Knowledge-Based Systems 220 (2021) 106907

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Automatic trend detection: Time-biased document clustering

Sahar Behpour [a,*], Mohammadmahdi Mohammadi [b], Mark V. Albert [a], Zinat S. Alam [a], Lingling Wang [c], Ting Xiao [a]

[a] University of North Texas, 3940 N Elm St, Denton, TX 76207, USA
[b] Whip Mobility, Petaling Jaya, Selangor, Malaysia
[c] University of Connecticut, 2100 Hillside Rd, Storrs, CT 06269, USA

ARTICLE INFO

Article history:
Received 8 October 2020
Received in revised form 25 February 2021
Accepted 25 February 2021
Available online 2 March 2021

Keywords:
Text mining
Trend detection
Temporal biased clustering
Machine learning

ABSTRACT

Identifying the trending topics in journals and conferences is valuable for understanding the role of authors, institutions, and funding agencies in the progression of knowledge produced in the field. However, many available clustering methods do not accommodate a desire for temporally clustered results that are typical of trends, in part because time of publication is often neglected as a feature. As a demonstration of how time can be emphasized in trend detection, we use a novel approach of introducing a weighted temporal feature to bias a topic clustering toward articles in a similar time frame; this is performed over a set of finance journal abstracts from 1974 to 2020. Latent Dirichlet Allocation (LDA) is used to parameterize each abstract, followed by dimensionality reduction using Singular Value Decomposition (SVD). We detect trending finance topics that are not identifiable when we use a standard clustering approach with no temporal bias. To identify trending topics, we utilize a metric of the silhouette score divided by the standard deviation of clusters over time. We then isolate topics identified by this metric and validate them using expert judgment. Our clustering strategy using temporal bias can be readily utilized in other fields for discovering the rise and fall of trends.

1. Introduction

Understanding the evolution of research topics is crucial in detecting emerging trends in science. Being able to estimate trends in advance could be useful for identifying trendsetters or preparing for upcoming research trends. For instance, predicting topics inside various research areas enables funding agencies to efficiently allocate resources toward those promising research trends [1]. Moreover, finding and examining trending topics is a way to truly understand the nature of an emerging scientific field [2]. Another application of topic evolution in scientific literature shows how one study influences other studies on related topics. In fact, it helps in understanding the lineage of topics. For example, in the sociology of science, topic evolution analysis can help us understand and objectively evaluate the contribution of a scientist or an article. Topic evolution analysis can also result in developing high-performance information retrieval tools with the objective of citation recommendation for scientific researchers [3]. The importance of trends and trend evolution analysis can simply be seen through the widely used phrase, "the trend is your friend except at the end where it bends" [4]. Therefore, in general, knowledge of the rise and fall of scientific topics and their evolutions can benefit communities and organizations in both academia and industry [5].

There is no universal definition of a trend, though most studies consider trends as relatively sudden changes caused by a single event or monotonic shifts during several time intervals [6]. These two types of trends, sudden and monotonic, are seen in individual members of a population or the population as a whole [7,8]. According to Kontostathis et al. [9], a topic area that grows in interest and usage over time is called an emerging trend. The growth in the number of journal articles related to XML, from 3 in the year 1994 to 374 in the year 1999, is a suitable example of this definition. In this study, we adopt Kontostathis et al. [9] definition of trending topics while using the abstracts of finance journals as the sample for our study. Therefore, defining a trend according to the studies' objectives is the first step toward trend analysis and making a trend detection algorithm.

There are several methods of trend detection. Some of the existing trend detection systems are semi-automatic, requiring user input to begin processing. Others are entirely automated, taking a corpus as its input and developing a list of topics. A human reviewer validates the topics generated by the automated process and determines which topics can be considered trends. Trend detection in the scientific literature using text mining techniques,

* Corresponding author.
E-mail addresses: sahar.behpour@unt.edu (S. Behpour), mohammadi.m@aut.ac.ir (M. Mohammadi), mark.albert@unt.edu (M.V. Albert), zinat.alam@unt.edu (Z. Alam), lingling.wang@uconn.edu (L. Wang), ting.xiao@unt.edu (T. Xiao).

https://doi.org/10.1016/j.knosys.2021.106907

specifically topic modeling and clustering algorithms, has been a common technique among researchers [10,11]. For example, Latent Dirichlet Allocation (LDA) is a highly effective unsupervised learning methodology and a popular tool for topic modeling [12]. In LDA, each document is represented as a combination of topics, where each topic corresponds to a multinomial distribution over words. Document-topic and topic-word distributions that are learned by LDA provide the most relevant and meaningful topics for documents and the most descriptive words for each topic, respectively.

LDA has also been applied to a corpus of finance literature to group relevant research on a topic, with illustrations of the evolution of some of the topics over time [13]. However, the LDA model used in Aziz et al. [13] is not capable of simultaneously considering the time-series evolution and capturing the topics in an automated way. Further, in most clustering models, observed trends are a byproduct of topic clustering without the influence of time in identifying the cluster.

Although many performance evaluation metrics are offered for both supervised and unsupervised learning algorithms, there is a lack of suitable evaluation metrics that can automatically and effectively evaluate the trending topics. Roy et al. [5] offer two methods – citation traces and web resources – for emerging trend detection; however, their models must be adjusted manually. With the fast growth of digital documents, automated systems that rely on manual efforts to identify emerging trends are not feasible.

Erten et al. [14] use citations and correlations between documents to visualize temporal patterns of datasets for trend detection. Chang and Blei [15] have introduced the *relational topic model*, a hierarchical model of both network citation structure and node attributes, where the attributes of each document are its words. They apply LDA to the citation network to identify trends based on citations. Similarly, He et al. [3] combine LDA and citation networks in order to address the problem of trending topic evolution. Their approach detects topics in independent subsets of a corpus and leverages citations to connect topics in different time frames. As the authors mentioned in their study, their primary purpose is not to detect topics evolving over short time scales but to show the evolution of some topics that are already identified using the citation network. However, a trend may exist outside the citation network, such as the Covid-19 pandemic [16], which has undoubtedly affected the direction of many researchers since the pandemic has started [17].

In this paper, we propose a novel automated temporal trend detection framework. The proposed method is based on a time-biased clustering algorithm. We first use Latent Dirichlet Allocation (LDA) to parameterize each abstract and Singular Value Decomposition (SVD) for dimensionality reduction. We added time as an additional feature to the document vector representation, and adjusted the weight of time relative to other features in the representation. Finally, we create a trend score metric that can capture the most suitable clustering runs and clusters for each model's iteration. We use human experts to validate that our proposed model and evaluation metrics produce suitable trending topics. We also compare K-means clustering with another clustering algorithm — DBSCAN and demonstrate the advantage of our time-biased clustering approach.

In light of the emerging literature discussed above, the main contributions of our biased clustering approach can be summarized in three major points. First, our approach is an automated time-biased clustering model that can extract the trending topics and terms by considering time as an additional element in the model. Second, the model creates a quality score metric to automatically evaluate candidate clusters in each iteration of clustering runs. And last, this method works on journal abstracts and can be readily leveraged to other fields and documents of other forms. This can be a collection of news articles or research papers in the field of study — any document form with publication dates. To further explore the advantages and disadvantages of our framework, we shall provide later in the text a thorough comparison between our time-biased trend detection approach and some of the existing trend detection approaches.

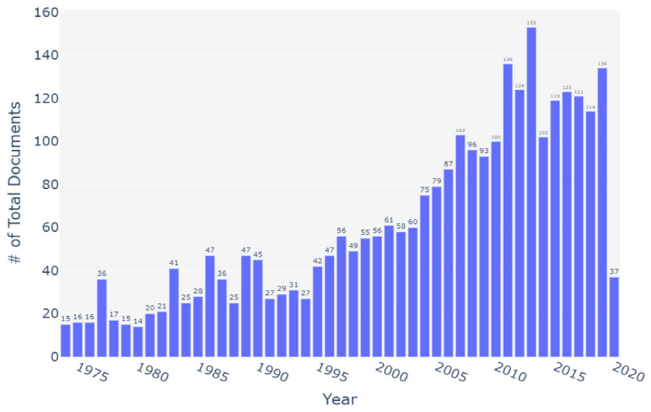


Fig. 1. Article Distribution by Year for the Journal of Finance Economics. The *X*-axis shows the year of publications from 1974 to the first quarter of 2020. *Y*-axis represents the total number of publications in each year.

The remainder of the paper is organized as follows. In Section 2, we discuss our Time-Biased Clustering Framework. In Section 3, we apply our methodology, validate the trending topics in finance Journal articles, and discuss the scope for future research. In Section 4, we discuss the extant literature and the results of experiments comparing time-biased clustering with other clustering approaches. Section 5 concludes.

2. Time-biased clustering approach

2.1. Data collection

We analyze the abstracts of the articles from the *Journal of Financial Economics,* a premier journal in the field of finance. We collect the abstracts of finance articles for a period of 47 years from 1974 (the year of the first issue) until the first quarter of 2020. The abstracts are collected from ScienceDirect, which provides subscription-based access to a large database of academic journals and e-books. ScienceDirect maintains the articles and abstracts in electronic format. It also maintains volumes and issues in chronological order.

We extract the abstracts using Python's Beautiful Soup web scraping library. After the required pre-processing, our data includes a total of 2858 documents published by 1819 unique authors. Fig. 1 shows the distribution of articles by year. As seen, the total number of publications in the year 2013 has reached 153, the highest among all years. From the extracted datasets, "Abstract," "Author," "Year of Publication," and "Affiliation" fields are collected and stored into a Comma-separated Value (CSV) format. Good to mention that no labeling or target class is required for this dataset. We only need the corpus of abstracts, while the time stamp for each abstract is provided by the *Journal of Financial Economics* website.

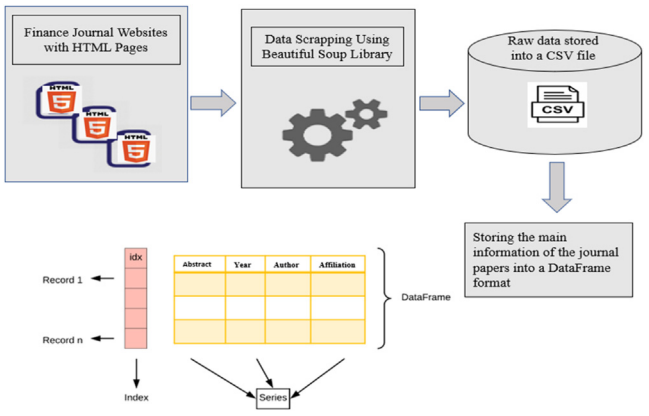


Fig. 2. Data collection and preparation flowchart. In this step, the data is scraped from one of the top finance journals and then stored in the CSV format. The data is then placed in a suitable data structure provided by the Numpy library of Python, DataFrame, while maintaining a unique abstract ID. The ID is later used to retrieve information about abstracts assigned to different clusters.

The following steps are taken toward data collecting and preparation, as shown in Fig. 2:

1. Web scraping to extract information from journal papers.
2. Making a CSV file that contains all the volumes of finance journal abstracts and relevant fields.
3. Storing the extracted information into a PANDAS DataFrame data structure for efficient processing.

2.2. Pre-processing and framework

Preprocessing, which is one of the most critical parts of the automatic trend detection system (Fig. 3), is done by tokenization; lowercasing the tokens; removing punctuation, stopwords, and numbers; and stemming the tokens. This process yields a vocabulary size of over 12,000 distinct words. We apply LDA using Python's Sklearn library to derive the terms that carry the main content of the text. We extract roughly 50 different features (number of topics) that appear in the entire corpus of all the finance journal abstracts. In the next step, we compute vector representations for each abstract using Term Frequency–Inverse Document Frequency (TF–IDF). The result of this effort produces a matrix of 2,858 rows (one row for each document) and 12,000 columns (one column for each distinct word). Thus, we build up a new corpus representation from the extracted terms for further analysis. The vector representations of the abstracts are high dimensional and sparse, thus introduce noise in the clustering approach. We address this issue by using dimensionality reduction techniques, including Singular Value Decomposition (SVD) and Nonnegative Matrix Factorization (NMF), and compare the effectiveness of each. After pre-processing, we use two versions of clustering, ***Standard clustering*** and ***Temporal (Time-Biased) clustering,*** to compare the results. We implement various clustering algorithms, find out the effectiveness of each, and choose the most efficient clustering algorithm among all. In the rest of this section, we will explain the model complexity and detailed information of each experiment used in our study.

In standard clustering, where time is not a feature, a set of articles is represented by a matrix with m rows and n columns, as shown in Eq. (1). We call this SVD (after applying the SVD dimensionality reduction) matrix a ***standard corpus representation***. Each row of the matrix is a representation of a journal abstract. If we use a clustering algorithm in this step, the resulting clusters are driven by the TF–IDF representations without an explicit role of time in clustering. The results of this scenario are presented and discussed in Section 3 of the paper.

$$A_{(m\times n)} = A_{(2858\times 10)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix} \quad (1)$$

In our time-biased clustering, time is added as a feature prior to clustering. In this step, time is introduced and plays a significant role. If we only add the year of publication to the previous SVD matrix, the impact of the year of clustering is highly dependent on the variance of the year feature relative to the other reduced features. To control the relative impact of the year of publications relative to the SVD matrix values, we introduce an automatically adjustable threshold and call it a bias parameter. To normalize the corpus representation, we divide each element of the matrix by the standard deviation of that specific column, including the year column. Thus, the standard deviation for each column is equal to 1. With normalized feature columns, the relative impact of the year is approximately equal to other feature columns with a bias coefficient multiplication scaling the importance of the year in determining the clusters. After that, we generate a candidate list of bias amounts starting with 0 incrementing by 0.01 and ending at 100. Then, we iteratively multiply the normalized year of publication by a bias amount to arrive at a new document matrix for clustering (Eq. (2)). This corpus is now called a ***biased corpus representation***.

$$A_{m\times(n+1)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} & \cdots & a_{1(n+1)} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} & \cdots & a_{2(n+1)} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} & \cdots & a_{i(n+1)} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} & \cdots & a_{m(n+1)} \end{bmatrix} \quad (2)$$

2.3. Trend score metric

After creating the temporal (time-biased corpus) representation, we aim to systematically detect trending topics and evaluate different temporal bias variations. To identify the trending topics for expert analysis, we seek metrics that consider the topics' quality based on the abstract descriptions along with an emphasis on topics which are localized in a short span of years. Variables used to calculate such evaluation metrics are silhouette scores and standard deviations of years for each cluster. The silhouette score displays a measure of how close each point inside a cluster is to the points in the neighboring clusters [18]. This score has a range of [−1, 1], where a high silhouette score indicates that a sample suitably matches the neighboring terms within a cluster. In contrast, a low silhouette score, near -1, indicates that a sample is assigned to a wrong cluster. The standard deviation shows how articles are distributed among different years — the lower the standard deviation, the more localized a cluster's structure and the trending topics.

We create a metric to detect and evaluate suitable thresholds that should be applied to the system to find suitable time biased clustering runs with trending topics. We use the average *trend score,* which is the silhouette score of an individual cluster divided by the standard deviation of the year for a specific timeline corresponding to that cluster. An average over the *trend scores* provides

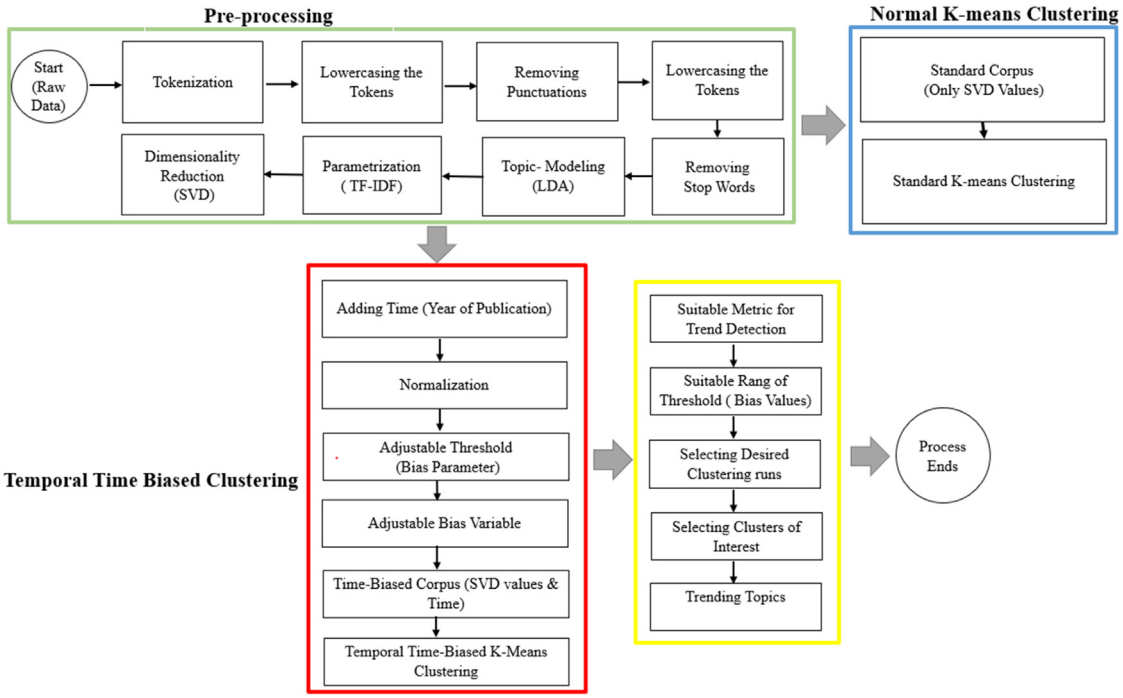


Fig. 3. Schematic diagram of the trend detection framework. At the pre-processing stage, a series of natural language pre-processing tasks are followed by topic modeling and dimensionality reduction techniques to produce reduced vector representation for each abstract. Next, two scenarios — Normal Clustering and Temporal (Time-Biased) Clustering, are considered. In Temporal (Time-Biased) Clustering, we add the year of publication to the resulting corpus, introduce an adjustable bias, and perform clustering on this biased representation.

an approximate range of time-biased amounts suitable for different clustering runs. This range of bias amount that applies to distinct clustering runs results in clusters with high-quality terms and localized in time. According to the average trend score presented in Fig. 4, for bias amounts well beyond 10, the clusters are shaped based on time, not based on the information in the corpus representation. We identify low, moderate, and high bias amounts of 0.57, 3.4, and 9.8, respectively, as candidates for trend detection. These three biases are recognized as local picks in the average trend score illustration. Also, clustering runs with bias amounts between 0–6 (low and moderate bias amounts) contain more trending topics than clustering runs with a bias between 6–10.

Next, we create a summary metric to identify individual clusters that are highly interpretable according to TF–IDF reduced representations as well as localized in time according to the year's standard deviation for documents in that cluster. We observe that 0.4, the maximum average trend score, is a suitable cutoff to identify individual clusters based on trend scores. In the next section, we explain how we choose some clusters of interests and validate some trending topics using this summary metric.

2.4. Experimental setup and model complexity

This study evaluates and validates an automated trend detection system. For data collection, we use Python to scrape the data from the HTML files of the *Journal of Financial Economics*. Title, abstract, year of publication, affiliation, authors are all stored in a CSV format. In the data pre-processing, the LDA model alpha parameter is set to 0.1. This parameter determines both the distribution and concentration of the Dirichlet. According to parameter tuning, if we set the number of topics to 50 and the number of documents to 10, the LDA has a high performance.

After applying different dimensionality reduction techniques, SVD shows a higher performance (according to the average silhouette score of the clustering models) compared to the Principle Component Analysis (PCA) and Non-Negative-Matrix (NMF).

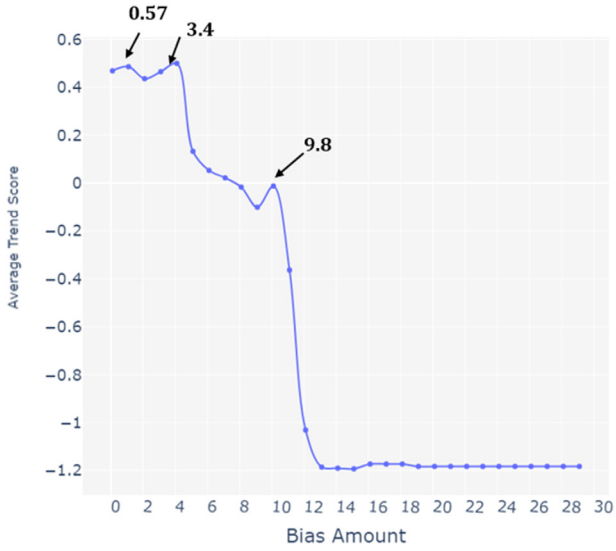


Fig. 4. An evaluation metric to detect the candidate time-biased clustering runs. The graph shows the average trend score metric to detect and evaluate the most suitable time-biased clustering runs. The X-axis shows bias amounts applied to different individual clustering runs. The Y-axis shows the average trend score, silhouette score divided by the standard deviation of years of publication, for every run of K-means clustering. Three time-biased clustering runs (0.57, 3.4, and 9.8), corresponding to low, moderate to high bias amounts, are being detected and highlighted as the candidate clustering runs that contain trending topics.

Table 1 summarizes the average silhouette score of each run of K-means clustering while applying each of the dimensionality reduction techniques. We test our experiment with different numbers of clusters (K). By comparing the average silhouette score of each clustering run, we find that the average silhouette

Table 1
Comparison among the K-means clustering performance using the silhouette score after applying Singular Value Decomposition (SVD), Non-Negative-Matrix (NMF), and Principle Component Analysis (PCA).

	SVD	NMF	PCA
Average Silhouette Score (Standard K-means Clustering)	0.86	0.76	0.85
Silhouette Score (Time-biased K-means clustering)	∼0.6–0.4	∼0.5–0.3	∼0.6-0.4

score of the K-means algorithm does not change significantly with K = 10 or K = 15 if we use a small number for alpha.

3. Historical trends in the field of finance

In this section, we relate our trend detection results with what is known in the field of finance.

3.1. Standard clustering approach

As discussed in Section 2, we first apply topic modeling followed by dimensionality reduction and standard clustering on the corpus. Table 2 shows the terms associated with the top 10 out of 15 topic clusters. These terms are identified by the TF–IDF scores; higher TF–IDF corresponds to higher association among terms. We can measure the quality of these clusters using the silhouette score as well as manually checking their interpretability according to our domain knowledge. The average silhouette score is high for all clusters, with the minimum score being 0.50.

Note that some generic finance terms such as *relate*, *firm*, *firms*, *model*, *price*, and *financial* are relatively frequent in the corpus and cannot distinguish the topics. However, there are some distinct terms, e.g., *hedge*, *fund*, *CEO*, *board*, *bank*, *predict*, *announce*, *bidder*, *target*, etc., that can identify specific finance research topics. For example, in Cluster 2, *CEO*, *board*, and *compensation* together can identify corporate governance-related research; *bank*, *debt*, *loan*, *credit*, and *institution* refer to research related to financial intermediaries; and terms such as *merger* and *target* are sufficient to reliably distinguish the merger and acquisition literature. The limitation of standard clustering is that it tends to combine more than one area of finance research in the same cluster. We can easily identify at least four different streams of finance research, (a) corporate governance, (b) financial intermediaries, (c) merger and acquisitions, and (d) investments within Cluster 2. In other words, a standard clustering is limited in identifying timely research trends in the area of finance. Table 2 also reveals that the number of abstracts is not evenly distributed among the clusters; Clusters 3 and 5 include 1316 and 716 abstracts, respectively. Each of the bottom two clusters (not reported) based on the silhouette score has only one abstract. Though the average silhouette score is relatively high for the top-two topic clusters, this is mainly due to a higher number of papers in those clusters.

Fig. 5 presents one run of the K-means (K = 15) standard clustering without any time bias. Each cluster is shown in a different color. In Fig. 5, Clusters 3 and 5 are distinguishable from the rest simply because these two clusters have relatively higher numbers of papers compared to other clusters. In this type of clustering, clusters are distributed over time, and there is no automatic way to capture a trending topic through the observed graphs.

3.2. Temporal (time-biased) clustering approach

The main goal of developing a biased clustering model is to identify the micro-level research trends over a period of time. Kuhn [19] first developed a model to recognize a paradigm shift in scientific research by analyzing the shifts in vocabulary and framing. However, we are interested in the change of research focus in finance over time, followed by economic shocks, new

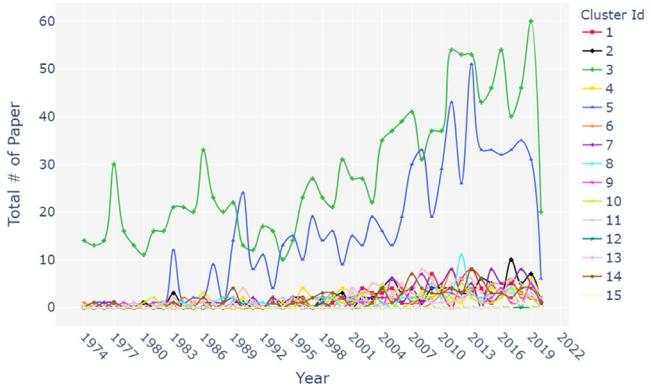


Fig. 5. Standard K-means clustering. Without temporal bias, 15 distinct clusters distinguished by their ID numbers, colors, terms, and structure are produced. The *X*-axis shows the years the papers are published. The *Y*-axis indicates the number of published papers categorized in each cluster. There are only two clusters (5 and 12) with little or no overlap with other clusters. The rest of the clusters cover the same structure in their graphs, and so we assume they carry similar terms. All the clusters are spread over the journal publication years from 1974 to 2020. The correlations between graphs and the terms are high and thus make the trend detection task inefficient and unfair. The start and the end date of a specific topic that a cluster represents is also unclear. Most of the graphs, at a minimum, represent 5 picks during different time intervals. In most of the graphs, the pick is highly overlapping, e.g., the overlap of the picks in clusters 4 and 14 during time intervals 1984–1986 and 1992–1995.

regulations, change in business model and technology, availability of new sources of data, etc. Therefore, we introduce temporally biased clustering to identify distinct trends in published articles. First, we add a low threshold of time bias of 0.57 in Table 3 and lay out the terms associated with the top 10 among 15 topics clusters with a trend score of at least 0.40.

In comparison with the standard clustering in Table 2, we notice a few differences. First, there are fewer key terms extracted for each cluster. Second, the number of articles is more evenly distributed among clusters. Third, the average silhouette scores of the clusters are lower but reasonably high to have a meaning. Fourth and most importantly, with the introduction of bias, the topics of the journal articles are better grouped, and thus we are able to recognize some trends in finance research as identified in Table 3.

To observe the impact of higher temporal bias, we introduce a moderate amount of bias (=3.9) to the clustering process and present the results in Table 4. We then present results based on a high level of bias (=9.8) in Table 5. We observe that as the bias level increases, the number of well-defined clusters declines. Specifically, the number of topics with the silhouette score above 0.4 declines from 10 in Tables 3–4, and further to 2 in Table 5.

Next, we create visualizations of the cluster structures to observe the trends. Fig. 6 (a–c) shows three different types of cluster distributions over time by three different time biased thresholds. We detect three time-biased clustering runs with low to high bias amounts (0.57, 3.9, 9.8) that have desirable average trend scores. To choose clusters of interest for trend detection and trending topic generation, we use the average trend score of equal or greater than 0.4 as the standard cutoff. For a bias amount of 0.57,

Table 2
Terms associated with Top 10 topic clusters with standard-means (K = 15) clustering.

Cluster ID	3	5	2	6	7	9	13	14	11	1
	future time liquid bond effect test premium predict estimate paper rate portfolio returns find relate inform fund expect investor volatility option use asset trade risk market stock model price	perform examine public loan govern use higher result value relate evidence director target firms manage control debt ownership financial effect increase board shareholder corporate CEO bank	loan stock higher industries effect cash CEO board compensation firms offer credit shareholder institution market financial product ownership find capital invest public manage target increase equities merger bank debt	new increase return effect capital corporate time cost investor tax model execute finance consist fund risk relate result use find market invest price manage financial debt option stock firm	change target option paper model contract tax incentive offer show reduce firms significantstock market increase inform cost equities capital return announce financial manager fund effect risk loan bank firm	predict bank spread return associate security share default cost hedge effect use capital liquid inform relate debt find offer investor manage equities invest market financial model risk price stock firm	leverage trade capital find bank investor test effect paper asset relate use stock corporate analyst tax offer model price invest fund market issue manage risk debt rate inform firm	interest share return develop model investor relate loan use takeover financial borrow effect ownership fund value increase higher inform find sale credit stock invest rate cost bank capital market firm	trade board competition capital firms effect target model public relate inform venture result liquid tax stock increase find use bidder cost invest corporate CEO credit manage debt market bank	invest cost trade high relate predict asset cashflow use effect return offer capital cash find option hedge investor increase financial fund model inform liquid stock credit risk market price firm
# of abstracts	1312	716	98	69	100	56	51	87	102	71
Silhouette Score	0.99	0.99	0.60	0.57	0.56	0.56	0.54	0.54	0.53	0.52

Table 3
Topic clusters with low temporal bias (bias = 0.57) and a trend score greater than 0.4.

Cluster ID	10	8	11	1	3	6	7	4	9	5
	result abnormal relate negate cash high effect positive evidence earn market increase significant price find firm repurchase stock announce return	change activity trade market effect positive find earn relate funds hedge stock invest investor return perform mutual firm manage	announce return increase acquisition firms invest market find gain bank stock effect acquire tax cash merger value capital target firm	effect estimate find returns structure use term asset rate volatility dividend predict expect risk market inform price stock model	firms liquid inform credit manage effect model increase find industries equities invest financial cost leverage cash capital finance debt	investor impact trader use examine abnormal insider provide earn evidence large find return firm inform liquid market price stock trade	market investor portfolio forward call future paper valuation value process exercise term interest formula bond derivative model rate price option	turnover ownership effect perform incentive increase relate outsider CEO corporate pay find chief officer director compensation board executive firm	condition explain find predict liquid cross section price portfolio beta asset relate model returns volatility market expect stock premium risk	perform model data factor paper empirical statistic capital jump portfolio risk CAPM return market use estimate asset price test
Topic Trend – Main Areas	Asset Pricing	Investment	Corporate Finance	Asset Pricing	Corporate Finance	Asset Pricing	Asset Pricing	Corporate Governance	Asset Pricing	Asset Pricing
Topic Trend – Subareas	Information & Market Efficiency	Fund Management	Mergers & Acquisition	Factor Modeling	Capital Structure	Market Efficiency	Financial Markets & Derivatives	Executive Compensation	Factor Modeling	Pricing Models
Trend Timeline	1975, 1990–1992, 1996, 2001, 2010–2014, 2019	1976, 1980, 1983, 1985–1988, 2006–2008, 2010, 2013–2014, 2017	1977, 1987, 1993–1996, 1999, 2006–2011, 2013–2016	1987, 1990, 1995–1998, 2001, 2007–2010, 2013, 2016–2017	1990, 1992, 1996, 2002, 2006–2012, 2015, 2019	1994–1997 2001, 2004, 2014–2016	1981–1982, 1990–1996, 2002–2003, 2005, 2008–2009, 2012, 2014, 2015	1983, 1985–1988, 2003–2004, 2007, 2011, 2014, 2017	1978, 1987–1988, 1996–2000, 2004	1975–1981
# of Abstracts	142	272	137	249	246	191	210	179	229	174
Silhouette Score	0.54	0.51	0.51	0.50	0.47	0.47	0.47	0.46	0.43	0.40
Avg. Yearly Std. Dev of Bias	0.96	0.88	0.91	0.97	0.77	1.02	1.21	0.69	0.95	1.19
Trend score	0.56	0.58	0.56	0.52	0.62	0.46	0.39	0.67	0.45	0.35

we can visually observe 10 clusters out of 15 that are potential candidate clusters carrying trending topics. The clusters show some picks in a specific time interval, though they are spread over a longer time interval. For example, clusters with ID number 8 include journal papers that were published between 1979 and 2019, but they still have different trend timelines (1976, 1980, 1983, 1985–1988, 2006–2008, 2010, 2013–2014, 2017). These trend timelines in the cluster's structure represent the effect of the time-bias amount. Once we increase the bias amount to 3.4 in Fig. 6b, the effect of this threshold on the system is more visible. Now the clusters have fewer overlaps compared to the clusters with a bias amount of 0.57 (Fig. 6a). Fig. 6c provides a visualization with a higher bias amount, which primarily groups the clusters by the year of publication.

The findings of the temporal clustering approach are summarized as follows. First, the topic trends become more localized in time as the bias increases. This is evident by the increasing trend scores and the decreasing number of years present in the trend timeline from Tables 3 to 5. By increasing the threshold amount of temporal bias, the topic clusters become more localized in time (Fig. 6). For high temporal bias amounts, each cluster has a clear beginning, peak, and end. As the bias amount increases, the effect is localized over a narrower period of time. For example, the highest temporal bias amount of 9.8 localizes the topic trends to roughly four to five years (Fig. 6c).

Second, the well-defined topic trends are stable across different bias levels. For example, Cluster 4: Corporate Governance and Executive Compensation in Table 3 corresponds to Cluster 12 in Table 4 and Cluster 15 in Table 5, respectively. Similarly, Cluster 8: Investments and Fund Management in Table 3 corresponds to Cluster 4 in Table 4 and Cluster 7 in Table 5, respectively. Thus, the temporal bias helps to refine the periods in which there are greater interests in these topics. In the next section, we apply a trend validation strategy by studying the clusters of interests with bias amounts, 0.57, 3.9, and 9.8, discussed in this section.

3.3. Trend validation

We validate the trend detection of our bias clustering model in two ways. First, we survey academic literature in finance to check if the trends discussed in the literature are consistent with the trends identified in our model. Second, we check if there is a significant change in the economic, business, and regulatory environment prior to the trend identified in our model. Academic research by financial economists often follows these changes.

Among all the clusters presented in Tables 3–5 and Fig. 6(a–c), only two clusters of each clustering run – *Corporate Governance/Executive Compensation* and *Fund Management* - are reported in Fig. 7(a–c) to validate the trends. For our purposes, validating a few clusters is sufficient to confirm that our time-biased clustering system is capable of detecting the trending topics. The X-axis of Fig. 7(a–c) plots the years of publication. The left Y-axis represents the number of papers published in each cluster in each year (bar chart). The right Y-axis shows the percentages of the papers published in each cluster in each year (line). We introduce a low amount of bias (0.57) in Fig. 7a and record the publishing trends in Executive Compensation (Cluster 4) and Fund Management (Cluster 8). Next, we introduce a medium amount of bias (3.9) and capture the trends in Executive Compensation (Cluster 12) and Fund Management (Cluster 4) in Fig. 7b. Finally, we introduce a high amount of bias (9.8) and record the publishing trends in Executive Compensation (Cluster 15) and Fund Management (Cluster 7) in Fig. 7c.

Next, we compare if observed publishing trends in Fig. 7(a–c) coincide with the changes observed in the economic, business, and regulatory environment. There is a heightened interest among finance researchers in corporate governance and executive compensation after the financial crisis of 2008 and the subsequent enactment of the Dodd-Frank Wall Street Reform and Consumer Protection Act in 2010. In Tables 4 and 5, our model shows that the research trend in corporate governance and executive compensation peaks between 2013–2018. Fig. 7(b–c) further demonstrates this trend. This 3- to 5- year gap is notable as it can take many years for interest in a field to build and the publications in finance journals to follow; our identified trend period matches the timing of the shifts in economic conditions and regulation landscape.

Then, we evaluate the accuracy of the trend period we identified for Fund Management. A survey of the academic literature suggests that while there are only 16 papers on hedge funds published in the premier finance journals prior to 2005, these journals have published 105 papers on hedge funds since 2005 [20]. The earlier research in this area, termed as Investments — Fund Management, has focused on the *performance* and comparison of *hedge* funds with *mutual* funds and the factors that drive hedge fund *returns*. Another strand of research looks at the operational and *liquidity* risks of hedge funds. More recent studies focus on managerial *skill*, fund *management*, and the impact of hedge fund *activity* on *market* stability and *stock return*. Table 3, 4, and 5 list all such terms, including *hedge*, *mutual*, *fund*, *performance*, *stock*, *return*, *liquidity*, *skill*, *manage*, in Clusters 8, 4, and 7, respectively.

Though Alfred Winslow Jones established the first hedge fund in 1949, the real revolution of the hedge fund industry started between 2000 and 2002 when the dot.com bubble burst. During that period, hedge funds consistently generate better-than-market returns, which results in a heavy inflow of capital from institutional investors. Hedge Fund Research (HFR) estimates that the total assets under management of the hedge fund industry increased from $39 billion in 1990 to more than $2.97 trillion by 2015. The research trend for Fund Management identified in our model covers roughly the period from 2005–2016 for the temporal bias amount of 0.57 and 3.9 in Fig. 7a and Fig. 7b, respectively. The research trend for Fund Management peaks at 2009 and lasts till 2015 for the temporal bias amount of 9.8. These trend periods are consistent with the timeline of the rise of the hedge fund industry and the survey of academic literature.

4. Related literature and comparison of clustering techniques

4.1. Related works

A number of temporal trending topic detection and evolution models are available in the literature. Erten et al. [14] offer a temporal graph visualization of the categorization of articles in the Association for Computing Machinery (ACM) conference proceedings using an algorithm for visualization and combined graph, TGRIP, to find the most frequently used topics in computing. Their novel visualization method can identify steadily declining areas and rapidly growing ones. Their category graphs, which take advantage of the ACM categorization of articles, consisted of multiple time slices. To identify the trends in each year, the top five frequently used title words are chosen. Words such as "design", "system", and "simulation" are persistently found in the list over time, whereas "ada", "database", and "parallel" are listed only for a few years before they disappeared. Since this study uses the ACM portal's manually labeled dataset, the visualization cannot be easily replicated on a different unlabeled dataset.

Also, some models are not easy to replicate due to their required manual corrections or reliance on additional data structure. Roy et al. [5] offers two methods for trend detection — citation traces and web resources to identify emerging trends.

Table 4
Topic clusters with medium temporal bias (bias = 3.9) and a trend score greater than 0.4.

Cluster ID	13	12	4	8	3	9	1
	share negate relate recommend significant institution announce high positive effect earn short price market increase find firm repurchase stock return	evidence govern effect option CEO Increase Incentive manager relate pay corporate find director chief officer compensation board executive firm	tax earn effect use positive activity relate find trade stock return invest perform investor funds firm hedge mutual manage	expect trade investor credit asset term find returns use structure volatility rate predict risk market stock inform price model	corporate firms liquid effect increase model find equities industries leverage financial cost invest capital cash finance debt	estimate use deal return acquisition invest bank firms tax market effect find stock merger acquire value cash capital target firm	return analyst use trading impact order earn evidence investor trader provider firm insider find inform, price market stock liquid trade
Topic Trend – Main Areas	Asset Pricing	Corporate Governance	Investments	Asset Pricing	Corporate Finance	Corporate Finance	Asset Pricing
Topic Trend – Subareas	Information & Market Efficiency	Executive Compensation	Fund Management	Factor Modeling	Capital Structure	Mergers & Acquisition	Information & Market Efficiency
Trend Timeline	1999–2000, 2002–2003, 2005–2006, 2011, 2013–2016	2006–2008, 2010–2011, 2014–2016, 2017–2018	2002–2005, 2007–2009, 2010–2018	2003–2017, 2013, 2011–2019	1995–2000, 2002–2004, 2007–2009	1994–1999, 2001–2004, 2006–2010	1998–2001,2004–2006, 2009–2012
# of Abstracts	96	183	184	181	190	107	119
Silhouette Score	0.58	0.58	0.51	0.51	0.50	0.49	0.46
Avg. Yearly Std. Dev of Bias	0.45	0.37	0.38	0.39	0.41	0.48	0.46
Trend Score	1.27	1.44	1.35	1.32	1.21	1.02	1.00

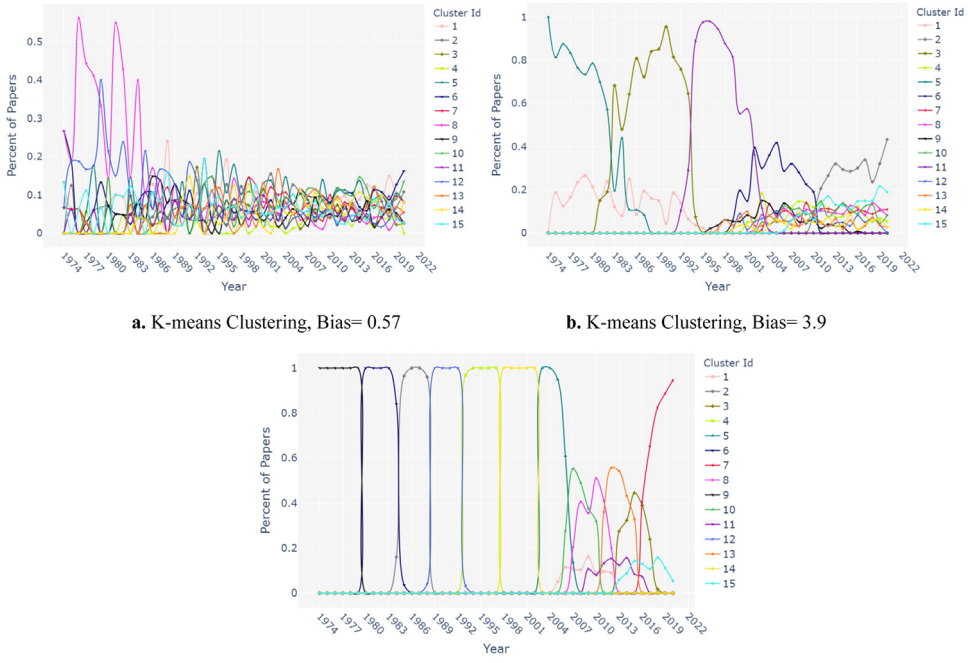
Table 5
Topic clusters with high temporal bias (bias = 9.8) and a Trend score greater than 0.4.

Cluster ID	15	7
Terms	Value higher incentive management financial option pay director find relate increase corporate CEO board chief officer executive compensation firm	Institution information market relate activity use liquid find skill funds return stock invest investor firm perform manage mutual hedge fund
Topic Trend – Main Areas	Corporate Governance	Investments
Topic Trend – Subareas	Executive Compensation	Fund Management
Trend Timeline	2013–2018	2009–2015
# of Abstracts	97	114
Silhouette Score	0.57	0.45
Avg. Yearly Std. Dev of Bias	0.18	0.17
Trend Score	3.16	2.58

Although the two methodologies are interesting with good performance, the authors emphasize that their models are only manually performed. Manual efforts by human experts to identify emerging trends are no longer feasible as the number of digital documents has dramatically increased over the past decade [9].

Detecting trending topics from high-dimensional textual datasets is a challenging task due to considerable ambiguities and redundancies in the existing text [21]. The quality of topics produced by a text mining model can be poor, given the noise in high-dimensional vector representations of abstracts [22]. In response to this limitation, Boushaki et al. [23] propose a dynamic and incremental method based on the Latent Semantic Indexing (LSI) and Cuckoo Search (CS) optimization. The authors show that their model can improve clustering with more precision and less computational space by reducing the dimensionality of the data. Also, Deng et al. [24] present a comprehensive survey of existing Soft Space Clustering (SSC) algorithms and their recent developments as promising technology involving clusters that are identified based on their association with subspaces in high-dimensional spaces.

Similarly, a topic modeling implementation prior to applying clustering techniques can decrease the noise inherent in document vector representations [25]. Topic modeling can project documents into a dimensionality-reduced topic space that could facilitate effective document clustering. Integrating these two tasks, topic modeling and clustering, into a unified framework is reflected in the Multi-Grain Clustering Model (MGCTM) by



a. K-means Clustering, Bias= 0.57

b. K-means Clustering, Bias= 3.9

c. K-means Clustering, Bias = 9.8

Fig. 6. Cluster distributions over time as the amount of temporal bias increases. The *X*-axis shows the year of publications. The *Y*-axis indicates the percentage of published papers categorized in each cluster in a given year. 6a. Low bias = 0.57 creates several topics spread over different time intervals. 6b. Medium bias = 3.9 creates several topics that are grouped by a set of years. 6c. High bias = 9.8 demonstrates clusters that are primarily driven by the year of publication.

Xie and Xing [26]. We show that using LDA followed by a TF–IDF parameterization and then Singular Value Decomposition can effectively and efficiently reduce the dimensionality of the original corpus representation from 2858 by 12,000 to 2558 by 3. According to the quality metric, trend score, that we have created, our results support the usefulness of LDA prior to clustering as proposed by Lu et al. [25].

Many of the current trend detection systems are domain specific. Consequently, one may not be able to use the same model to capture emerging trends for a new context. Bolelli et al. [27] introduce a Segmented Author-Topic model by adding the temporal ordering in the documents. Morinaga and Yamanishi [28] employ a probabilistic Finite Mixture Model to represent the structure of topics and analyze the changes in time of the extracted components to track emerging topics. However, their evaluation rests on an email corpus; thus, it is not clear how it would perform on a scientific corpus. One of the advantages of our biased clustering strategy is its applicability to other fields. Other journals or research areas can be similarly analyzed for trending topics.

When we use the standard clustering approach without a temporal bias, each cluster spans across multiple research areas in finance and, therefore, finding the trending topics through the structure of clusters is not possible. Once we introduce a temporal bias, even a very small amount of bias, to the clustering process, clearer topic trends start to emerge. The sharpness of the topic trends improves as we increase the magnitude of the bias. Although the number of well-defined clusters (topic score > 0.4) declines as the bias increases, the quality of the well-defined clusters remains high even at the very high amount of bias. Overall, our findings highlight the benefits of incorporating time bias into topic clustering and modeling.

One of the advantages of our biased clustering strategy is its applicability to other fields. Other journals or research areas can be similarly analyzed for trending topics. For example, our method can be applied in analyzing the trends in the news or the trends in stock market risk and return. In fact, this automatic trend detection system can be performed on a variety of textual data.

Table 6
A comparison between the maximum average trend scores of K-means and DBSCAN. The K-means clustering has higher performance compared to DBSCAN.

Biased Clustering	K-means	DBSCAN
Average Trend score	0.40	0.25

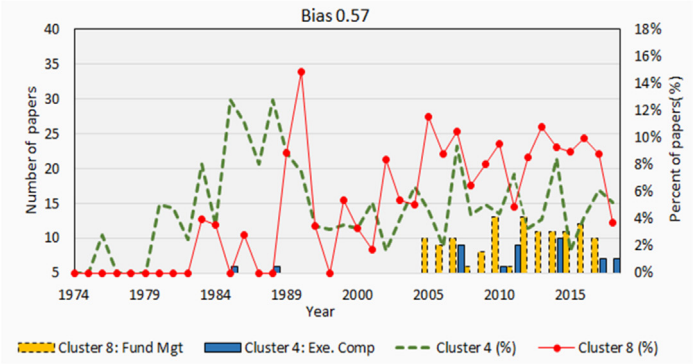
4.2. Time-biased K-means versus DBSCAN

There are a large number of strategies that can perform clustering. K-means is a simple centroid-based approach while Density-Based Spatial Clustering of Application (DBSCAN) is a common representative among density-based clustering algorithms. The K-means clustering algorithm is described in detail by Hartigan [29] and DBSCAN by Januzaj et al. [30]. A thorough review of the DBSCAN method can be found in Khan et al. [31].

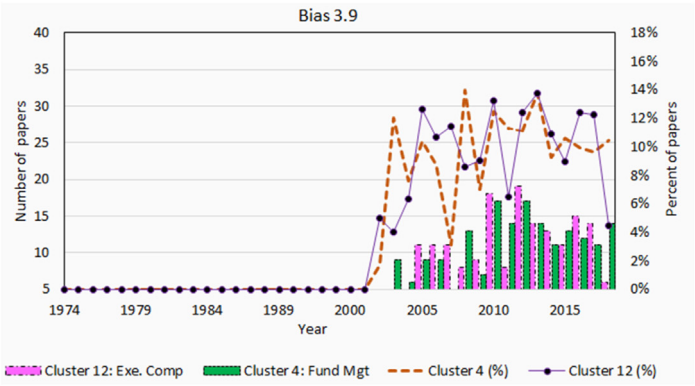
Both models, K-means and DBSCAN, have their own pros and cons; K-means is fast, easy to implement, efficient in terms of computing with linear complexity. However, it requires users' insight and effort in defining the number of clusters. K-means also starts with a random choice of cluster centers, which affects replicability. DBSCAN does not require a pre-selected number of clusters and is also able to find outliers. However, this model's major drawback is that the density estimation parameters often must be tuned to the data set. DBSCAN often does not perform well when clusters have varying densities.

To find a clustering model that works best in fulfilling our objectives, we test and evaluate both K-means clustering and DBSCAN. The results in Table 6 show that K-means clustering can produce a higher average trend score compared to the DBSCAN.

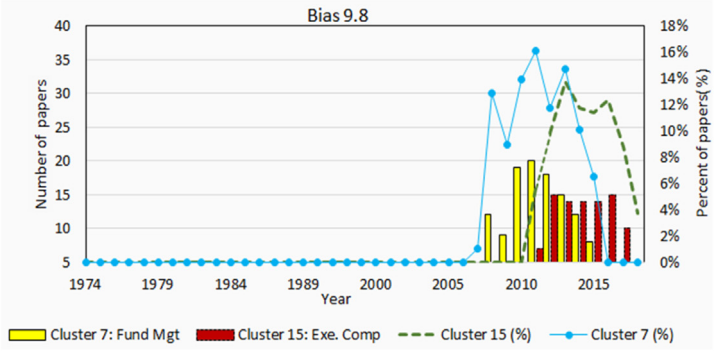
Since K-means clustering is a single view clustering method, it clusters one dimension of the data matrix based on similarities along another dimension to improve the time-biased clustering system. However, we can still improve the performance of our system by using other parameterization methods for abstract representations. For example, Multi-View clustering developed



a. Trends in Executive Compensation (Cluster 4) and Fund Management (Cluster 8), Bias = 0.57



b. Trends in Executive Compensation (Cluster 12) and Fund Management (Cluster 4), Bias = 3.9



c. Trends in Executive Compensation (Cluster 15) and Fund Management (Cluster 7), Bias = 9.8

Fig. 7. Trends corresponding to each time-biased clustering run. The *X*-axis plots the year of publication. The left *Y*-axis represents the number of papers published in each cluster in each year (bar chart). The right *Y*-axis shows the percentages of the papers published in each cluster in each year (line). 7a. Bias = 0.57, trends in Executive Compensation (Cluster 4) and Fund Management (Cluster 8) are captured. 7b. Bias = 3.9, trends in Executive Compensation (Cluster 12) and Fund Management (Cluster 4) are captured. 7c. Bias = 9.8, trends in Executive Compensation (Cluster 15) and Fund Management (Cluster 7) are captured.

with the Cooperation of Visible and Hidden views, i.e., MV-Co-VH [24], a method to deal with unlabeled data, has received a lot of attention in recent years. Another model known as Co-clustering for Co-occurrence Data (MV-ITCC) proposed by Xu et al. [32] claims that the MV-ITCC model can exploit the agreement and disagreement among different views by sharing a common clustering result and provide a correct balance between agreement and disagreement, with a mechanism of maximum entropy. The model is being tested on both images and text datasets and has demonstrated their model's superiority. Although the objective of Xu et al. [32] is not to find trending topics of a textual dataset which has time dependency, it is still worthwhile to consider their model to improve the quality of clusters of our model. In future studies, we plan to test the performance of our model using MV-Co-VH [24] and MV-ITCC [32] clustering methods.

5. Conclusion

In this work, we develop and evaluate the effectiveness of our temporal (time-biased) clustering approach to detect trends in a corpus of journal abstracts. We find that both LDA and Singular Value Decomposition are practical in the parameterization of the

finance journal abstracts. We create a trend score for automatic detection of a trend by utilizing the silhouette score for topic interpretability and the standard deviation of years to quantify localization in time. This approach allows the identification of interpretable trends over a wide range of temporal bias values. Notably, the amount of bias can be varied to suit the needs of a user — higher bias can be used to localize topic clusters to a specific time, while lower bias values can simply be used as an additional factor to improve the interpretability of standard topic modeling.

For our time-biased clustering approach, there is a number of variations possible for some steps that can potentially improve the quality of the identified clusters of interest. The importance of using alternate clustering strategies is discussed in Section 4.2. Similarly, in addition to using the TF–IDF and LDA for document vector parameterization, we can also consider other document parameterization techniques for improving the performance. For instance, advanced word embedding techniques using BERT [33] and later transformer variants may help improve contextual inference. Additionally, we can utilize various interface designs to improve the usability of the system — particularly for domain experts without an understanding of clustering. We are in the process of designing a user interface to allow domain experts to observe the clusters as they are generated and expose the small set of parameters mentioned in this paper for iterative observation of improvement. The range and options of clusters provided on the interface could be streamlined with more feedback from domain experts to enable a more seamless detection of the top 3 to 5 high-quality clusters without the need for parameter choices.

By formally introducing the role of time in topic clustering, we are able to identify historical trends, which ultimately enables a better prediction of the direction of academic research going forward. Finally, a more thorough analysis of trends that are of particular relevance to an individual would be warranted, including visualizations of the trending topics and corresponding information such as the authors and affiliations of the trendsetters for particular topics.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] V. Prabhakaran, W.L. Hamilton, D. McFarland, D. Jurafsky, Predicting the rise and fall of scientific topics from trends in their rhetorical framing, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, 2016, pp. 1170–1180.

[2] G.D. Webster, P.K. Jonason, T.O. Schember, Hot topics and popular papers in evolutionary psychology: Analyses of title words and citation counts in evolution and human behavior, 1979 – 2008, Evolut. Psychol. 7 (3) (2009) 348–362.

[3] X.J. He, Y.F. Hsu, S. Zhu, A.T. Wierzbicki, O. Pontes, C.S. Pikaard, H.L. Liu, C.S. Wang, H. Jin, J.K. Zhu, An effector of RNA-directed DNA methylation in arabidopsis is an ARGONAUTE 4-and RNA-binding protein, Cell 137 (3) (2009) 498–508.

[4] J.D. Schwager, Hedge Fund Market Wizards, John Wiley & Sons, Hoboken, New Jersey, 2012.

[5] S. Roy, D. Gevry, W.M. Pottenger, Methodologies for trend detection in textual data mining, in: Proceedings of the Textmine, vol. 2, 2002, pp. 1-12.

[6] R.M. Hirsch, R.B. Alexander, R.A. Smith, Selection of methods for the detection and estimation of trends in water quality, Water Resour. Res. 27 (5) (1991) 803–813.

[7] N.S. Urquhart, S.G. Paulsen, D.P. Larsen, Monitoring for policy-relevant regional trends over time, Ecol. Appl. 8 (2) (1998) 246–257.

[8] T.L. McDonald, Review of environmental monitoring methods: survey designs, Environ. Monit. Assess. 85 (3) (2003) 277–292.

[9] A. Kontostathis, L.M. Galitsky, W.M. Pottenger, S. Roy, D.J. Phelps, New York, A Survey of Emerging Trend Detection in Textual Data Mining, Springer, 2004, pp. 185–224,

[10] D. Gevry, Detection of emerging trends: Automation of domain expert practices, (Master's Thesis), Department of Computer Science and Engineering, Lehigh University, 2002.

[11] Q. Mei, C. Zhai, Discovering evolutionary theme patterns from text: an exploration of temporal text mining. in: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 2005, pp. 198-207.

[12] A. Estabrooks, N. Japkowicz, A mixture-of-experts framework for text classification, in Proceedings of the 2001 workshop on Computational Natural Language Learning Association, 2001, pp. 1-8.

[13] S. Aziz, M.M. Dowling, H. Hammami, A. Piepenbrink, Machine learning in finance: A topic modeling approach, SSRN Electron. J. (2019) http://dx.doi.org/10.2139/ssrn.3327277.

[14] C. Erten, P.J. Harding, S.G. Kobourov, K. Wampler, G. Yee, Exploring the computing literature using temporal graph visualization, Proc. Visualiz. Data Anal. 5259 (2004) 45–56.

[15] J. Chang, D.M. Blei, Hierarchical relational models for document networks, Ann. Appl. Stat. 4 (1) (2010) 124–150.

[16] M. Ciotti, S. Angeletti, M. Minieri, M. Giovannetti, D. Benvenuto, S. Pascarella, C. Sagnelli, M. Bianchi, S. Bernardini, M. Ciccozzi, COVID-19 outbreak: an overview, Chemotherapy 64 (5–6) (2019) 215–223.

[17] S. Verma, A. Gustafsson, Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach, J. Bus. Res. 118 (2020) 253–261.

[18] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[19] T.S. Kuhn, The Structure of Scientific Revolutions, University of Chicago Press, 1962.

[20] V. Agarwal, K. Mullally, N. Naik, Hedge funds: A survey of the academic literature, Found. Trends Finance 10 (1) (2015) 1–107.

[21] M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques, in: KDD Workshop on Text Mining, 2000.

[22] S. Balbi, Beyond the curse of multidimensionality: High dimensional clustering in text mining, Statist. Appl.-Italian J. Appl. Statist. 22 (1) (2012) 53–63.

[23] S.I. Boushaki, N. Kamel, O. Bendjeghaba, High-dimensional text datasets clustering algorithm based on cuckoo search and latent semantic indexing, J. Inform. Knowl. Manage. 17 (3) (2018) 1–24.

[24] Z. Deng, R. Liu, P. Xu, K. Choi, W. Zhang, X. Tian, T. Zhang, L. Liang, B. Qin, S. Wang, Multi-view clustering with the cooperation of visible and hidden views, IEEE Trans. Knowl. Data Eng. (2020).

[25] Y. Lu, Q. Mei, C. Zhai, Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA, Inf. Retr. 14 (2) (2011) 178–203.

[26] P. Xie, E.P. Xing, Integrating document clustering and topic modeling, in: Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence, 2013, pp. 694–703.

[27] L. Bolelli, Ş. Ertekin, C.L. Giles, Topic and trend detection in text collections using latent dirichlet allocation, in: European Conference on Information Retrieval, Springer, Berlin, Heidelberg, 2009, pp. 776–780.

[28] S. Morinaga, K. Yamanishi, Tracking dynamics of topic trends using a finite mixture model, in: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 811-816.

[29] J.A. Hartigan, Clustering Algorithms, John Wiley & Sons, Hoboken, New Jersey, 1975.

[30] E. Januzaj, H.P. Kriegel, M. Pfeifle, Dbdc: Density based distributed clustering, in: Proceedings of the Advances in Database Technology, vol. 2992, 2004, pp. 88-105.

[31] K. Khan, S.U. Rehman, K. Aziz, S. Fong, S. Sarasvady, DBSCAN: Past, present and future, in: the 5th International Conference on the Applications of Digital Information and Web Technologies (ICADIWT), 2014, pp. 232-238.

[32] P. Xu, Z. Deng, K.-S. Choi, L. Cao, S. Wang, Multi-view information-theoretic co-clustering for co-occurrence data. in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33(01), 2019, 379-386.

[33] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

Sahar Behpour is a Data Science Ph.D. candidate and a Teaching Fellow in the College of Information, Department of Information Science, University of North Texas. Her primary research interests include Machine Learning, Artificial Intelligence, and Data Mining. Her research papers have been accepted for presentation at the American Physical Society, Association for Library and Information Science, International Conference of Knowledge Management, and ACM Richard Tapia Celebration of Diversity in Computing.

Mohammadmahdi Mohammadi is an AI/Data Engineer at Whip Mobility in a Malaysian company. He is a Data Scientist with +5 years of hands-on experience in Data related projects and +9 years of experience in Web development. His research interest is using machine learning to create new data-related solutions for challenging contexts.

Mark V. Albert is an assistant professor of Computer Science and Engineering at the University of North Texas. He is the director of the Biomedical Artificial Intelligence lab which uses machine learning to better understand and inform clinical care. His work has been published in various journals including Proceedings of the National Academy of Sciences, Journal of Neuroscience Methods, Journal of Cognitive Neuroscience, and Journal of Physical Medicine and Rehabilitation.

Zinat Alam is an Assistant Professor in the G. Brint Ryan College of Business, University of North Texas. Her primary research interests include capital structure, corporate governance and machine learning in Finance. Her research papers have been published in the *Journal of Financial and Quantitative Analysis and Management Science* and have been accepted for presentation at the American Finance Association Meetings, Financial Management Association Meetings, Financial Intermediation Research Society Conference, Conference on Financial Economics and Accounting, European Association of Law and Economics Conference, Conference on Empirical Legal Studies, EFMA Asian Finance Symposium, Academic Conference on Corporate Governance, and CRSP Forum.

Lingling Wang is an assistant professor of finance at School of Business, University of Connecticut. Her primary research interests are in international finance and corporate governance. Her work has been published in various finance and accounting journals, including the *Journal of Financial Economics, Review of Financial Studies, Review of Finance, and Contemporary Accounting Research.*

Ting Xiao is a research assistant professor of Computer Science and Engineering at the University of North Texas. She applies statistical and machine (deep) learning tools to extract meaningful information from large data sets including recent projects processing video, audio, and wearable sensors. Her deep learning approaches focus on distilling high-throughput data using a combination of unsupervised (autoencoders) and supervised (transfer learning) dimensionality reduction techniques. To fully utilize these results, she also creates the interfaces to interpret and visualize the resulting information. She specializes in developing models and software used by clinical collaborators with user testing, feedback, and the subsequent impact on human decision-making all part of the approach.