	1
1978	R. E. Williamson, *Does Relevance Feedback Improve Document Retrieval Performance?*	28, 26, 41, 3, 19, 10	4
1986	C. BERRUT, P. PALMER, *Solving Grammatical Ambiguities within a Surface Syntactical Parser for Automatic Indexing*	40, 42, 43, 8, 23, 21	3
1982	A.S. Fraenkel, M. Mor, Y. Perl, *IS TEXT COMPRESSION BY PREFIXES AND SUFFIXES PRACTICAL?*	43, 13, 29, 10, 22, 6	0
1990	E. A. Fox, Qi Fan Chen, A. M. Daoud, L. S. Heath, *Order Preserving Minimal Perfect Hash Functions and Information Retrieval*	10, 36, 3, 13, 23, 5	3
1982	V. V. Raghavan, M.Y.L. Ip, *TECHNIQUES FOR MEASURING THE STABILITY OF CLUSTERING: A COMPARATIVE STUDY*	24, 5, 10, 37, 28, 3	4
1979	V. V Raghavan, K. Birchard, *A Clustering Strategy Based on a Formalism of the Reproductive Process in Natural Systems*	24, 5, 6, 3, 23	20
1981	Y. Kambayashi, T. Hayashi, Sh. Yajima, *DYNAMIC CLUSTERING PROCEDURES FOR BIBLIOGRAPHIC DATA*	24, 21, 3, 14, 10, 5	0
1985	M. D. Gordon, *A LEARNING ALGORITHM APPLIED TO DOCUMENT REDESCRIPTION*	26, 3, 19, 5, 1	2
1985	D. M. Arnow, A. M. Tenenbaum, C. Wu, *P-Trees: Storage Efficient Multiway Trees*	23, 11, 36, 13, 43, 10	1

3.11 Top 3 Papers for Selected Years – WWW

Year	Top Papers	Topics	Citations
1995	S. Glassman, M. Manasse, M. Abadi, P. Gauthier, P. Sobalvarro, *The Millicent Protocol for Inexpensive Electronic Commerce*	1, 17, 38, 36, 25, 27	
	J. E. Pitkow, C. M. Kehoe, *Results from the Third WWW User Survey*	7, 6, 11	
	M. Peirce, D. O'Mahony, *Scaleable, Secure Cash Payment for WWW Resources with the PayMe Protocol Set*	9, 17, 19	
2000	N. Abe, T. Kamba, *A web marketing system with automatic pricing*	9, 6, 27, 8, 38, 10	
	C. Hölscher, G. Strube, *Web Search Behavior of Internet Experts and Newbies*	1, 40, 19, 34, 38, 15	
	J. A. Tomlin, *An Entropy Approach to Unintrusive Targeted Advertising on the Web*	16, 11, 33, 3, 6, 1	
2005	Q. Li, B. M. Kim, S. H. Myaeng, *Clustering for Probabilistic Model Estimation for CF*	27, 21	4
	X.-F. Su, H.-J. Zeng, Zh. Chen, *Finding Group Shilling in Recommendation System*	27, 17, 24, 32, 9, 12	13
	C.-N. Ziegler, S. M. McNee, J. A. Konstan, G. Lausen, *Improving Recommendation Lists Through Topic Diversification*	14, 22, 10, 21, 15, 19	402
2010	S. Rendle, C. Freudenthaler, L. Schmidt-Thieme, *Factorizing Personalized Markov Chains for Next-Basket Recommendation*	27, 11	140
	Zh. Wen, Ch.-Y. Lin, *How Accurately Can One's Interests Be Inferred From Friends?*	27, 1, 10, 28, 37, 34	3
	L. Backstrom, E. Sun, C. Marlow, *Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity*	27, 1, 32, 10	161
2015	F. Zhang, K. Zheng, N. Jing Yuan, X. Xie, E. Chen, X. Zhou, *A Novelty-Seeking based Dining Recommender System*	27, 32, 16, 1	6
	J. Wang, D. Hardtke, *User Latent Preference Model for Better Downside Management in Recommender Systems*	27, 18, 1, 32, 39, 38	2
	I. Kloumann, L. Adamic, J. Kleinberg, Sh. Wu, *The Lifecycles of Apps in a Social Ecosystem*	27, 1, 7, 32, 20, 37	6

3.12 Top 3 Papers for Selected Years – SIGIR

Year	Top Papers	Topics	Citations
1980	L. C. Smith, *'Memex' as an image of potentiality in information retrieval research and development*	21, 1, 25, 3, 19, 30	3
	R. C. Schank, J. L. Kolodner, G. DeJong, *Conceptual information retrieval*	4, 5, 1, 42, 30, 19	27

	C. D. Hafner, *Representation of knowledge in a legal information retrieval system*	5, 12, 27, 31, 42, 19	8
1985	M. D. Gordon, *A LEARNING ALGORITHM APPLIED TO DOCUMENT REDESCRIPTION*	26, 3, 19, 5, 1	2
	D. M. Arnow, A. M. Tenenbaum, C. Wu, *P-Trees: Storage Efficient Multiway Trees*	23, 11, 36, 13, 43, 10	1
	T. Ito, C. T. Yu, *OPTIMIZATION OF A HIERARCHICAL FILE ORGANIZATION FOR SPELLING CORRECTION*	22, 43, 13, 41, 6	1
1990	E. A. Fox, Qi Fan Chen, A. M. Daoud, L. S. Heath, *Order Preserving Minimal Perfect Hash Functions and Information Retrieval*	10, 36, 3, 13, 23, 5	3
	A. Bookstein, Sh. T. Klein, *Construction of Optimal Graphs for Bit-Vector Compression*	13, 10, 42, 34, 3, 23	3
	C. Stanfill, *Partitioned Posting Files: A Parallel Inverted File Structure for Information Retrieval*	13, 43, 20, 36, 3, 41	20
1995	H. Schütze, D. A. Hull, J. O. Pedersen, *A Comparison of Classifiers and Document Representations for the Routing Problem*	26, 37, 29, 20, 3, 42	141
	J. Allan, *Relevance Feedback With Too Much Data*	26, 28, 43, 42, 3, 41	30
	C. Buckley, G. Salton, *Optimization of Relevance Feedback Weights*	26, 3, 41	76
2000	Ch. Zhai, P. Jansen, D. A. Evans, *Exploration of a Heuristic Approach to Threshold Learning in Adaptive Filtering*	26, 28, 11, 39, 36, 6	1
	X. Zhu, S. Gauch, *Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web*	26, 38, 28, 34, 4, 9	53
	M. Iwayama, *Relevance Feedback with a Small Number of Relevance Judgements: Incremental Relevance Feedback vs. Document Clustering*	26, 28, 24, 41, 34	33
2005	Gui-Rong Xue, Chenxi Lin, Qiang Yang, WenSi Xi, Hua-Jun Zeng, Yong Yu, Zheng Chen, *Scalable Collaborative Filtering Using Cluster-based Smoothing*	27, 24, 39	160
	L. Wang, Ch. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, Y. Li, *Detecting Dominant Locations from Search Queries*	34, 32, 14, 8, 38, 37	42
	K. Yu, Sh. Yu, V. Tresp, *Multi-Label Informed Latent Semantic Indexing*	10, 37, 18, 29	73
2010	N. Kawamae, *Serendipitous Recommendations via Innovators*	27, 34, 14, 9	15
	N. Lathia, S. Hailes, L. Capra, X. Amatriain, *Temporal Diversity in Recommender Systems*	27, 34, 28, 1	66
	F. Zhong, D. Wang, G. Wang, W. Chen, Y. Zhang, Zh. Chen, H. Wang, *Incorporating Post-Click Behaviors into a Click Model*	11, 26, 13, 34	14
2015	X. Liu, W. Wu, *Learning Context-aware Latent Representations for Context-aware Collaborative Filtering*	27	1
	X. Li, G. Cong, X.-L. Li, T.-A. Nguyen Pham, Sh. Krishnaswamy, *Rank-GeoFM: A Ranking based Geographical Factorization Method for Point of Interest Recommendation*	27	30
	P. Wang, J. Guo, Y. Lan, J. Xu, Sh. Wan, X. Cheng, *Learning Hierarchical Representation Model for Next Basket Recommendation*	27, 9, 16	25

Pavel Savov, Adam Jatowt, and Radoslaw Nielek. Innovativeness Analysis of Scholarly Publications by Age Prediction using Ordinal Regression. *International Conference on Computational Science*, pages 646-660. Springer, Cham, 2020.

Chapter 4

Innovativeness Analysis of Scholarly Publications by Age Prediction using Ordinal Regression

In this paper we refine our method of measuring the innovativeness of scientific papers. Given a diachronic corpus of papers from a particular field of study, published over a period of a number of years, we extract latent topics and train an ordinal regression model to predict publication years based on topic distributions. Using the prediction error we calculate a real-number based innovation score, which may be used to complement citation analysis in identifying potential breakthrough publications. The innovation score we had proposed previously could not be compared for papers published in different years. The main contribution we make in this work is adjusting the innovation score to account for the publication year, making the scores of papers published in different years directly comparable. We have also improved the prediction accuracy by replacing multiclass classification with ordinal regression and Latent Dirichlet Allocation models with Correlated Topic Models. This also allows for better understanding of the evolution of research topics. We demonstrate our method on two corpora: 3,577 papers published at the International World Wide Web Conference (WWW) between the years 1994 and

2019, and 835 articles published in the Journal of Artificial Societies and Social Simulation (JASSS) from 1998 to 2019.

4.1 Introduction

Citation analysis has been the main method of measuring innovation and identifying important and/or pioneering scientific papers. It is assumed that papers having high citation counts have made a significant impact on their fields of study and are considered innovative. This approach, however, has a number of shortcomings: Works by well-known authors and/or ones published at well-established publication venues tend to receive more attention and citations than others (the rich-get-richer effect) [7]. According to Merton [8], who first described this phenomenon in 1968, publications by more eminent researchers will receive disproportionately more recognition than similar works by less-well known authors. This is known as the *Matthew Effect*, named after the biblical Gospel of Matthew. Serenko and Dumay [10] observed that old citation classics keep getting cited because they appear among the top results in Google Scholar, and are automatically assumed as credible. Some authors also assume that reviewers expect to see those classics referenced in the submitted paper regardless of their relevance to the work being submitted. There is also the problem of self-citations: Increased citation count does not reflect the work's impact on its field of study.

We addressed these shortcomings in our previous work [172] by proposing a machine learning-based method of measuring the innovativeness of scientific papers. Our current method involves training a Correlated Topic Model (CTM) [49] on a diachronic corpus of papers published at conference series or in different journal editions over as many years as possible, training a model for predicting publication years using topic distributions as feature vectors, and calculating a real number innovation score for each paper based on the prediction error.

We consider a paper innovative if it covers topics that will be popular in the future but have not been researched in the past. Therefore, the more recent the publication year predicted by our model compared to the actual year of publication, the greater the paper's score. We showed in [172] that our innovation scores are positively correlated with citation counts, but there are also highly scored papers having few citations. These papers may be worth looking into as potential "hidden gems" – covering topics researched in the future but relatively unnoticed. Interestingly, we have not found any highly cited papers with low innovation scores.

4.2 Related Work

The development of research areas and the evolution of topics in academic conferences and journals over time have been investigated by numerous researchers. For example, Meyer et al. [43] study the Journal of Artificial Societies and Social Simulation (JASSS) by means of citation and co-citation analysis. They identify the most influential works and authors and show the multidisciplinary nature of the field. Saft and Nissen [77] also analyze JASSS, but they use a text mining approach linking documents into thematic clusters in a manner inspired by co-citation analysis. Wallace et al. [36] study trends in the ACM Conference on Computer Supported Cooperative Work (CSCW). They took over 1,200 papers published between the years 1990 and 2015, and they analyzed data such as publication year, type of empirical research, type of empirical evaluations used, and the systems/technologies involved. [78] analyze trends in the writing style in papers from the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) published over a 36-year period.

Recent research on identifying potential breakthrough publications includes works such as Schneider and Costas [3, 4]. Their approach is based on analyzing citation networks, focusing on highly-cited papers. Ponomarev et al. [5] predict citation count based on citation velocity, whereas Wolcott et al. [6] use random forest models on a number of features, e.g. author count, reference count, H-index etc. as well as citation velocity. These approaches, in contrast to ours, take into account non-textual features. They also define breakthrough publications as either highly-cited influential papers resulting in a change in research direction, or "articles that result from transformative research" [6].

A different approach to identifying novelty was proposed by Chan et al. [104]. They developed a system for finding analogies between research papers, based on the premise that "scientific discoveries are often driven by finding analogies in distant domains". One of the examples given is the simulated annealing optimization algorithm inspired by the annealing process commonly used in metallurgy. Identifying interdisciplinary ideas as a driver for innovation was also studied by Thorleuchter and Van den Poel [109]. Several works have employed machine learning-based approaches to predict citation counts and the long-term scientific impact (LTSI) of research papers, e.g., [94] or [9].

Examples of topic-based approaches include Hall et al. [58]. They trained an LDA model on the ACL Anthology, and showed trends over time like topics increasing and declining in popularity. Unlike our approach, they hand-picked topics from the generated model and manually seeded 10 more topics to improve field coverage. More recently Chen et al. [65] studied the evolution of topics in the field of information retrieval (IR). They trained a 5-topic LDA model on a corpus of around 20,000 papers from *Web of Science*.

Sun and Yin [59] used a 50-topic LDA model trained on a corpus of over 17,000 abstracts of research papers on transportation published over a 25-year period to identify research trends by studying the variation of topic distributions over time. Another interesting example is the paper by Hu et al. [70] where Google's Word2Vec model is used to enhance topic keywords with more complete semantic information, and topic evolution is analyzed using spatial correlation measures in a semantic space modeled as an urban geographic space.

Research on document dating (timestamping) is related to our work, too. Typical approaches to document dating are based on changes in word usage and on language change over time, and they use features derived from temporal language models [112, 116], diachronic word frequencies [119, 120], or occurrences of named entities. Examples of research articles based on heuristic methods include: [118], [124] or [123]. Jatowt and Campos [117] have implemented the visual, interactive system based on n-gram frequency analysis. In our work we rely on predicting publication dates to determine paper innovativeness. Ordinal regression models trained on topic vectors could be regarded as a variation of temporal language models and reflect vocabulary change over time. Aside from providing means for timestamping, they also allow for studying how new ideas emerge, gain and lose popularity.

4.3 Datasets

The corpora we study in this paper contain 3,577 papers published at the International World Wide Web Conference (WWW) between the years 1994 and 2019, and 835 articles published in the Journal of Artificial Societies and Social Simulation (JASSS)[1] from 1998 to 2019. We have studied papers from the WWW Conference before [172], which is the reason why we decided to use this corpus again, after updating it with papers published after our first analysis, i.e. ones in the years 2018 and 2019. We chose JASSS as the other corpus to analyze in order to demonstrate our method on another major publication venue in a related but separate field, published over a period of several years. It is publicly available in HTML, which makes it straightforward to extract text from the documents.

In an effort to extract only relevant content, we performed the following preprocessing steps on all texts before converting them to Bag-of-Words vectors:

1. Discarding page headers and footers, *References*, *Bibliography* and *Acknowledgments* sections as "noise" irrelevant to the main paper topic(s)

[1] `http://jasss.soc.surrey.ac.uk/`

2. Conversion to lower case
3. Removal of stopwords and punctuation as well as numbers, including ones spelled out, e.g. "one", "two", "first" etc.
4. Part-of-Speech tagging using the Penn Treebank POS tagger (NLTK) [164] – This step is a prerequisite for the WordNet Lemmatizer, we do not use the POS tags in further processing
5. Lemmatization using the WordNet Lemmatizer in NLTK

4.4 Method

4.4.1 Topic Model

In our previous work [172] we trained Latent Dirichlet Allocation (LDA) [38] topic models. In this paper, however, we have decided to move towards Correlated Topic Models (CTM) [49] and only built LDA models as a baseline. Unlike LDA, which assumes topic independence, CTM allows for correlation between topics. We have found this to be better suited for modeling topics evolving over time, including splitting or branching. We used the reference C implementation found at `http://www.cs.columbia.edu/~blei/ctm-c/`.

In order to choose the number of topics k, we have built a k-topic model for each k in a range we consider broad enough to include the optimum number of topics. In the case of LDA this range was $\langle 10, 60 \rangle$. We then chose the models with the highest C_V topic coherence. As shown by Röder et al. [64], this measure approximates human topic interpretability the best. Furthermore, according to Chang et al. [61], topic model selection based on traditional likelihood or perplexity-based approaches results in models that are worse in terms of human understandability. The numbers of topics we chose for our LDA models were 44 for the WWW corpus and 50 for JASSS. Because CTM supports more topics for a given corpus [49] and allows for a more granular topic model, we explored different ranges of k than in the case of LDA: $\langle 30, 100 \rangle$ for WWW and $\langle 40, 120 \rangle$ for JASSS. As before, we chose the models with the highest C_V.

4.4.2 Publication Year Prediction

Because publication years are ordinal values rather than categorical ones, instead of One-vs-One or One-vs-Rest multiclass classifiers, which we had used previously, we have implemented ordinal regression (a.k.a. ordinal classification) based on the framework

proposed by Li and Lin [173], as used by Martin et al. [174] for photograph dating. An N-class ordinal classifier consists of $N-1$ *before-after* binary classifiers, i.e. for each pair of consecutive years a classifier is trained, which assigns documents to one of two classes: "year y or before" and "year $y+1$ or after". Given the class membership probabilities predicted by these classifiers, the overall classifier confidence that paper p was published in the year Y is then determined, as in [174], by Eq. 4.1:

$$conf(p,Y) = \prod_{y=Y_{min}}^{Y} P(Y_p \leqslant y) \cdot \prod_{y=Y+1}^{Y_{max}} (1 - P(Y_p \leqslant y)) \tag{4.1}$$

where Y_{min} and Y_{max} are the first and last year in the corpus, and Y_p is the publication year of the paper p.

We used topic probability distributions as k-dimensional feature vectors, where k is the number of topics. Due to the small size of the JASSS corpus, we trained a separate model to evaluate each document (Leave-one-out cross-validation), whereas in the case of the WWW corpus we have settled for 10-fold cross-validation. We have implemented ordinal regression using linear Support Vector Machine (SVM) classifiers.

4.4.3 Paper Innovation Score

Following [172], we define our innovation score based on the results from the previous step - classifier confidence - as the weighted mean publication year prediction error with classifier confidence scores as weights:

$$S_P(p) = \frac{\sum_y conf(p,y) \cdot (y - Y_p)}{\sum_y conf(p,y)} \tag{4.2}$$

where Y_p is the year paper p was published in and $conf(p,y)$ is the classifier confidence for paper p and year y. Unlike the score defined in [172], the denominator in Eq. 4.2 does not equal 1, since the scores $conf(p,y)$ defined in Eq. 4.1 are not class membership probabilities.

As illustrated in Fig. 4.1, the higher the publication year of paper p, the lower the minimum and maximum possible values of $S_P(p)$. In order to make papers from different years comparable in terms of innovation scores, $S_P(p)$ needs to be adjusted to account for the publication year of paper p.

Suppose the prediction error for papers published in the year Y is a discrete random variable Err_Y. Based on the actual prediction error distributions for the WWW and

JASSS corpora (see Fig. 4.3), let us define the expected publication year prediction error for papers published in the year Y as:

$$E(Err_Y) = \sum_{n=Y_{min}-Y}^{Y_{max}-Y} n \cdot Pr(Err_Y = n) \tag{4.3}$$

where Y_{min} and Y_{max} are the minimum and maximum publication years in the corpus, and $Pr(Err_Y = n)$ is the observed probability that the prediction error for a paper published in the year Y is n. To calculate $Pr(Err_Y = n)$ we use the distribution from Fig. 4.3 truncated to the range $\langle Y_{min} - Y, Y_{max} - Y \rangle$, i.e. the minimum and maximum possible prediction errors for papers published in the year Y.

Let us then define the adjusted innovation score as the deviation of $S_P(p)$ from its expected value divided by its maximum absolute value:

$$S'_P(p) = \begin{cases} \frac{S_P(p) - E(Err_{Y_p})}{E(Err_{Y_p}) - (Y_{min} - Y_p)} & \text{if } S_P(p) < E(Err_{Y_p}) \\ \frac{S_P(p) - E(Err_{Y_p})}{Y_{max} - Y_p - E(Err_{Y_p})} & \text{if } S_P(p) \geqslant E(Err_{Y_p}) \end{cases} \tag{4.4}$$

where Y_p is the publication year of the paper p.
$S'_P(p)$ has the following characteristics:

1. $-1 \leqslant S'_P(p) \leqslant 1$
2. $S'_P(p) = 0$ if paper p's predicted publication year is as expected
3. $S'_P(p) < 0$ if paper p's predicted publication year is earlier than expected
4. $S'_P(p) > 0$ if paper p's predicted publication year is later than expected

4.5 Results

Fig. 4.2 shows the relation between the number of topics k and coherence C_V for CTM models trained on each of our corpora. Topic coherence initially peaks for values of k close to the optimal values found for LDA, then after a dip, it reaches global maxima for k equal to 74 and 88 for WWW and JASSS, respectively.

As shown in Tab. 4.1, publication year prediction accuracy expressed as Mean Absolute Error (MAE) is markedly improved both by using CTM over LDA and ordinal regression over a standard One-vs-One (OvO) multiclass SVM classifier. The best result we achieve for the WWW corpus was 2.56 and for JASSS: 3.56.

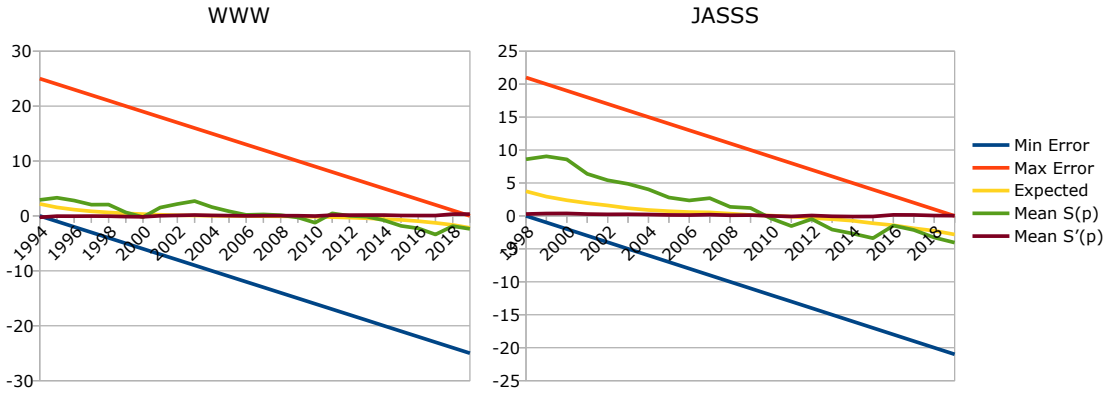

FIGURE 4.1: Minimum and maximum prediction errors decrease as the publication year increases and so does the mean unadjusted score (S_P). To make papers from different years comparable in terms of innovation score, the adjusted innovation score (S'_P) measures the deviation of the prediction error from its expected value.

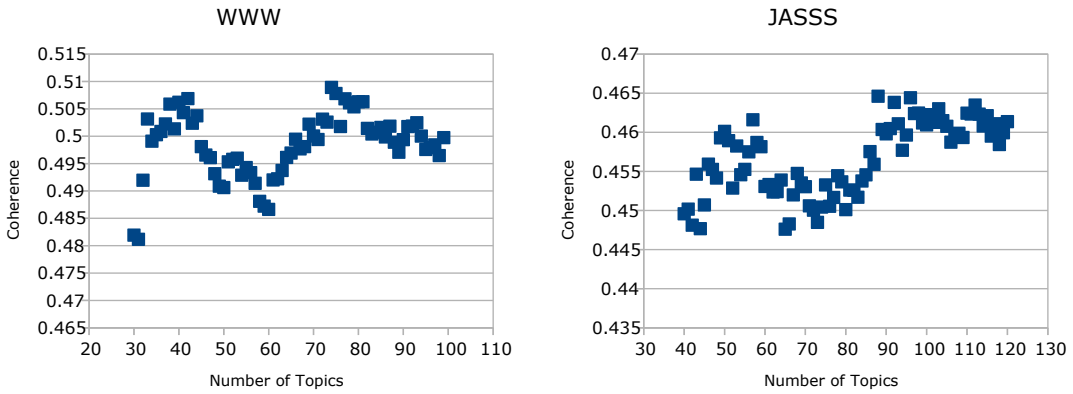

FIGURE 4.2: C_V Topic coherence by number of topics. We chose the CTM models with the highest values of C_V coherence as described in Sec. 4.4.1.

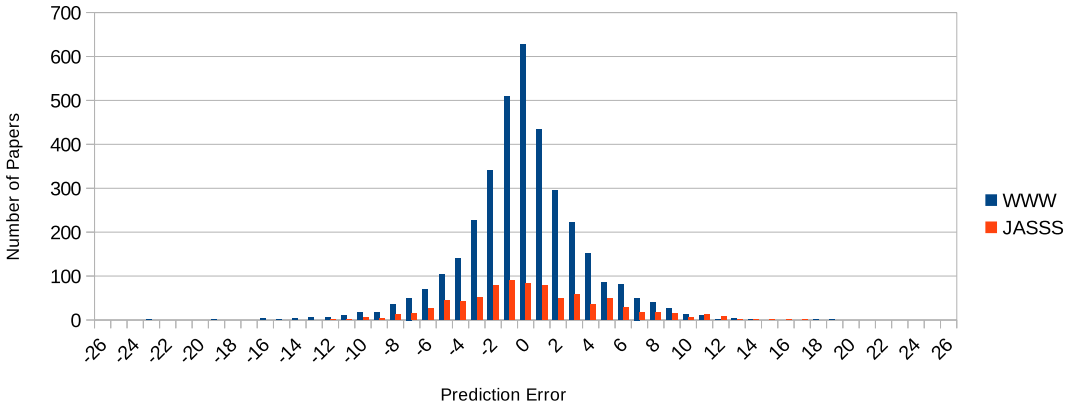

FIGURE 4.3: Distribution of publication year prediction errors for both corpora. We use these distributions to calculate the expected prediction error for each year and adjust paper innovation scores for their publication years.

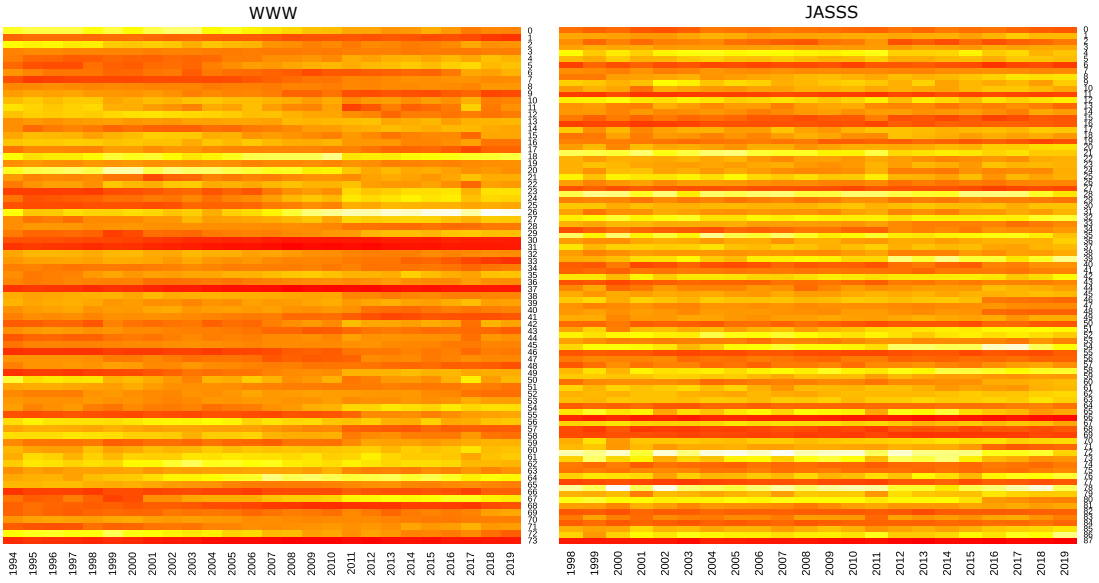

FIGURE 4.4: Topic popularity over time. The color of the cell in row t and column y represents the mean proportion of topic t in papers published in the year y. Bright red represents maximum values, white means zero.

TABLE 4.1: Mean absolute prediction errors: CTM vs. LDA and Multiclass SVM vs. Ordinal Regression

	Multiclass SVM		Ordinal Regression	
	WWW	JASSS	WWW	JASSS
LDA	4.14	6.09	3.34	4.38
CTM	3.02	4.22	2.56	3.56

Tab. 4.3 shows the top 3 papers with the highest innovation scores for both corpora. For each of those papers we list the number of citations and some of their most significant topics. All of them have been cited, some of them widely. The more a paper's topic distribution resembles the topic distributions of papers published in the future and the less it resembles that of papers from the past, the higher the innovation score. Some examples of highly scored, fairly recently published papers having few citations include:

- WWW, 2019: *Multiple Treatment Effect Estimation using Deep Generative Model with Task Embedding* by Shiv Kumar Saini et al. – no citations, 6^{th} highest score (0.946), topics covered: #10, #28, #33, #57 (see: Tab. 4.2)
- JASSS, 2017: *R&D Subsidization Effect and Network Centralization: Evidence from an Agent-Based Micro-Policy Simulation* by Pierpaolo Angelini et al. – 2 citations, 20^{th} highest score (0.634), topics covered: #4, #48, #65 (see: Tab. 4.2)

Fig. 4.5 illustrates the correlation between Innovation Scores and citation counts. Because the number of citations is expected to grow exponentially [171], we have used

TABLE 4.2: Selected latent topics described by their top 30 words.

	No.	Top 30 Words
WWW	2	cluster similarity algorithm set use measure intent result document number group base approach different information click give distance web method similar user problem find represent clustering term session figure follow
	4	object information web model multimedia use content provide base presentation retrieval type structure medium metadata represent show level image also system support relationship value order different part present define point
	9	network node link sample edge method random walk graph model degree social use distribution show figure matrix number result value set base prediction parameter time performance follow order neighbor problem
	10	ad advertiser click advertising use target bid user model ctr impression show search revenue advertisement online value campaign per number domain display keywords learn keyword rate conversion bundle sponsor base
	12	user tweet twitter post account social spam use number follower content campaign network follow also show feature detection find detect study medium group identity figure abusive information identify spammer time
	15	social network tag co information author people user use paper friend relationship group person web measure similarity name interest annotation base team profile number system share find relation concept work
	26	service web ontology use process model concept base composition approach rule qos set description state constraint example define provider provide system information owl may instance context execution describe match axiom
	28	treatment claim source effect group causal true data variable control model experiment use truth estimate distribution value fact set make prior match outcome unit credibility parameter reliability figure evidence assertion
	33	model feature learn performance dataset network attention layer neural sequence prediction train use method datasets propose state task deep baseline representation lstm vector input base embed figure time interaction information
	38	email influence flow information model user time chain diffusion reply use work company network number figure factor transition base job sender data receive social also give process probability study show
	41	user social cascade facebook post feature group number network time model friend figure hashtags show discussion distribution content comment activity also study large online predict use set observe size share
	52	mobile apps app device use performance application network time model energy data show dl user figure android developer result signal browser different permission run number deep platform measurement support cloud
	56	event news time topic blog medium temporal information story source trend use attention show feed series post interest analysis different content set detection data figure country article work goal day
	57	rating user model use preference item rank comment restaurant data method movie show value set matrix base latent distribution high group approach rat number low give result learn different bias
	72	feature classifier label classification class set use train learn data score training accuracy tree performance positive instance sample number base category svm example detection dataset test approach method result bias
JASSS	0	model democracy society polity complex system social political simple world state dynamic country power global non democratic change data economic theory war simulation time see development peasant transition complexity also
	4	model income policy economic tax level region household rate consumption result increase base agent high change market doi firm price cost economy work effect low al et value parameter distribution
	6	agent belief model resource level time simulation social number society may population communication set probability case experiment information environment collective state make action process base system initial result also increase
	21	model agent household data flood base house use et simulation al number housing level year population process figure time area change result urban different city location new center homeowner income
	24	simulation method data output algorithm number match use microsimulation fit set example variable probability table result test alignment mean prediction sample observation pair time show order weight different distance measure
	48	bank interbank financial loss risk network institution asset al et doi system figure channel contagion data market default ast cross systemic liability rule total customer use banking shareholding show increase
	65	social research science simulation model review journal scientist agent community scientific base number fund proposal year jasss project author paper system publication study result topic network time funding publish society
	71	opinion model social influence agent doi time group dynamic polarization et al value show different individual network change journal effect evolution simulation figure base result interaction confidence cluster process event
	72	energy model agent system electricity decision social base technology use al et change charge policy different value simulation figure scenario demand environmental household actor diffusion factor power result information transition

TABLE 4.3: Top 3 papers with the highest innovation scores in both corpora with citation counts and topics covered.

	Year	Author(s) and Title	Score	Citations	Topics
WWW	2011	C. Budak, D. Agrawal, A. El Abbadi, *Limiting the Spread of Misinformation in Social Networks*	0.971	607	9, 12, 38, 41, 56
	2010	A. Sala, L. Cao, Ch. Wilson, R. Zablit, H. Zheng, B. Y. Zhao, *Measurement-calibrated Graph Models for Social Network Experiments*	0.963	189	2, 9, 15, 41, 52
	2018	H. Wu, Ch. Wang, J. Yin, K. Lu, L. Zhu, *Sharing Deep Neural Network Models with Interpretation*	0.955	7	33, 72
JASSS	2001	K. Auer, T. Norris, *"ArrierosAlife" a Multi-Agent Approach Simulating the Evolution of a Social System: Modeling the Emergence of Social Networks with "Ascape"*	0.868	13	6, 21
	2000	B. G. Lawson, S. Park, *Asynchronous Time Evolution in an Artificial Society Model*	0.841	13	6, 24, 71
	2008	R. Bhavnani, D. Miodownik, J. Nart, *REsCape: an Agent-Based Framework for Modeling Resources, Ethnicity, and Conflict*	0.788	51	0, 72

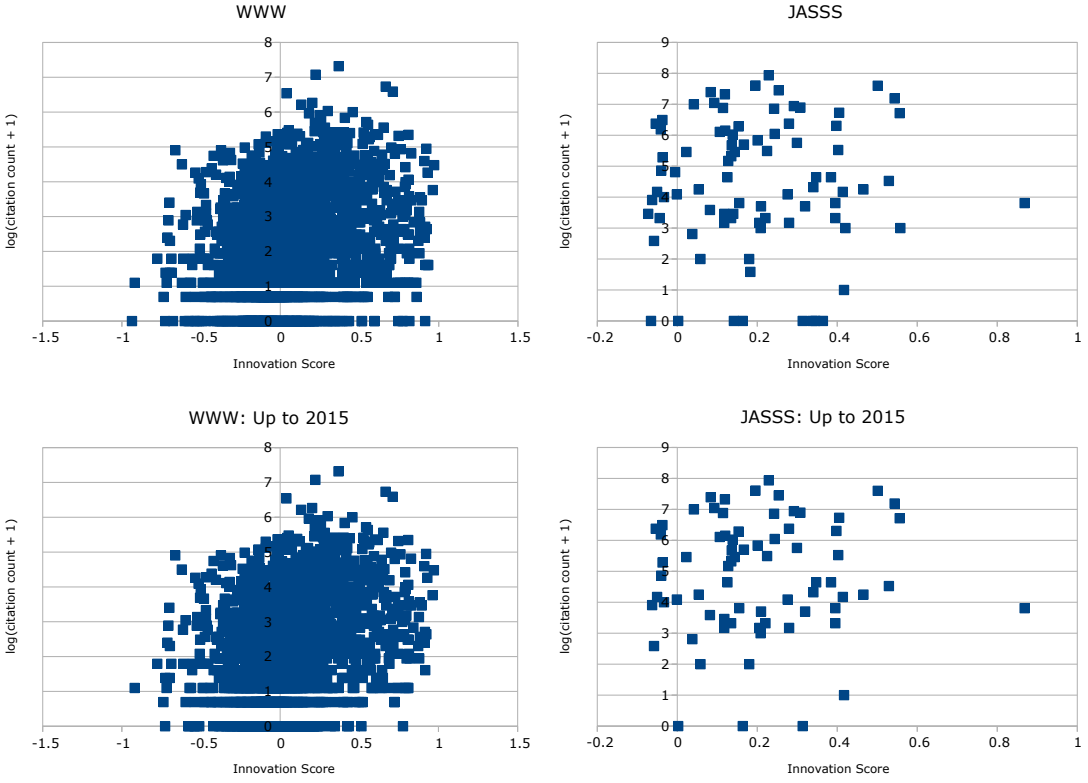

FIGURE 4.5: Innovation score vs. Citation count for all papers (above) and papers at least 5 years old (below).

$log_2(citation\ count + 1)$ instead of raw citation counts. The value of this expression is zero if the number of citations is zero and grows monotonically as the number of citations increases. The citation data for the WWW corpus come from ACM's Digital Library[2], however publications from the JASSS journal are not available in the ACM DL. We were also unable to scrape complete citation data from Google Scholar. We have therefore manually collected citation counts for 5 randomly selected papers from each year. We

[2] `http://dl.acm.org/`

have calculated Spearman's ρ correlation coefficients between the innovation scores and citation counts. The results are: 0.28 with a p-value of $1.21 \cdot 10^{-41}$ for the WWW corpus and 0.32 with a p-value of $1.91 \cdot 10^{-6}$ for JASSS. The innovation scores are, therefore, weakly correlated to the citation counts. The correlation coefficients are slightly higher for papers at least 5 years old: 0.3 for WWW and 0.37 for JASSS. This may be explained by the fact that newer papers have not yet accumulated many citations regardless of their innovativeness.

4.6 Conclusion and Future Work

We have shown a simple yet significant improvement to our novel method of measuring the innovativeness of scientific papers in bodies of research spanning multiple years. Scaling the innovation score proposed in our previous research has enabled us to directly compare the scores of papers published at different years. We have also improved the prediction accuracy by employing ordinal regression models instead of regular multiclass classifiers and Correlated Topic Models instead of LDA. It may be argued that this makes our method more reliable, as deviations of the predicted publication year from the actual one are more likely to be caused by the paper actually covering topics popular in the future rather than just being usual prediction error. Moreover, CTM allowed to better model and understand the evolution of research topics over time.

In the future we plan to explore non-linear ways to scale the innovation scores, taking into account the observed error distribution (Fig. 4.3) to give more weight to larger deviations from the expected value. We also plan to use word embeddings or extracted scientific claims [175] as well as other means of effectively representing paper contents and conveyed ideas besides topic models as features to our methods.

Pavel Savov, Adam Jatowt, and Radoslaw Nielek. Predicting the Age of Scientific Papers. *International Conference on Computational Science*. Springer, Cham, 2021.

Chapter 5

Predicting the Age of Scientific Papers

In this paper we show how the age of scientific papers can be predicted given a diachronic corpus of papers from a particular domain published over a certain time period. We first train ordinal regression models for the task of predicting the age of individual sentences by fine-tuning series of BERT models for binary classification. We then aggregate the prediction results on individual sentences into a final result for entire papers. Using two corpora of publications from the International World Wide Web Conference and the Journal of Artificial Societies and Social Simulation, we compare various result aggregation methods, and show that the sentence-based approach produces better results than the direct document-level method.

5.1 Introduction

Document dating or timestamping is the process of inferring the age of a document, if it is either unknown or unreliable, based on its textual content. In the scientific domain, publication dates of documents are usually known, but the results of document timestamping may be used to complement traditional scientometric methods in assessing the innovativeness of research papers [176] or identifying novelty. At a basic level, the larger the difference between the actual timestamp and the predicted timestamp of a target scientific document, the higher is its potential innovativeness or novelty of the target paper. This may be useful to non-expert readers of technical documents, such as potential investors or decision makers at funding bodies, who wish to know how new or innovative the ideas or methods covered by these documents were at the time of their creation.

Furthermore, in practical scenarios, the timestamping models specialized for scientific corpora can also be applied to other types of documents that may discuss scientific technology and domain-focused research, or quote content from scientific papers. Such documents may not have explicit timestamps (e.g., web pages) and the determination of their age (as well as the related concept of timeliness) can be useful in many cases. Thus, in general, scientific document age prediction can be used for discovering the content parts in a scientific publication that are novel or innovative, or perhaps obsolete/outdated when considering the document publication date [176] as well as for determining the age of science-related content in non-scholarly documents that lack timestamps.

In this paper we focus on improving the accuracy of scientific paper age prediction by using state-of-the-art word embedding models trained on two corpora of papers from related but distinct domains, published at leading publication venues in their respective fields. Typical approaches to automatic document dating are based on modeling language change over time and shifts in word usage. Examples of temporal language models, i.e. time series of statistical language models include [112, 116]. Jatowt and Campos [117] have implemented an online visual and interactive system based on n-gram frequency analysis. Garcia-Fernandez et al. [118] used SVM classifiers on feature vectors of word and n-gram frequencies. Ordinal regression models were used for document dating by Niculae et al. [121], or Popescu and Strapparava [122]. Another approach to temporal language modeling are neural language models based on word embeddings such as Word2Vec [108]. Kim et al. [125] studied the shift in word semantics over time by training a model for each time interval and then plotting the words' cosine similarities to their reference points. Soni et al. [177] used diachronic word embeddings to show that scientific papers using words in their newer meanings tend to receive more citations. Vashishth et al. [178] proposed a deep learning approach to document dating, exploiting syntactic and temporal document graph structures. Unlike the above-mentioned methods, which work mainly on news articles or generic documents, we focus on a particular genre of scholarly publications. We also approach the document dating task at a sentence-level, and we test several sentence aggregation approaches.

5.2 Datasets

We study the following two corpora: (1) *WWW*: 3,896 papers published at the International World Wide Web Conference between 1994 and 2020, containing 1,037,051 sentences, (2) *JASSS*: 884 articles published in the Journal of Artificial Societies and Social Simulation[1] between 1998 and 2020, containing 321,589 sentences. Both corpora

[1] `http://jasss.soc.surrey.ac.uk/`

contain entire papers. However, we have removed page headers and footers, *References*, *Bibliography* and *Acknowledgments* sections as "noise" irrelevant to the papers' contents. All papers published in the JASSS journal are available in HTML at the journal's website[1]. Papers from the proceedings of the WWW conference are available at `https://thewebconf.org/` in different formats for different years. Most are available in PDF, some in HTML and a small number of older papers in PostScript. We used the *pdftotext* tool[2] to extract plain text from PDF documents. We divided the documents into sentences using the Punkt sentence tokenizer for the English language implemented in the Natural Language Toolkit (NLTK) Python library [179]. Conversion to lower case and tokenization were performed by the BERT tokenizer.

5.3 Method

We propose to approach the problem of scientific document's age prediction by first predicting the age of its sentences. Thanks to focusing on sentences instead of entire documents we can use more labelled data instances for training, which is quite important for relatively narrow scientific domains with constrained datasets (e.g., proceedings of conferences dedicated to a particular research sub-field). Thus, our approach is composed of two steps: (1) predicting the age of sentences and (2) aggregating sentence age to determine the document age. We describe these two steps below.

5.3.1 Predicting Sentence Age

As time units are clearly ordinal values, we predict the age of individual sentences by means of Ordinal Regression, a.k.a. Ordinal Classification, based on the framework proposed by Li and Lin [173]. Ordinal Regression was also used by Martin et al. [174] for photograph dating. An N-class ordinal regression model consists of $N-1$ *before-after* binary classifiers, i.e. for each pair of consecutive years a classifier is trained, which assigns sentences to one of two classes: "year y or before" and "year $y+1$ or after". Given the class membership probabilities predicted by these classifiers, the overall classifier confidence that sentence s was written in the year Y is then determined, as in [174], by Eqs. 5.1 and 5.2:

$$conf(s,Y) = \prod_{y=Y_{min}}^{Y} P(Y_s \leqslant y) \cdot \prod_{y=Y+1}^{Y_{max}} (1 - P(Y_s \leqslant y)) \tag{5.1}$$

[2] `https://www.xpdfreader.com/pdftotext-man.html`

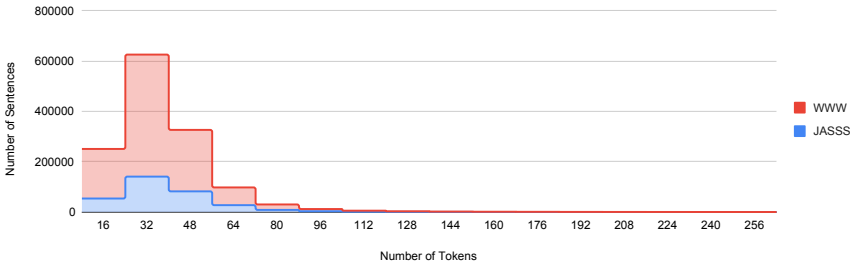

FIGURE 5.1: Number of Tokens per Sentence

where Y_{min} and Y_{max} are the first and last year in the corpus, and Y_s is the publication year of the paper that s comes from.

Thus, the predicted year for the sentence s is:

$$\hat{Y_s} = \underset{y \in [Y_{min}, Y_{max}]}{\text{argmax}} \; conf(s, y) \tag{5.2}$$

Unlike the approaches of [173] and [174], we used the Huggingface Transformers[3] [180] Python library to fine-tune SciBERT models [181] for sequence classification in binary *before-after* classification. SciBERT is a BERT [182] model trained on 1.14M scientific papers from the `semanticscholar.org` corpus. The maximum sequence length supported out-of-the-box is 512, however over 95% of the sentences in our corpora contain up to 64 tokens (see Fig. 5.1). We have, therefore, decided to cap the maximum sequence length at 64. We have not observed any significant differences in the predictive performance of the models, expressed as Mean Absolute Error, for maximum sequence lengths of 64, 128, and 512 tokens. We trained each model for two epochs, the batch size was 32, and the learning rate: 2e-5. The BERT authors recommend fine-tuning the models for 2 to 4 epochs, but we have found our models to overfit the training data when fine-tuned for more than 2 epochs. In most cases the differences in average loss and accuracy on the validation set for models trained for two epochs vs. one were minimal.

We have made an 80/20 split on the document level so as to make sure of the clean separation of training and testing sentences. Although our approach yielded poor prediction results on the sentence-level (4.49 years for JASSS and 3.56 years for WWW, see Fig. 5.2), as we will show later, the final prediction of document age produces quite good results.

[3] `https://huggingface.co/transformers/`

5.3.2 Predicting Document Age

As stated above, we predict the age of entire papers by aggregating the results of individual sentence age prediction using various aggregation functions. We have experimented with rejecting sentences for which the model's confidence was below a certain threshold in the range from 0 to 0.5. For values greater than 0.5 in some documents no sentences exceeded that threshold.

Newest Sentence As a baseline approach we assume the age of the paper p equals the age of its newest sentence. Since most papers contain at least one sentence the most probable age of which is predicted as 0 years, we only take into account the sentence predicted as the newest among those, for which the model's confidence exceeds 0.5. This value was chosen, as it gave the best results.

Topic distribution based classifier As another baseline approach, which works purely on the document-level, we used a method based on SVM classifier on vectors of latent topic distributions derived from document collections [176].

Arithmetic Mean In this approach we calculated the predicted age of paper p as the mean predicted age of all its sentences.

Weighted Mean w/Sentence Offset We assumed that the sooner a sentence appears in the paper, the more important it is. We, therefore, defined the predicted age of paper p as the weighted mean predicted age of its sentences, where the weight of each sentence was its ordinal number within the paper p divided by the number of sentences in p:

$$\hat{Y_p} = \frac{\sum_{s \in p} \hat{Y_s} \cdot \frac{n_s}{|\{s \in p\}|}}{\sum_{s \in p} \frac{n_s}{|\{s \in p\}|}}$$

where n_s is the ordinal number of the sentence s within p.

This concept is a simplified approach to weighted zoning [183], where each sentence is assigned a weight, depending on which section of the paper it appears in, e.g. Abstract: 1, Introduction: 0.8, Related Work: 0.3, everything else: 0.5.

Weighted Mean w/TextRank TextRank by Mihalcea and Tarau [154] is an unsupervised graph-based algorithm for keyword extraction and text summarization, based on PageRank [18]. Its variant for text summarization finds the most important sentences by running a variation of PageRank on a graph, whose vertices represent the document's sentences. Each edge has a weight corresponding to the similarity of the sentences represented by the vertices connected by that edge. In contrast to PageRank, the graph constructed by TextRank is undirected, since the similarity between sentences is symmetric. Various sentence similarity measures may be used, but Barrios et al. [184] showed

that a variation of the Okapi-BM25 [185] ranking function, which is itself a variation of the TF-IDF model using a probabilistic model, yields the best results. We used the implementation of TextRank with the BM25 ranking function from the *gensim*[4] Python library to find importance scores for all sentences in each document. We then used these scores as weights to calculate the predicted publication year of each paper p defined as the weighted mean of the years of its sentences:

$$\hat{Y_p} = \frac{\sum_{s \in p} Imp_p^s \cdot \hat{Y_s}}{\sum_{s \in p} Imp_p^s}$$

where Imp_p^s is the TextRank importance score of s within the paper p.

Citation Removal In this approach we make the assumption that any sentences citing other papers are unimportant for the content of the paper being analyzed or introduce concepts and ideas from older papers (hence potentially negatively impacting the age detection process). Thus, we remove all sentences containing citations and proceed to calculate the predicted publication year using any of the approaches described above. As shown in Section 5.4, in most cases citation removal improves the prediction results in terms of Mean Absolute Error. Another possible extension could be removing entire *Related Work* sections.

5.4 Results

As stated before, the mean absolute age prediction error (MAE) for individual sentences is 4.49 years for the JASSS corpus and 3.56 for WWW. The prediction error distribution is shown in Fig. 5.2. Although these results are not satisfactory, we obtain much better results for entire documents. As shown in Tab. 5.1, the sentence-based approach aggregating individual predictions of many sentences gives much better results in predicting paper publication dates. Except for the naive *newest sentence* baseline, the MAE is always less than 1 year. Also the document level approach proposed in [176] performs much worse.

Weighting the sentence age predictions by sentence offsets performed better on the WWW corpus, while TextRank weights gave better results for JASSS. In all cases, however, removing sentences containing citations improved the document age predictions significantly. This supports our assumption that sentences citing other articles could introduce noise.

[4] `https://radimrehurek.com/gensim_3.8.3/index.html`

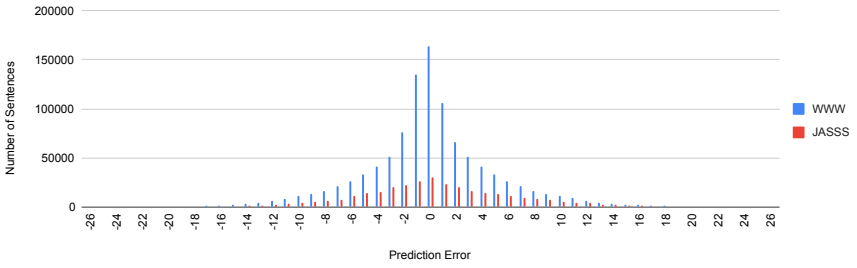

FIGURE 5.2: Sentence Prediction Error Distributions

TABLE 5.1: Results of prediction methods (Mean Absolute Error: #years).

	WWW		JASSS	
Document-level [176]	2.56		3.56	
Sentence-level	All Sentences	Citations Removed	All Sentences	Citations Removed
Newest Sentence	8.959	8.946	8.267	8.33
Arithmetic Mean	0.833	0.816	0.743	0.67
Weighted Mean w/Sentence Offset	0.709	**0.684**	0.738	0.645
Weighted Mean w/TextRank	0.741	0.725	0.67	**0.636**

5.5 Conclusions and Future Work

In this paper we have shown how the accuracy of scientific paper age prediction can be improved by using state-of-the-art word embedding models at the sentence level, and then aggregating the results. Interestingly, for all aggregation methods except for the most basic baseline approach, i.e. *newest sentence*, increasing the value of the confidence threshold led to worse results. This suggests that unless sentences are rejected based on domain-specific knowledge, e.g. rejecting sentences containing citations, the more predictions are aggregated into the final result the better, similarly to the "wisdom of the crowds" effect, where the aggregated predictions of multiple agents are far closer to the actual value than most of the individual predictions [186]. Finally, we note that as our approach works on the sentence-level, it could also be used to assess the age of text excerpts (e.g., in web pages) about specialized scientific topics, and, therefore, potentially help readers better understand their actual novelty and age.

Having achieved a mean prediction error of less than a year, we plan on experimenting with datasets having narrower time slices, e.g. the Covid-19 dataset from Kaggle[5]. We will also try weighting sentences containing scientific claims [175].

[5] `https://www.kaggle.com/imdevskp/corona-virus-report`

Chapter 6

Conclusions and Future Work

We have demonstrated a novel classification-based method of measuring innovation, which may be used to complement citation analysis in identifying potential breakthrough publications in bodies of research spanning multiple years. We proposed a real-number measure of paper innovativeness based on the prediction error of the publication year and the classifier's confidence. We also showed how to adjust this measure to allow for comparing the scores of papers published in different years, as well as aggregate them to compare the innovativeness of different years. In Chapter 5 we have shown that the accuracy of age prediction may be improved greatly by using BERT – a state-of-the-art word embedding model to predict the age of individual sentences and aggregating the results. Finally, we applied our method to three corpora of papers from leading publication venues in their respective fields, and compared our results with citation counts. The innovation scores and citation counts were moderately correlated for older papers and weakly correlated for newer ones. A possible explanation of this phenomenon would be the fact that more recent papers, even the "innovative" ones, have not yet accumulated enough citations to reflect their importance, as accumulating citations is a slow process[11].

The most important contribution and the main advantage of our method is its ability to fully automate the analysis of the corpus. None of its steps – topic model training and selection, classifier training and score calculation – require expert knowledge or manual intervention.

The main limitation of the proposed method is that it only captures a snapshot in time. As new papers in the studied domain (journal, conference, etc.) are published and new time slices are added, the topic model as well as the prediction model need to be retrained from scratch. The latent topics discovered in the updated corpus may change

completely. However, this problem disappears when the approach with topic modeling is replaced with word embeddings as in Chapter 5.

Previously calculated innovation scores (see Sections 3.4.4 and 4.4.3) may also change as new time slices are added to the corpus. Let us consider a scenario, where papers P_1 and P_2, both published in the same year Y cover topics T_1 and T_2 respectively, both of which appear in the year Y for the first time. Obviously, when analyzing the corpus up to Y, both papers will likely receive high scores. However, let us now suppose that the topic T_1 never appears again, and T_2 gains popularity and continues to be popular for a number of years. When analyzing the updated corpus after a few years, P_1's score will decrease, but P_2's score will remain high.

Another potential weakness of our method is its sensitivity to shifts in the scope of the analyzed publication venues. Early occurrences of a specific topic at a particular conference may not necessarily be groundbreaking, as this topic may have already been covered elsewhere. This may be remedied by training the models on papers from multiple venues in a given domain. On the other hand, as discussed in section 2.7.2, the occurrence of previously researched topics in a new context may indicate innovation.

The most obvious direction for the future is reformulating the paper innovation score proposed in Chapters 3 and 4 (Equations 3.2 and 4.4) using aggregated sentence age predictions (see Chapter 5) weighted by the predictions' probabilities. The aggregation function is to be proposed.

It would also be worth exploring normalizing the innovation scores by non-linear functions giving more weight to larger deviations of the predicted publication date from its expected value. As illustrated by Figures 3.3, 4.3, and 5.2, the observed prediction error distribution is non-linear, and small deviations from the expected value are more likely than large ones, and therefore - less significant.

Bibliography

[1] Eugene Garfield. Citation indexes for science. *Science*, 122(3159):108–111, 1955.

[2] Derek J De Solla Price. Little science, big science. 1963.

[3] Jesper W Schneider and Rodrigo Costas. Identifying potential 'breakthrough' research articles using refined citation analyses: Three explorative approaches. *STI 2014 Leiden*, page 551, 2014.

[4] Jesper W Schneider and Rodrigo Costas. Identifying potential "breakthrough" publications using refined citation analyses: Three related explorative approaches. *Journal of the Association for Information Science and Technology*, 68(3):709–723, 2017.

[5] Ilya V Ponomarev, Duane E Williams, Charles J Hackett, Joshua D Schnell, and Laurel L Haak. Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change*, 81:49–55, 2014.

[6] Holly N Wolcott, Matthew J Fouch, Elizabeth R Hsu, Leo G DiJoseph, Catherine A Bernaciak, James G Corrigan, and Duane E Williams. Modeling time-dependent and-independent indicators to facilitate identification of breakthrough research papers. *Scientometrics*, 107(2):807–817, 2016.

[7] Howard D White. Citation analysis and discourse analysis revisited. *Applied linguistics*, 25(1):89–116, 2004.

[8] Robert K Merton. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63, 1968.

[9] Mayank Singh, Ajay Jaiswal, Priya Shree, Arindam Pal, Animesh Mukherjee, and Pawan Goyal. Understanding the impact of early citers on long-term scientific impact. In *Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on*, pages 1–10. IEEE, 2017.

[10] Alexander Serenko and John Dumay. Citation classics published in knowledge management journals. part ii: studying research trends and discovering the google scholar effect. *Journal of Knowledge Management*, 19(6):1335–1355, 2015.

[11] Jason Priem, Dario Taraborelli, Paul Groth, and Cameron Neylon. Altmetrics: A manifesto. 2010.

[12] Derek J De Solla Price. Networks of scientific papers. *Science*, pages 510–515, 1965.

[13] Paul Benjamin Lowry, James Gaskin, Sean L Humpherys, Gregory D Moody, Dennis F Galletta, Jordan B Barlow, and David W Wilson. Evaluating journal quality and the association for information systems senior scholars' journal basket via bibliometric measures: Do expert journal assessments add value? *MIS quarterly*, pages 993–1012, 2013.

[14] Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178 (4060):471–479, 1972.

[15] Eugene Garfield. The impact factor and using it correctly. *Der Unfallchirurg*, 48 (2):413, 1998.

[16] Mike Rossner, Heather Van Epps, and Emma Hill. Show me the data, 2008.

[17] Carl T Bergstrom, Jevin D West, and Marc A Wiseman. The eigenfactor^{TM} metrics. *Journal of neuroscience*, 28(45):11433–11434, 2008.

[18] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[19] P Davis. Network-based citation metrics: Eigenfactor vs. sjr. *Scholarly Kitchen blog*, 28, 2015.

[20] Henk F Moed. Measuring contextual citation impact of scientific journals. *Journal of informetrics*, 4(3):265–277, 2010.

[21] Cameron Neylon and Shirley Wu. level metrics and the evolution of scientific impact. *PLoS biol*, 7(11):e1000242, 2009.

[22] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.

[23] Pablo D Batista, Mônica G Campiteli, and Osame Kinouchi. Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1):179–189, 2006.

[24] Jasleen Kaur, Filippo Radicchi, and Filippo Menczer. Universality of scholarly impact metrics. *Journal of Informetrics*, 7(4):924–932, 2013.

[25] Teja Tscharntke, Michael E Hochberg, Tatyana A Rand, Vincent H Resh, and Jochen Krauss. Author sequence and credit for contributions in multiauthored publications. *PLoS Biol*, 5(1):e18, 2007.

[26] Michael Schreiber. Restricting the h-index to a publication and citation time window: A case study of a timed hirsch index. *Journal of Informetrics*, 9(1): 150–155, 2015.

[27] Jevin D West, Michael C Jensen, Ralph J Dandrea, Gregory J Gordon, and Carl T Bergstrom. Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the American Society for Information Science and Technology*, 64(4): 787–801, 2013.

[28] Mark EJ Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12, 2008.

[29] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

[30] Ying Ding. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics*, 5(1):187–203, 2011.

[31] Rodrigo De Castro and Jerrold W Grossman. Famous trails to paul erdős. *The Mathematical Intelligencer*, 21(3):51–53, 1999.

[32] Jennifer Lin and Martin Fenner. Altmetrics in evolution: Defining and redefining the ontology of article-level metrics. *Information standards quarterly*, 25(2):20, 2013.

[33] Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. Do "altmetrics" correlate with citations? extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10):2003–2019, 2015.

[34] Zhichao Fang and Rodrigo Costas. Studying the accumulation velocity of altmetric data tracked by altmetric. com. *Scientometrics*, pages 1–25, 2020.

[35] John W Lounsbury, Karol G Roisum, Lois Pokorny, Abigail Sills, and Gregory J Meissen. An analysis of topic areas and topic trends in the community mental health journal from 1965 through 1977. *Community mental health journal*, 15(4): 267–276, 1979.

[36] James R. Wallace, Saba Oji, and Craig Anslow. Technologies, methods, and values: Changes in empirical research at cscw 1990 - 2015. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):106:1–106:18, December 2017. ISSN 2573-0142. doi: 10.1145/3134741. URL `http://doi.acm.org/10.1145/3134741`.

[37] Thomas Hoffman. Probabilistic latent semantic analysis. In *proc. of the 15th Conference on Uncertainty in AI, 1999*, 1999.

[38] David M. Blei, Andrew Y. Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4–5):993–1022, 2003.

[39] Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269, 1973.

[40] Henry Small. Co-citation context analysis and the structure of paradigms. *Journal of documentation*, 1980.

[41] Henry Small and Edward Greenlee. Citation context analysis of a co-citation cluster: Recombinant-dna. *Scientometrics*, 2(4):277–301, 1980.

[42] Henry Small. Macro-level changes in the structure of co-citation clusters: 1983–1989. *Scientometrics*, 26(1):5–20, 1993.

[43] Matthias Meyer, Iris Lorscheid, and Klaus G. Troitzsch. The development of social simulation as reflected in the first ten years of jasss: a citation and co-citation analysis. *Journal of Artificial Societies and Social Simulation*, 12(4):12, 2009. ISSN 1460-7425. URL `http://jasss.soc.surrey.ac.uk/12/4/12.html`.

[44] Jonas Hauke, Iris Lorscheid, and Matthias Meyer. Recent development of social simulation as reflected in jasss between 2008 and 2014: A citation and co-citation analysis. *Journal of artificial societies and social simulation*, 20(1), 2017.

[45] Bela Gipp and Jöran Beel. Citation proximity analysis (cpa): A new approach for identifying related work based on co-citation analysis. In *ISSI'09: 12th international conference on scientometrics and informetrics*, pages 571–575, 2009.

[46] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[47] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[48] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84, 2012.

[49] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[50] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584, 2006.

[51] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143859. URL `http://doi.acm.org/10.1145/1143844.1143859`.

[52] Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*, 2012.

[53] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

[54] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *2008 eighth IEEE international conference on data mining*, pages 3–12. IEEE, 2008.

[55] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[56] Andre Gohr, Alexander Hinneburg, Rene Schult, and Myra Spiliopoulou. Topic evolution in a stream of documents. In *Proceedings of the 2009 SIAM international conference on data mining*, pages 859–870. SIAM, 2009.

[57] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 198–207, 2005.

[58] David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the History of Ideas Using Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 363–371, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1613715.1613763`.

[59] Lijun Sun and Yafeng Yin. Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C: Emerging Technologies*, 77:49–66, 2017.

[60] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009.

[61] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc., 2009. URL `http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf`.

[62] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108, 2010.

[63] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, 2011.

[64] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3317-7. doi: 10.1145/2684822.2685324. URL `http://doi.acm.org/10.1145/2684822.2685324`.

[65] Baitong Chen, Satoshi Tsutsui, Ying Ding, and Feicheng Ma. Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics*, 11(4):1175–1189, 2017.

[66] Yee Whye Teh, Matthew J. Beal, Michael I. Jordan, and David M. Blei. Hierarchical dirichlet processes, 2003.

[67] Wei Li, David Blei, and Andrew McCallum. Nonparametric bayes pachinko allocation. *arXiv preprint arXiv:1206.5270*, 2012.

[68] David Mimno, Andrew McCallum, and Gideon S Mann. Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries (JCDL'06)*, pages 65–74. IEEE, 2006.

[69] Xuerui Wang and Andrew McCallum. A note on topical n-grams. Technical report, MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE, 2005.

[70] Kai Hu, Qing Luo, Kunlun Qi, Siluo Yang, Jin Mao, Xiaokang Fu, Jie Zheng, Huayi Wu, Ya Guo, and Qibing Zhu. Understanding the topic evolution of scientific literatures like an evolving city: Using google word2vec model and spatial autocorrelation analysis. *Information Processing & Management*, 56(4):1185–1203, 2019.

[71] David A Cohn and Thomas Hofmann. The missing link-a probabilistic model of document content and hypertext connectivity. In *Advances in neural information processing systems*, pages 430–436, 2001.

[72] Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101 (suppl 1):5220–5227, 2004.

[73] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, pages 233–240, 2007.

[74] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 957–966, 2009.

[75] Ramesh M Nallapati, Amr Ahmed, Eric P Xing, and William W Cohen. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550, 2008.

[76] Alan F. Smeaton, Gary Keogh, Cathal Gurrin, Kieran McDonald, and Tom Sødring. Analysis of papers from twenty-five years of sigir conferences: What have we been doing for the last quarter of a century? *SIGIR Forum*, 36(2): 39–43, September 2002. ISSN 0163-5840. doi: 10.1145/792550.792556. URL `http://doi.acm.org/10.1145/792550.792556`.

[77] Danilo Saft and Volker Nissen. Analysing full text content by means of a flexible co-citation analysis inspired text mining method - exploring 15 years of jasss articles. *International Journal of Business Intelligence and Data Mining*, 9(1):52–73, 2014.

[78] Henning Pohl and Aske Mottelson. How we guide, write, and cite at chi. 2019.

[79] Len Thomas. Monitoring long-term population change: why are there so many analysis methods? *Ecology*, 77(1):49–58, 1996.

[80] Brent A Coull, Joel Schwartz, and MP Wand. Respiratory health and air pollution: additive mixed model analyses. *Biostatistics*, 2(3):337–349, 2001.

[81] M Ugur Gudelek, S Arda Boluk, and A Murat Ozbayoglu. A deep learning based stock trading model with 2-d cnn trend detection. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE, 2017.

[82] Katharine Lynn Gray. Comparison of trend detection methods. 2007.

[83] Michael Färber and Adam Jatowt. Finding temporal trends of scientific concepts. In *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2019) co-located with the 41st European Conference on Information Retrieval (ECIR 2019), Cologne, Germany, April 14, 2019.*, pages 132–139, 2019. URL `http://ceur-ws.org/Vol-2345/paper12.pdf`.

[84] Michael Färber, Chifumi Nishioka, and Adam Jatowt. Scholarsight: visualizing temporal trends of scientific concepts. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 438–439. IEEE, 2019.

[85] Ketan K Mane and Katy Börner. Mapping topics and topic bursts in pnas. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5287–5290, 2004.

[86] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[87] Henry Small, Kevin W Boyack, and Richard Klavans. Identifying emerging topics in science and technology. *Research policy*, 43(8):1450–1467, 2014.

[88] April Kontostathis, Leon M Galitsky, William M Pottenger, Soma Roy, and Daniel J Phelps. A survey of emerging trend detection in textual data mining. In *Survey of text mining*, pages 185–224. Springer, 2004.

[89] Shenhao Jiang, Animesh Prasad, Min-Yen Kan, and Kazunari Sugiyama. Identifying emergent research trends by key authors and phrases. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 259–269, 2018. URL `https://aclanthology.info/papers/C18-1022/c18-1022`.

[90] Babak Sohrabi and Ahmad Khalilijafarabad. Systematic method for finding emergence research areas as data quality. *Technological Forecasting and Social Change*, 137(C):280–287, 2018.

[91] Levent Bolelli, Şeyda Ertekin, and C Lee Giles. Topic and trend detection in text collections using latent dirichlet allocation. In *European Conference on Information Retrieval*, pages 776–780. Springer, 2009.

[92] Yookyung Jo, Carl Lagoze, and C Lee Giles. Detecting research topics via the correlation between graphs and texts. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 370–379, 2007.

[93] Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 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, pages 1247–1252, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063757. URL `http://doi.acm.org/10.1145/2063576.2063757`.

[94] Rui Yan, Congrui Huang, Jie Tang, Yan Zhang, and Xiaoming Li. To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, pages 51–60, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1154-0. doi: 10.1145/2232817.2232831. URL `http://doi.acm.org/10.1145/2232817.2232831`.

[95] H Inhaber and KJSSoS Przednowek. Quality of research and the nobel prizes. *Social Studies of Science*, 6(1):33–50, 1976.

[96] Eugene Garfield. Do nobel-prize winners write citation-classics. *Current Contents*, (23):3–8, 1986.

[97] Eugene Garfield and Alfred Welljams-Dorof. Of nobel class: A citation perspective on high impact research authors. *Theoretical medicine*, 13(2):117–135, 1992.

[98] András Schubert, Wolfgang Glänzel, and Tibor Braun. Subject field characteristic citation scores and scales for assessing research performance. *Scientometrics*, 12 (5-6):267–291, 1987.

[99] Ilya V Ponomarev, Duane E Williams, Brian K Lawton, Di H Cross, Yvette Seger, Joshua Schnell, and Laurel L Haak. Breakthrough paper indicator: early detection and measurement of ground-breaking research. In *CRIS*, 2012.

[100] JJ Winnink and Robert JW Tijssen. Early stage identification of breakthroughs at the interface of science and technology: lessons drawn from a landmark publication. *Scientometrics*, 102(1):113–134, 2015.

[101] Kostya S Novoselov, Andre K Geim, Sergei V Morozov, D Jiang, Y_ Zhang, Sergey V Dubonos, Irina V Grigorieva, and Alexandr A Firsov. Electric field effect in atomically thin carbon films. *science*, 306(5696):666–669, 2004.

[102] B Hesse Mary. Models and analogies in science, 1966.

[103] Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.

[104] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. Solvent: A mixed initiative system for finding analogies between research papers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):31:1–31:21, November 2018. ISSN 2573-0142. doi: 10.1145/3274300. URL `http://doi.acm.org/10.1145/3274300`.

[105] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

[106] Joel Chan, Tom Hope, Dafna Shahaf, and Aniket Kittur. Scaling up analogy with crowdsourcing and machine learning. In *ICCBR Workshops*, pages 31–40, 2016.

[107] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 235–243, 2017.

[108] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[109] D. Thorleuchter and D. Van den Poel. Identification of interdisciplinary ideas. *Information Processing & Management*, 52(6):1074–1085, November 2016. ISSN 0306-4573. doi: 10.1016/j.ipm.2016.04.010. URL `https://doi.org/10.1016/j.ipm.2016.04.010`.

[110] Anthony FJ Van Raan. Sleeping beauties in science. *Scientometrics*, 59(3):467–472, 2004.

[111] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

[112] Franciska De Jong, Henning Rode, and Djoerd Hiemstra. Temporal language models for the disclosure of historical text. In *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, pages 161–168, Amsterdam, the Netherlands, 2005. Koninklijke Nederlandse Academie van Wetenschappen.

[113] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[114] Angelo Dalli and Yorick Wilks. Automatic dating of documents and temporal text classification. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 17–22, 2006.

[115] Nattiya Kanhabua and Kjetil Nørvåg. Improving temporal language models for determining time of non-timestamped documents. In *International Conference on Theory and Practice of Digital Libraries*, pages 358–370. Springer, 2008.

[116] Nattiya Kanhabua and Kjetil Nørvåg. Using temporal language models for document dating. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 738–741. Springer, 2009.

[117] Adam Jatowt and Ricardo Campos. Interactive system for reasoning about document age. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 2471–2474, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4918-5. doi: 10.1145/3132847.3133166. URL `http://doi.acm.org/10.1145/3132847.3133166`.

[118] Anne Garcia-Fernandez, Anne-Laure Ligozat, Marco Dinarelli, and Delphine Bernhard. When was it written? automatically determining publication dates. In *International Symposium on String Processing and Information Retrieval*, pages 221–236, Berlin, Heidelberg, 2011. Springer.

[119] Alina Maria Ciobanu, Anca Dinu, Liviu Dinu, Vlad Niculae, and Octavia-Maria Şulea. Temporal classification for historical romanian texts. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 102–106, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[120] Haritz Salaberri, Iker Salaberri, Olatz Arregi, and Benat Zapirain. Ixagroupehudiac: A multiple approach system towards the diachronic evaluation of texts. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 840–845, Denver, Colorado, 2015. Association for Computational Linguistics.

[121] Vlad Niculae, Marcos Zampieri, Liviu P Dinu, and Alina Maria Ciobanu. Temporal text ranking and automatic dating of texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 17–21, 2014.

[122] Octavian Popescu and Carlo Strapparava. Semeval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 870–878, 2015.

[123] Abhimanu Kumar, Matthew Lease, and Jason Baldridge. Supervised language modeling for temporal resolution of texts. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2069–2072, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063892. URL `http://doi.acm.org/10.1145/2063576.2063892`.

[124] Dimitrios Kotsakos, Theodoros Lappas, Dimitrios Kotzias, Dimitrios Gunopulos, Nattiya Kanhabua, and Kjetil Nørvåg. A burstiness-aware approach for document dating. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1003–1006, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2257-7. doi: 10.1145/2600428.2609495. URL `http://doi.acm.org/10.1145/2600428.2609495`.

[125] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*, 2014.

[126] Kurt D Bollacker, Steve Lawrence, and C Lee Giles. Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the second international conference on Autonomous agents*, pages 116–123, 1998.

[127] C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98, 1998.

[128] Peter N Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *Soda*, volume 93, pages 311–21, 1993.

[129] Jöran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. Research-paper recommender systems: a literature survey. *Int. J. on Digital Libraries*, 17(4):305–338, 2016. doi: 10.1007/s00799-015-0156-0. URL `https://doi.org/10.1007/s00799-015-0156-0`.

[130] Joeran Beel, Stefan Langer, Georgia Kapitsaki, Corinna Breitinger, and Bela Gipp. Exploring the potential of user modeling based on mind maps. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 3–17. Springer, 2015.

[131] Joeran Beel. Towards effective research-paper recommender systems and user modeling based on mind maps. *arXiv preprint arXiv:1703.09109*, 2017.

[132] Felice Ferrara, Nirmala Pudota, and Carlo Tasso. A keyphrase-based paper recommender system. In *Italian research conference on digital libraries*, pages 14–25. Springer, 2011.

[133] Steven Bethard and Dan Jurafsky. Who should i cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 609–618, 2010.

[134] Yichen Jiang, Aixia Jia, Yansong Feng, and Dongyan Zhao. Recommending academic papers via users' reading purposes. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 241–244, 2012.

[135] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186, 1994.

[136] Sean M McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K Lam, Al Mamunur Rashid, Joseph A Konstan, and John Riedl. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 116–125, 2002.

[137] David M Pennock, Eric J Horvitz, Steve Lawrence, and C Lee Giles. Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach. *arXiv preprint arXiv:1301.3885*, 2013.

[138] André Vellino. A comparison between usage-based and citation-based methods for recommending scholarly research articles. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–2, 2010.

[139] Herbert Van De Sompel and Johan Bollen. An architecture for the aggregation and analysis of scholarly usage data. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'06)*, pages 298–307. IEEE, 2006.

[140] Stefan Pohl, Filip Radlinski, and Thorsten Joachims. Recommending related papers based on digital library access records. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 417–418, 2007.

[141] Marcos Baez, Daniil Mirylenka, and Cristhian Parra. Understanding and supporting search for scholarly knowledge. *Proceeding of the 7th European Computer Science Summit*, pages 1–8, 2011.

[142] Onur Küçüktunç, Erik Saule, Kamer Kaya, and Ümit V Çatalyürek. Recommendation on academic networks using direction aware citation analysis. *arXiv preprint arXiv:1205.1143*, 2012.

[143] Yicong Liang, Qing Li, and Tieyun Qian. Finding relevant papers based on citation relations. In *International conference on web-age information management*, pages 403–414. Springer, 2011.

[144] Andrew Arnold and William W Cohen. Information extraction as link prediction: Using curated citation networks to improve gene detection. In *International Conference on Wireless Algorithms, Systems, and Applications*, pages 541–550. Springer, 2009.

[145] Ni Lao and William W Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81(1):53–67, 2010.

[146] Ding Zhou, Shenghuo Zhu, Kai Yu, Xiaodan Song, Belle L Tseng, Hongyuan Zha, and C Lee Giles. Learning multiple graphs for document recommendations. In *Proceedings of the 17th international conference on World Wide Web*, pages 141–150, 2008.

[147] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430, 2010.

[148] Sean M McNee, Roberto Torres, Joseph A Konstan, Mara Abel, and John Riedl. Enhancing digital libraries with techlens. 2004.

[149] Anshul Kanakia, Zhihong Shen, Darrin Eide, and Kuansan Wang. A scalable hybrid research paper recommender system for microsoft academic. In *The World Wide Web Conference*, pages 2893–2899, 2019.

[150] Benard Magara Maake, ZUVA Tranos, et al. A serendipitous research paper recommender system. *International Journal of Business and Management Studies*, 11(1):39–53, 2019.

[151] Peter D Turney. Learning to extract keyphrases from text. national research council. *Institute for Information Technology, technical report ERB-1057*, 1999.

[152] Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, and Craig G Nevill-Manning. Kea: Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*, pages 129–152. IGI global, 2005.

[153] Kuo Zhang, Hui Xu, Jie Tang, and Juanzi Li. Keyword extraction using support vector machine. In *international conference on web-age information management*, pages 85–96. Springer, 2006.

[154] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[155] Xiaojun Wan and Jianguo Xiao. Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, volume 8, pages 855–860, 2008.

[156] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1: 1–20, 2010.

[157] Florian Boudin. Unsupervised keyphrase extraction with multipartite graphs. *arXiv preprint arXiv:1803.08721*, 2018.

[158] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. A text feature based automatic keyword extraction method for single documents. In *European Conference on Information Retrieval*, pages 684–691. Springer, 2018.

[159] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

[160] Bastien Latard, Jonathan Weber, Germain Forestier, and Michel Hassenforder. Using semantic relations between keywords to categorize articles from scientific literature. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 260–264. IEEE, 2017.

[161] Pavel Savov, Adam Jatowt, and Radoslaw Nielek. Towards understanding the evolution of the www conference. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 835–836, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4914-7. doi: 10.1145/3041021.3054252. URL `https://doi.org/10.1145/3041021.3054252`.

[162] Weidong Zhao, Ran Wu, and Haitao Liu. Paper recommendation based on the knowledge gap between a researcher's background knowledge and research target. *Information Processing & Management*, 52(5):976–988, September 2016. ISSN 0306-4573. doi: 10.1016/j.ipm.2016.04.004. URL `https://doi.org/10.1016/j.ipm.2016.04.004`.

[163] Aravind Sesagiri Raamkumar, Schubert Foo, and Natalie Pang. Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems. *Information Processing & Management*, 53(3):577–594, 2017.

[164] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[165] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[166] Loni Hagen. Content analysis of e-petitions with topic modeling: How to train and evaluate lda models? *Information Processing & Management*, 54(6):1292–1307, 2018.

[167] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[168] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5 (Aug):975–1005, 2004.

[169] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10 (3):61–74, 1999.

[170] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics.

[171] Derek de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, 27 (5):292–306, 1976.

[172] Pavel Savov, Adam Jatowt, and Radoslaw Nielek. Identifying breakthrough scientific papers. *Information Processing & Management*, 57(2):102168, 2020.

[173] Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. In *Advances in neural information processing systems*, pages 865–872, 2007.

[174] Paul Martin, Antoine Doucet, and Frédéric Jurie. Dating color images with ordinal classification. In *Proceedings of International Conference on Multimedia Retrieval*, pages 447–450, 2014.

[175] Titipat Achakulvisut, Chandra Bhagavatula, Daniel Acuna, and Konrad Kording. Claim extraction in biomedical publications using deep discourse model and transfer learning. *arXiv preprint arXiv:1907.00962*, 2019.

[176] Pavel Savov, Adam Jatowt, and Radoslaw Nielek. Innovativeness analysis of scholarly publications by age prediction using ordinal regression. In *ICCS*, pages 646–660. Springer, 2020.

[177] Sandeep Soni, Kristina Lerman, and Jacob Eisenstein. Follow the leader: Documents on the leading edge of semantic change get more citations. *JASIST*, 2020.

[178] Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. Dating documents using graph convolution networks. In *Proceedings of ACL*, pages 1605–1615, 2018.

[179] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

[180] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*, pages 38–45, 2020.

[181] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: Pretrained language model for scientific text. In *EMNLP*, 2019.

[182] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[183] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Scoring, term weighting and the vector space model. *Introduction to information retrieval*, 100: 2–4, 2008.

[184] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*, 2016.

[185] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

[186] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.