Scientometrics (2020) 122:357–379
https://doi.org/10.1007/s11192-019-03274-x

Application of entity linking to identify research fronts and trends

Mauricio Marrone[1]

Received: 15 March 2019 / Published online: 1 November 2019

Abstract
Studying research fronts enables researchers to understand how their academic fields emerged, how they are currently developing and their changes over time. While topic modelling tools help discover themes in documents, they employ a "bag-of-words" approach and require researchers to manually label categories, specify the number of topics a priori, and make assumptions about word distributions in documents. This paper proposes an alternative approach based on entity linking, which links word strings to entities from a knowledge base, to help solve issues associated with "bag-of-words" approaches by automatically identifying topics based on entity mentions. To study topic trends and popularity, we use four indicators—Mann–Kendall's test, Sen's slope analysis, z-score values and Kleinberg's burst detection algorithm. The combination of these indicators helps us understand which topics are particularly active ("hot" topics), which are decreasing ("cold" topics or past "bursty" topics) and which are maturely developed. We apply the approach and indicators to the fields of Information Science and Accounting.

Keywords Natural Language Processing · Content Analysis and Indexing · Burstiness · Information Storage and Retrieval · Text analysis · Entity annotation

Introduction

As ever more academic articles are published, it becomes increasingly challenging for researchers to orient themselves and to remain informed of developments in their rapidly diversifying academic fields. Identifying "core" topics is of great interest to government, industry (Small et al. 2014) and academia (Lee and Kang 2018). As research fronts—the cluster of articles actively cited by researchers (Price 1965)—develop, researchers must grapple with questions, including how particular research fronts emerged, their current state-of-the-art, the critical paths for their evolution, and how research fronts are interconnected (Chen 2006). In analysing research fronts, the detection of "core" topics is of great interest to government, industry (Small et al. 2014) and

✉ Mauricio Marrone
mauricio.marrone@mq.edu.au

[1] Department of Accounting and Corporate Governance, Macquarie University, Building E4A, Room 339, Sydney 2109, Australia

academia (Lee and Kang 2018). Through analysing which topics are rising or falling in popularity, governmental funding boards can make decisions regarding grant allocation to promising areas, companies can design Research and Development (R&D) pursuits for promising technologies and researchers can identify promising topics upon which to focus their work (Lee and Kang 2018). Conducting topic analyses can promote knowledge transfer within and between research domains and assist funding agencies and decision-makers to remain updated about innovations and knowledge flows (Chen et al. 2017). The ability to understand and synthesize historical and emerging ideas, through the analysis of topics, is crucial for researchers to gain insights into how relevant modes of analysis, methods, theory and context are developing (Nederhof and Van Wijk 1997), to generate novel concepts and methods, and for their academic fields to progress (Westgate et al. 2015).

In recent years, studies have developed and presented automated methods to more easily identify emerging topics. These studies can be categorised into two main categories: (1) those based on citation analysis to create structure from datasets and the examination of topics that appear in the clusters; and (2) those that identify rapid growth of publications through text mining. The first category involves citation analysis and bibliometric coupling (Boyack and Klavans 2014; Hopcroft et al. 2004). However, these methods are limited, in so far as high citation counts may not necessarily imply quality; there are fundamental differences between research fields and authors may cite either their own or colleagues' contributions or studies from their target journals (Ivancheva 2008). The second group of methods includes topic modelling techniques such as LDA and LSA as well as the usage of controlled indexing vocabularies.

The approach presented in this paper can be categorised in the second group of methods. We present our method based on entity linking as a way of overcoming limitations associated with various existing methods in this category, including their employment of a "bag-of-words" approach, and their requirement for researchers to manually label categories, specify the number of topics to emerge from the data, and make assumptions about the distributions of words included in the documents (Lee and Kang 2018).

In Natural Language Processing, entity linking is an established approach that enables the automatic identification of topics rather than relying on researchers to manually categorise words. While we follow previous literature in using z-scores and the Kleinberg's burst-detection algorithm, we introduce the use of Mann–Kendall's test and Sen's slope analysis to examine topic trends. By combining these approaches, we aim to understand which research areas are particularly active ("hot" topics), which have decreased in prevalence ("cold" topics or past "bursty" topics) and which are approaching, or past being maturely developed.

The main contributions of the paper are: (1) to propose the use of entity linking to identify research fronts, and (2) to characterize topic trends using Mann–Kendall's test and Sen's slope analysis, enabling a statistically significant interpretation of the results.

The paper commences with a literature review, which analyses recent developments in the fields of Topic Detection and Tracking and the identification of research fronts. The paper then describes the methodology, which is based on entity linking and factor analysis. To highlight the generalisability of the proposed method, we then apply the methodology and indicators to both the top five Information Science (IS) and Accounting journals. The discussion section explores the advantages of using the proposed method.

Literature review

Over the last decade, researchers have been increasingly studying quantitative methods to identify and track research fronts as they evolve over time (Fujita et al. 2014). Researchers have become increasingly interested in using different text-mining techniques, including co-occurrence-based methods (Buzydlowski et al. 2002), automatic key-phrase extraction (Hasan and Ng 2014) and sequence-labelling algorithms, such as named-entity recognition (Nadeau and Sekine 2007; Tjong Kim Sang and De Meulder 2003), to analyse large corpora.

One way of understanding how a field is evolving is by examining citation clusters (Braam and Moed 1991; Jarneving 2007; Small and Griffith 1974). Various citation analysis techniques have been studied for delineating research fronts, including document co-citation (Small 1973), author co-citation (White and Griffith 1981) and those based on coupled networks (Liu et al. 2013). Authors argue that the structural properties of co-citation networks can characterize the emergence, development, application and demise of research areas (Small and Upham 2008), and that co-citation clusters can be used to track the emergence and growth of research areas and their short-term future change (Small 2006).

However, tracking a field using co-citation analysis presents several issues. Bibliometrics can only be used for studying academic literature (Kim and Chen 2015) and it may take years for an article to be cited many times and become widely recognized, with this process taking even longer for fields with smaller scales and/or that develop more slowly (Liu et al. 2013). Some widely-used citation analysis tools, such as Co-cited Networks (CCN), also exacerbate the lag-effect issue. In CCN, the connection between two articles is established by both being cited by a third article, which is published later. The author(s) of the third article must first read the two articles before citing them, exacerbating the time lag in predicting trends (ibid, 2013). There is, therefore, a need for approaches that work for a wide range of literature types and that accelerate the speed at which the dynamics of a research front can be studied.

An ideal approach would enable the identification of research fronts by also observing publications such as policy documents, patents and research-grant soliciting requests. One way of satisfying this purpose is the topic-modelling approach, which has gained popularity in recent years. The approach enables the efficient discovery of meaningful categories, called "topics", from collections of documents, and can be used to understand how topics and research fields change over time. Based on the topic-modelling literature, topics are either a collection of events (non-probabilistic topic model) or represented by a topic model, which is a group of semantically related words and is represented by a probabilistic distribution of the words (probabilistic topic model) (Zhou et al. 2017).

Topic models are statistical algorithms designed to automatically identify the core topics and main themes in large and unstructured collections of documents (Blei 2012). In the most common form, the algorithms are considered "generative" models as they assume documents contain multiple topics according to a probabilistic distribution. Each document is generated by selecting words from topics according to a certain probabilistic document generation process. By analysing the appearances of words in documents, topic modelling algorithms discover the topics of the documents, how the topics are interconnected and how they change over time (Blei 2012).

Most document collections can be organized chronologically and thus characteristics such as topic content and topic frequencies can be seen to evolve over time. To capture this dynamic behaviour, researchers have developed topic models with timestamps (Chen

et al. 2017). Thomas et al. (2014) describe topic evolution models as a branch of topic models that consider time in some way to detect and analyse how topics change or evolve over time. Topic evolution indicates how a topic changes over time, including whether it is maturely developed, imports knowledge from other topics, merges or splits into other topics, and whether topics are increasing or decreasing in importance (Chen et al. 2017). It has been recognized that topics in an academic field not only generate different amounts of interest but the level of interest can increase or decrease over time, reflecting their changing importance (Griffiths and Steyvers 2004).

One seminal work is the Dynamic Topic Model by Blei and Lafferty (2006), which organizes documents into time slices, while Wang et al. (2012) and Gohr et al. (2009) further developed Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis, respectively, for mining chronologically-ordered document streams.

Limitations of existing topic modelling methods

Topic modelling (including extensions to topic evolution) faces a number of key challenges. One of the most challenging tasks is how to name unlabelled topics (Lee and Kang 2018). The statistical algorithms generate groups of terms with statistical correlations (defined as topics), which must be identified and labelled by the researchers after the computational analysis (Schober et al. 2018). While topics tend to be labelled based on the most frequent words in the group of terms, designating a topic based on word frequency and probability distributions is not straightforward (Lee and Kang 2018). The ability to conduct this process effectively depends on the researchers' understanding of the corpora being analysed (Westgate et al. 2015) and requires teams of experts with advanced experience in the academic field (Lee and Kang 2018). Any controversy between researchers or inconsistency in terminology may reduce the usefulness of such automated processes (Westgate et al. 2015).

The selection of the number of topics necessitated by LDA and similar topic modelling approaches also has an arbitrary element. There is no rule that specifies how many topics researchers should examine. Researchers are, instead, left to explain their choice of number of topics.

The underlying assumption of LDA, and probability-generative models more generally, is that a topic is a "bag-of-words" (i.e., a distribution of words) (Zhou et al. 2017). However, while "bag-of-words" methodologies may be appropriate for term-level topic modelling, they may not be effective for clustering documents written by different authors (Michelson and Macskassy 2010).

When analysing topic evolutions, models make strong assumptions about the corpora and how much a topic can evolve. While the Dynamic Topic Model assumes topics evolve following a normal distribution, the Topics Over Time model assumes a topic's lifetime follows a beta distribution (Hall et al. 2008). The assumptions made by these models can be overly restrictive and may not allow for the actual evolution of topics to be identified.

Hot and cold topics

A primary purpose of Topic Detection and Tracking is to identify the "hot" topics in an academic field at a particular point in time (Griffiths and Steyvers 2004), to assist with detecting research fronts (Chen and Guan 2011). Nederhof and van Wijk (1997) identify three types of topics. "Hot" topics are those for which the number of publications is

increasing significantly. For "cold" topics, the number of publications is decreasing significantly. Finally, for "stable" topics, the number of publications is neither significantly increasing nor decreasing. A topic may become "hot" if new, relevant developments occur or if, after becoming "cold", they attract new attention. A topic may become "cold" if associated research becomes less interesting, returns on investment of effort and funding decrease or relevant funding halts. The label "cold topic" does not indicate a topic is dead. Rather, it signifies that the topic was once "hot", but focus on the topic has sharply decreased over time (Garousi and Mäntylä 2016).

Quantitative measures that describe the prevalence of specific types of research are useful for historical reasons (to identify how a field has evolved over time) and to determine targets for research funding (Griffiths and Steyvers 2004). By identifying whether topics are increasing in use ("hot" topic) or decreasing in use ("cold" topic), researchers can direct their attention to topics that can be understood as emerging research areas rather than those that are declining. However, to enable the identification of "hot" and "cold" topics and to distinguish between the two types, quantitative measures are necessary.

Traditionally, mainly linear regression techniques have been used to understand how topic "hotness" changes over time. In the simplest form, topic popularity can be modelled by identifying the impact of the change over time (time dummy variable as the explanatory variable) on the number of published articles per topic (dependent variable) for each topic (Westgate et al. 2015). Another possible dependent variable is the average probabilities of topics (the topic distributions) (Garousi and Mäntylä 2016). In these models, topics with positive random intercepts are understood as having a higher-than-average number of articles written about them over the given time period. A positive random slope signifies that the number of articles discussing a topic has increased significantly over the time period (Westgate et al. 2015).

However, this parametric approach requires the assumption that changes in topic "hotness" over time follow a linear fashion. Higher order parametric forms make similar assumptions; that is, that the trend can be approximated by a fitted regression line. For parametric approximations to be appropriate, however, samples must: be randomly drawn from a normally distributed population; roughly resemble a normal distribution; be identified on an interval or ratio (rather than ordinal or nominal) measurement scale; include only independent observations, apart from paired values; and the population must have approximately equal variances (Corder and Foreman 2014). Given that researchers may wish to study topic evolutions over a small amount of time, with few data points—for example, yearly over a five-year period—observations are likely to evolve in a way that does not satisfy equinormality nor represent a scattering of data points that can be parametrically approximated. These issues highlight the need for nonparametric approximation, which only require that the data are independent (Hamed and Rao 1998).

Method

By analysing the current status of topic modelling tools, we identified four main limitations: (1) the need for researchers to manually label categories; (2) the need to manually specify the number of topics/categories an algorithm should identify; (3) the use of a "bag-of-words" approach; and (4) strong assumptions about the evolution of topics. Our discussion of the identification of "hot" and "cold" topics further highlights the need for a method that accounts for the nonparametric evolution of topics over time.

Fortunately, improvements in Information Retrieval have allowed for the creation of entity linkers, which are employed in this paper to extract topics grouped by time period. By coupling entity linkers with nonparametric approximations of topic evolutions to identify topic trends and popularity, we propose a method that overcomes the identified limitations.

The proposed methodology involves two main stages: (1) the use of entity linking to extract key topics from documents and (2) the use of non-parametric tests to determine "hot" and trending topic areas. In this section, the paper details the methodology before applying it to the top five IS journals, as determined by Google Scholar Metrics. While our discussion focuses on the IS journals, we exemplify the generalizability of the method in our results by also applying it to the top five Accounting journals.

In comparison to traditional topic modelling approaches, the proposed methodology requires fewer assumptions and does not require manual labelling of topics or the specification of a specific number of topics or categories. By using entity linking, the method links appearances of word strings in texts to unambiguous entries in a knowledge base. Mentions with similar meanings are automatically grouped by considering the context in which they appear, rather than implementing a "bag-of-words" approach. By disambiguating terms in this way, entities are automatically categorized in terms of topics similar to the type described in the non-probabilistic topic modelling literature.

First step: data selection

The first step involves extracting the relevant data from a database such as Scopus. The queries used to extract the data are shown in "Appendix". Once all relevant publications are downloaded, the output should be exported as a comma separated values (CSV) file containing year, title and abstract information for each article. Next, the researchers delete any duplicate publications and merge the titles of the publications with the abstracts.

Table 1 provides a breakdown of the sources of articles included in the analysis.

	2014	2015	2016	2017	2018	Total
Information Science Journals						
Journal of Academic Librarianship	93	108	94	69	105	469
Journal of Informetrics	92	84	104	106	92	478
Journal of the Association for Information Science and Technology	162	201	225	225	146	959
Online Information Review	53	52	63	64	109	341

Table 1 Example of using TAGME to annotate texts. *Source* Marrone and Hammerle (2017)

Publication (Bender 2014)	*Sedgman is a **mining engineering** company based in **Brisbane** with offices globally and mine operation sites across **Australia**. It has more than 650 staff. Reid said his IT responsibilities include operations, **service desk**, service management, network, **server**, engineering applications for the company globally.*
TAGME result	Mining engineering, Brisbane, Australia, service desk (ITSM), server (computing)

	2014	2015	2016	2017	2018	Total
Scientometrics	397	325	379	393	424	1918
Total	797	770	865	857	876	4165
Accounting Journals						
Accounting Review	81	93	72	62	88	396
Contemporary Accounting Research	46	66	61	74	80	327
Journal of Accounting and Economics	33	35	49	44	50	211
Journal of Accounting Research	36	31	32	34	34	167
Review of Accounting Studies	48	49	34	50	46	227
Total	244	274	248	264	298	1328

Second step: topic extraction

Our approach implements entity linking to identify keywords. Such keywords are meaningful notions that represent the text and are henceforth referred to as "topics". As the extracted topics represent the content of the publications, they can highlight themes and changes in a research topic or journal (Bayramusta and Nasir 2016).

Entity linking (EL) is the task of identifying short and meaningful sequences of terms (entities) in an input text and annotates (disambiguates) these entities with unambiguous identifiers from a catalogue (Cornolti et al. 2013). Disambiguation is achieved by establishing a link to a relevant entry in a knowledge base (catalogue), which uniquely identifies the entity and provides further information about it (Cuzzola et al. 2015). One of the most widely used catalogues for EL is Wikipedia, as it covers an enormous and ever-increasing number of entities, has wide content coverage and includes special features such as "disambiguation" pages and unique identifiers for each page (Cornolti et al. 2013; Ferragina and Scaiella 2010; Khalid et al. 2008).

As an example, entity linking would remove any ambiguity concerning the term "Paris" by linking it to the abstract topic "city" (Uren et al. 2006). Different terms referring to the same topic are linked and treated as this topic, meaning that "U.S.", "USA" and "United States" would be normalized to "United States of America" (Khalid et al. 2008). As synonyms and ambiguous entities used in different documents are normalized to unambiguous topics, analysing discourses about topics across articles becomes more effective and efficient.

For our entity linker, we use TAGME, a software application that annotates term sequences using hyperlinks to Wikipedia articles (Ferragina and Scaiella 2010). Outperforming other software for short texts (Ferragina and Scaiella 2010; Kulkarni et al. 2009), TAGME is particularly suitable for annotating documents such as journal article abstracts and newspaper articles.

To better understand how TAGME works, Table 1 provides an example of how TAGME has annotated a sentence on the topic of Information Technology Service Management (ITSM). The words in brackets contextualize the mention by taking into account how it is referred to in the sentence. In applying TAGME, the values for the area-under-the-curve F-measure were set as the stochastic setting of tuneable parameters

(long_text 10, epsilon 0.427, $q = 0.1613$), following Cuzzola et al. (2015). These values define the annotation process.

Third stage: topic cleansing

After applying TAGME, all false positives—topics that make little meaningful sense given the context in which they appear—were deleted. For example, TAGME linked "to show" with the American television series "The T.O. Show". We also deleted terms including *research* and *publishing,* which are too general to describe the IS field, and *Emerald, Elsevier, Hungary* and *Budapest*, which were copyright information at the end of some abstracts. Finally, we set a minimum threshold frequency of ten, meaning that topics had to appear at least ten times in the journal articles to be considered in the analysis.

Fourth stage: trend analysis

To estimate trends (and thus "hot" and "cold" topics), researchers should estimate the frequency of the topic clusters and their corresponding topics (Walshe 2009). The package *matplotlib* in *Python* can plot the frequencies of each topic. We calculated the normalised frequency for each topic by dividing the number of appearances of the keyword of interest in a particular time period by the total number of keywords for that time period.

Fifth stage: determining topic trends and popularity

Topic trends

To investigate topic trends, this paper uses both Mann–Kendall's test and Sen's slope analysis, both of which have primarily been used for meteorological time series analysis. As nonparametric techniques, they require fewer assumptions than parametric trend tests and are robust to outliers in the data (Hamed and Rao 1998).

Mann–Kendall's test and Sen's slope analysis have predominantly been applied in meteorological studies (e.g., Gocic and Trajkovic 2013; Zhang and Lu 2009) such as for analysing temperature fluctuations (e.g., Kaushal et al. 2010) and rainfall and river flow chronological trends (e.g., da Silva et al. 2015). However, none of the pioneers of these tests (e.g., Mann 1945; Sen 1968), mention that they are only suitable for meteorological studies, preferring to focus discussions instead on the properties of the indicators. Their use to study chronological trends in other areas highlights their potential to be more widely applicable. For example, both tests have been used to analyse trends in a recreational lobster fishery (Sharp et al. 2005) and to assess the impact of technological developments on radiologists' workload (McDonald et al. 2015). The example most similar to our purposes is that of McDonald et al. (2010), who identify trends in the average number of authors per article using Sen's slope. Given the capacity of Sen's slope analysis to identify the speed of change of a time series, this paper argues that it can also be applied to identifying those topics that are increasing or decreasing in popularity most quickly.

Mann–Kendall's test The Mann–Kendall test assesses the correlation between the rank order of observed values and their temporal ordering (Hamed and Rao 1998). The null hypothesis is that the sample data is independent and identically distributed. The alternative hypothesis states that the data sample has a monotonic trend (Zhang and Lu 2009).

The Mann–Kendall test statistic is robust against non-normally distributed, censored and missing data (Yue et al. 2002) and is powerful compared to parametric competitors (Zhang and Lu 2009). It is also asymptotically normal (Hamed and Rao 1998) and can be used for sample sizes above four.

Nonparametric tests require that the data are independent (Hamed and Rao 1998). However, in many real situations, including the present scenario with topic trends, the observed data are likely autocorrelated, resulting in misinterpretation of trend-test results (ibid 1998) and an increased likelihood of falsely finding statistical significance without a trend being apparent, in the presence of positive serial correlation (Cox and Stuart 1955). While several possibilities exist to deal with this autocorrelation, this paper uses the Variance Correction Approach by Hamed and Rao (1998). Using this approach, Mann–Kendall trend-test results are robust in the presence of autocorrelation (Hamed and Rao 1998), resulting in the test being used where researchers are concerned about autocorrelation (e.g., Han et al. 2014; Stojković et al. 2014).

Sen's slope analysis While the Mann–Kendall statistic is used to identify trends, Sen's slope analysis is used to estimate the magnitude of trends detected by the Mann–Kendall test (Zhang and Lu 2009), thus calculating the speed of change of a time series trend. Sen's slope is the associated slope estimate for the Mann–Kendall statistical test (ibid 2009). After applying Mann–Kendall's test, this paper uses Sen's slope to identify which topic trends are increasing or decreasing the quickest.

Sen's slope is measured as a change in observed values per unit of time (Zhang and Lu 2009) and is calculated as the median of the set of (linear) slopes joining time-ordered pairs of points (Sen 1968). The sign of Sen's slope for each topic indicates whether the trend is increasing (positive) or decreasing (negative), while its absolute value indicates how quickly the trend is changing.

Topic popularity

A central problem for text mining is extracting meaningful structure from document streams that can be ordered chronologically. The published literature for a particular research field is characterized by topics that appear, increase in intensity over some period of time and then gradually disappear (Kleinberg 2003). To investigate topic popularity, this paper uses both z-scores to identify "hot" and "cold" topics and the burst-detection algorithm established by Kleinberg (2003) to highlight "bursty" topics.

Kleinberg's burst detection algorithm Alongside considering the textual features of the documents, another way to characterize topics is to examine the pattern of topic appearances over particular time periods, exposing greater fine-grained structure (Kleinberg 2003). Kleinberg (2003) developed a bursting algorithm based on the understanding that the appearance of a topic in a document stream is signalled by a "burst of activity" and that, as the topic emerges, certain features also rapidly increase. The algorithm aims to extract global structure, identifying bursts only if they have sufficient intensity and

based on the understanding that topics may appear in document streams in non-uniform patterns. This paper argues that, by detecting "bursty" topics, it is also possible to identify which topics are "hot" or "cold" over time.

This paper uses the Kleinberg algorithm as a way to robustly and efficiently identify bursts in topic appearances, thus providing an organizational framework to analyse document streams.

***z*-scores** The z-score transformation procedure is a widely-used statistical method for normalizing data (Cheadle et al. 2003). In the general sense, it expresses how far a value is from the population mean (how "unusual" a particular value is), expressing this difference in terms of numbers of standard deviations (Kirkwood and Sterne 2003).

More recently, researchers have applied an adjusted z-score to identify topics that experience sharp temporal increases (i.e. "hot" topics) (Huang et al. 2017). Their studies measure term novelty as the normalized difference between the term's predicted frequency (from past realizations) and its actual frequency:

$$z = \frac{\left(\text{current trend} - \left[\text{average of historic trends}\right]\right)}{\text{standard deviation of historic trends}}$$

Given a topic's observed frequency in the current time period, it is possible to compare this frequency with the average frequencies of the topic in the past and express this difference in terms of standard deviations. The more variable the topic frequency has been in the past, the less likely the z-score value will be large and the less likely the topic will be identified as "hot" or "cold". The "hottest" topics are likely to have experienced a recent, sharp increase in frequency, with previously minimal fluctuations in use.

Results

In the results section, the paper first considers trending topics before examining "hot" and "cold" topics over the last 5 years in the IS literature.

Topic Trends

Mann–Kendall and Sen's slope analysis

Based on Mann–Kendall and Sen's slope analysis, Tables 2 and 3 highlight the top ten most quickly increasing and decreasing topics in terms of their frequency of use in the IS and Accounting literature. p values less than 0.05 signify a monotonic trend. If the p value is positive, there exists a strictly increasing trend; if the p value is negative, there is a decreasing trend.

Although there are more topics with significantly decreasing trends than with significantly increasing trends, the significantly increasing topics increased at a faster rate over the time period. The most quickly increasing topic, *academia*, is increasing at a magnitude more than twice as fast as the most quickly decreasing topic is decreasing in use.

While some topics steadily decrease over time, other, related topics increase. This relationship can be observed, with *curriculum* decreasing while *undergraduate education* increases. This dichotomy exemplifies the level of detail our method provides as

Table 2 Most quickly increasing topics, based on Mann–Kendall and Sen's slope analysis

	Mann–Kendall z-score	Mann–Kendall p value	Sen's slope
Information science topics			
Scopus	2.2045	0.0275	0.0005
Academia	2.2045	0.0275	0.0005
Microsoft	2.2045	0.0275	0.0004
Consumer	2.2045	0.0275	0.0003
Data analysis	2.2045	0.0275	0.0002
Internet forum	2.2045	0.0275	0.0002
Undergraduate education	2.2045	0.0275	0.0002
Public relations	2.0212	0.0433	0.0002
Valence (psychology)	2.0212	0.0433	0.0001
Intuition	2.0212	0.0433	0.0001
Hybrid open access journal	2.0212	0.0433	0.0001
Cardiology	2.0212	0.0433	0.00005
Accounting topics			
Regulation	1.7450	0.0275	0.008
Profit	1.3594	0.0275	0.002
Fundamental analysis	1.3563	0.0275	0.002
Probability	1.2823	0.0275	0.002
Narcissism	1.2395	0.0433	0.003

These topics are all those with a statistically significant (5% level) increasing trend, as per the Mann–Kendall test

Table 3 Most quickly decreasing topics, based on Mann–Kendall and Sen's slope analysis

	Mann–Kendall z-score	Mann–Kendall p value	Sen's slope
Information Science topics			
Science citation index	− 2.2045	0.0275	− 0.0004
Curriculum	− 2.2045	0.0275	− 0.0003
Book	− 2.2045	0.0275	− 0.0002
G-index	− 2.2045	0.0275	− 0.0002
Uniform resource locator	− 2.2045	0.0275	− 0.0002
Big science	− 2.2045	0.0275	− 0.0002
Environmental science	− 2.2045	0.0275	− 0.0001
Web search query	− 2.2045	0.0275	− 0.0001
Web 2.0	− 2.2045	0.0275	− 0.0001
Biotechnology	− 2.2045	0.0275	− 0.0001
Performance indicator	− 2.2045	0.0275	− 0.0001
Reliability engineering	− 2.2045	0.0275	− 0.0001
Query expansion	− 2.0212	0.0433	− 0.0002
Returns to scale	− 2.0212	0.0433	− 0.0002
Accounting topics			
Auditor independence	− 1.0081	0.0275	− 0.0014

These topics are all those with a statistically significant (5% level) decreasing trend, as per the Mann–Kendall test

compared to traditional topic modelling tools. For example, as researchers manually constrain the number of categories LDA produces, it may have grouped both *curriculum* and *undergraduate education* in a common *education* 'topic'. While our approach shows how one topic increases in popularity while the other declines, the LDA 'topic' *education* may have appeared steadily over the analysis period.

The topic *undergraduate education* is often combined with the word *employment* (e.g., Adams et al. 2016) and with the administration of surveys to undergraduate students (Chao and Yu 2018; Sun et al. 2017). Researchers are also interested in analysing undergraduate student performance (Soria et al. 2014) and how they integrate into university life, including their experiences of anxiety (Sinnasamy and Karim 2014). While few papers discuss teaching methods in combination with *undergraduate education*, researchers who examine the topic *curriculum* also assess the effectiveness of teaching methodologies (e.g., Schmidt and English 2015). Researchers are interested in how to integrate information literacy in the educational curriculum (Moselen and Wang 2014; Salmerón et al. 2016) and the connections between *curriculum* and real-world experience in teaching information literacy (Young and Maley 2018).

Topic popularity

Kleinberg's burst detection algorithm (2003)

To visualize the results of Kleinberg's bursting algorithm, Figs. 1 and 2 are heat maps highlighting the "bursty" topics alongside their "bursting" years. The scale indicates the intensity of the topic appearances. The lightest blue boxes signal that a topic was very rarely spoken about in the corresponding time period, while the darkest blue boxes highlight the "bursting" periods for associated topics.

The figures show how different topics burst at different times over the past 5 years, and also had different proportions of time over which they burst. While most topics only burst for 1 year, the longest "bursting" topics in the IS literature are *Science Citation Index, Institute for Scientific Information, automation, Microsoft* and *Research Gate*. Unlike "bag-of-words" methods, the proposed method groups phrases in texts with the same meaning. For example, *Science Citation Index* is referred to as "Science Citation Index", "SCI", "Science Citation Index-Expanded", "SCI Expanded", "SCI-E" and "SCIE" by IS articles (Elango et al. 2016; Fernández et al. 2016; Zhou and Lv 2015). While the method proposed groups these terms, a "bag-of-words" approach would consider the terms as separate mentions, thereby reducing the effectiveness of comparing documents that describe the same concept using different terminology.

A key need raised by Chen et al. (2017) is to better understand how terms transition over time. While the figure highlights that citation indexes have been a "hot", overarching, topic of discussion over the past 5 years, different associated topics have burst at different times. By observing the figure, one comes to understand how results evolved over time. For example, to describe the varying focuses of the IS field, *infrastructure* and *Triple Helix* were first popular before *consumer* and *business product* became more popular.

There has been an increase in the number of literature reviews published in the included journals, explaining the burst in the topic *systematic review*. Among other areas, reviews have been published on public relations intelligence (Santa Soriano

Fig. 1 "Bursty" topics in Information Science, based on Kleinberg's burst detection algorithm

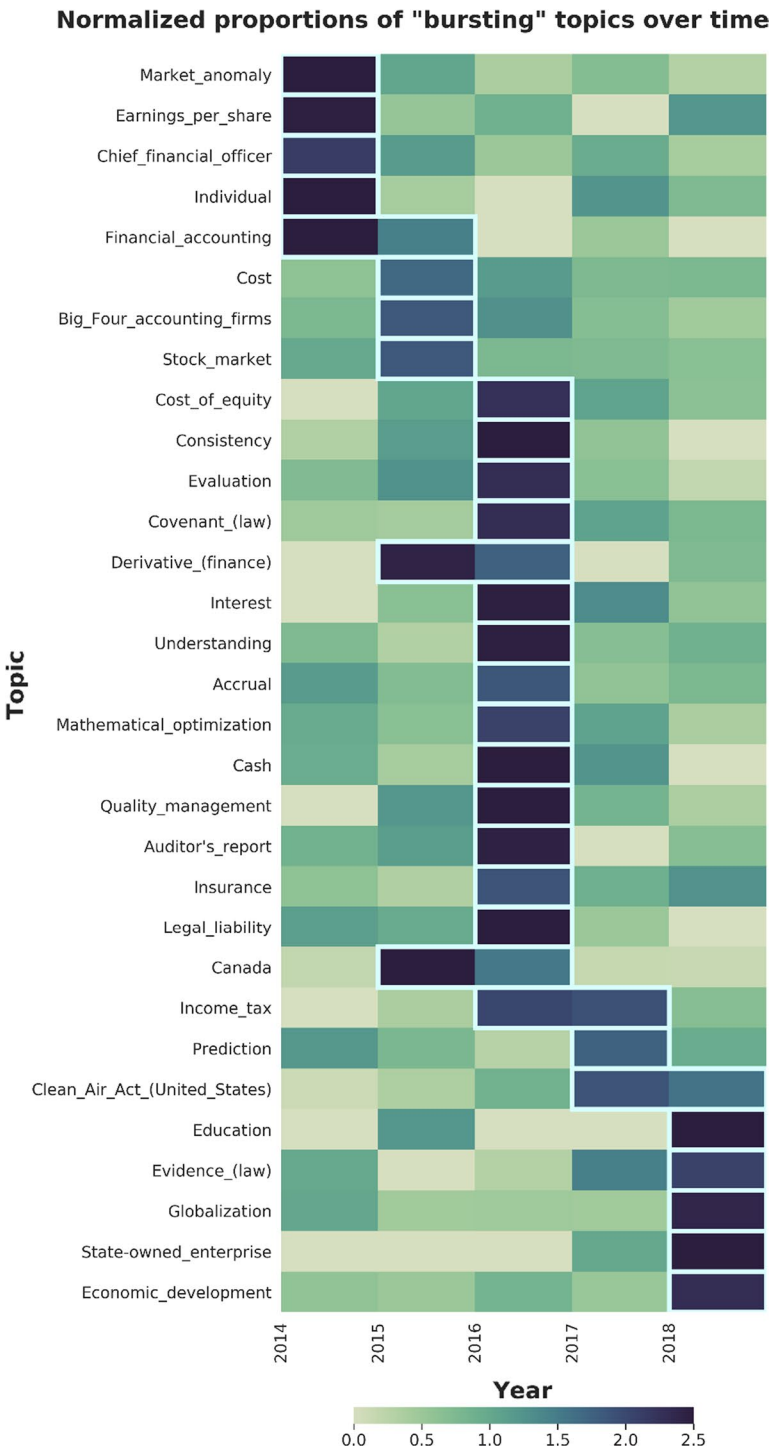

Fig. 2 "Bursty" topics in Accounting, based on Kleinberg's burst detection algorithm

et al. 2018), innovation research (Rossetto et al. 2018), and engineering information literacy instruction (Phillips et al. 2018).

z-scores

Compared to Kleinberg's burst detection algorithm, *z*-score analysis focuses in greater detail on topics that have burst in the most recent year. Another way of understanding "bursty" topics is to describe them as "hot": they have experienced sharp increases in use compared to their past trends. As the Kleinberg's algorithm is different from *z*-score analysis and may only capture the most "bursty" topics, there may be slight differences in the results.

Table 4 presents the top ten "hottest" and "coldest" topics. Based on the comparative magnitudes of the *z*-scores, Table 4 indicates a larger increase in use compared to historical trends for the "hottest" topics than there was a decrease in use for the "coldest" topics. The *z*-score for the "hottest" topic in IS, *Eugene Garfield*, is nearly 2.5 times larger than that for the "coldest" topic in IS, *scientific modelling*. This finding suggests that the "hottest" topics are further away, in terms of historic standard deviations, from their historic means than the "coldest" topics.

Table 4 "Hottest" and "coldest" topics, based on *z*-score values

"Hottest" topics in Information Science	*z*-score	"Coldest" topics in Information Science	*z*-score
Eugene Garfield	53.5	Scientific modelling	−26.0
Predation	28.5	Collaboration	−8.0
Risk	19.1	Binary relation	−6.1
Anger	17.5	Mathematical optimization	−4.3
PLOSONE	16.5	Social network analysis	−3.9
Sleeping Beauty (paper with delayed recognition)	15.5	Social capital	−3.7
Sexism	14.9	Thomson Reuters	−2.7
Memory	11.8	Book	−2.7
Online shopping	9.7	Heuristics	−2.6
Social Influence	9.6	Operations Research	−2.5
"Hottest" topics in Accounting	*z*-score	"Coldest" topics in Accounting	*z*-score
Economic development	2.0	Price	−1.8
State-owned enterprise	2.0	Social influence	−1.8
Consultant	2.0	Cost of capital	−1.7
Economic indicator	2.0	Pricing	−1.7
Capital (economics)	2.0	Financial statement	−1.7
Twitter	2.0	Auditor	−1.6
Inflation	2.0	Financial audit	−1.6
Natural experiment	1.9	Emotion	−1.6
Education	1.9	Stakeholder (corporate)	−1.6
Securitisation	1.9	Perception	−1.6

By examining the instances in the IS literature in which particular topics appear, an increase in the use of the topic *predatory*—to describe predatory science, academia and publications (fake journals)—becomes apparent. Researchers examine the spread of predatory publications (Demir 2018; Perlin et al. 2018) and regard predatory academia as a growing problem (da Silva and Tsigaris 2018; Perlin et al. 2018). Others argue that there is a lack of knowledge and awareness of predatory publications and urge academics to exercise caution in selecting conferences (Lang et al. 2018).

Topics labelled as "cold" are those that have experienced a disproportionate decrease in 2018. When discussing the topic *collaboration,* IS researchers typically examine collaboration patterns, networks and dynamics. However, the topic decreased significantly in 2018, appearing only 64 times compared to an average of a very stable 81 times over the four previous years.

Discussion

In this section, the paper focuses on the differences between how this paper investigates and visualizes topic trends and popularity, as compared to existing literature. The paper provides a new approach to identifying research fronts using entity linkers and applies a variety of non-parametric indicators to develop broad insights into document streams and how topics evolve over time. In particular, the paper (1) introduces entity linking to provide a more granular view of topics when studying research fronts; and (2) is the first to combine Mann–Kendall's test and Sen's slope analysis to understand how topics evolve over time. The approach could potentially be useful for a wide range of groups, notably researchers, decision makers, corporations and research students, to detect and visualize trends and rapid changes in different literature types over time.

The transient and rapidly evolving nature of a research front presents unique challenges for researchers, policy makers and other groups to remain up-to-date with changes. Thus, researchers have conducted scientific topic evolutions to keep updated with their fields and associated topics (Chen et al. 2017). Understanding the dynamics of a research front helps facilitate knowledge transfer within and across research domains, assists various groups to remain updated with changes in the field and presents a wealth of ideas (Ding and Stirling 2016). Monitoring research trends also helps with resource allocation and technological forecasting and is therefore particularly interesting to policy makers (Chen and Guan 2011). With the rapid increase in information, academic areas have become specialized and segmented, presenting both a challenge and an opportunity for analysing research fronts. While it has become increasingly difficult for researchers to understand their specialized fields as a whole, opportunities for cross-disciplinary studies have emerged (Fujita et al. 2014). Identifying relationships between funding trends and existing knowledge bases have become increasingly important for scientists, universities and research laboratories who wish to remain competitive (Chen 2006).

Our research contributes to the discussion on how to identify and track research fronts as they evolve over time. According to Chen (2006), common questions regarding research fronts include how they emerged, their current status and the critical paths in their evolution. To address these questions, it is necessary to detect and analyse trends and rapid changes over time by examining a research front's literature base, understanding significant turning points as the research front evolves, and discovering interconnections between

different research fronts. Topics selected based on rapid changes in popularity measures—in this paper, "hot" and "cold" topics and topics identified using Kleinberg's bust detection algorithm—are particularly appropriate for understanding how research fronts evolve (ibid, 2006).

The use of entity linking provides a more granular approach to topic modelling. Rather than relying on a "bag-of-words" approach, equivalent terms are identified and grouped, which reduces the potential for researcher bias and subjectivity to influence results, thus increasing comparability between studies. Kleinberg (2003) argues that the bursts for *data, base* and *bases* arise because *database* appeared as two words in a number of paper titles during the period. This example highlights how the "bag-of-words" approach focuses on the terms themselves, rather than elevating the analysis to consider the context in which terms appear. Using the entity linking methodology proposed in this paper, these terms would be grouped, most likely appearing as *database*. Kleinberg (2003) also shows that the "bag-of-words" approach identifies such topics as *some, on, improved* and *how* as "bursty". While these topics may indeed be representative of titling conventions fashionable during certain periods, as Kleinberg (2003) suggests, in themselves, they do not provide researchers much insight into the complexities of the topics dealt with by the documents. Using our methodology, these terms would most likely be replaced by overarching topics that better depict the content complexity.

With typical topic modelling tools, analysts must apply extensive sense-making efforts to efficiently synthesize word profiles into clusters. However, cluster labels formed by aggregating words are often too broad to be useful (Chen 2006). Words associated with emerging trends could also be lost amidst larger and more persistent themes. For example, a rapid increase in interest for healthcare, combined with biological and chemical weapon threats, could be overshadowed by a larger and more dominant biological weapon cluster (ibid, 2006). While some attempts have been made to automate cluster identification (e.g. Wallace et al. 2009), these attempts are unable to eliminate semantic ambiguities during keyword extraction and clustering (Liu et al. 2013). Most methods for detecting trends are independent within single clusters, and thus place inadequate attention on the connections between disciplines, fields and knowledge clusters that are widely discussed in knowledge research and acquisition processes (Griffith et al. 1974; Small and Griffith 1974). Other work that has been done to avoid semantic ambiguities and increase the effectiveness of keyword clustering (e.g., Kontostathis and Pottenger 2006; van Eck et al. 2010) requires complex computations and involves difficulties regarding mass data processing (Liu et al. 2013). In comparison, the proposed entity linking method automatically considers both polysemy and synonymy to identify topic appearances in different literature types. The method automatically groups words with the same meanings and attributes mentions with different meanings with different annotations, without the need for explicit clustering.

Aligned with the need for researchers to manually label categories, traditional topic modelling tools also require researchers to specify the number of topics their algorithm should find. When conducting a LDA topic extraction, for example, Doumit and Minai (2012) set the algorithm to generate ten topics for each news source and to characterize each topic with ten words to form a topic signature. It is challenging for researchers to decide how many topics should ideally be drawn from any given document. As an alternative, our approach requires assumptions about the optimal parameter values, however, does not require researchers to specify the number of topics.

Finally, our approach allows for the topic distributions to emerge from the data without requiring strong assumptions about the evolution of topics. Common topic modelling tools

often assume either particular distributions for topics or use a linear regression approach to identify the "hotness" or "coldness" of topics (Hall et al. 2008; Westgate et al. 2015). However, particularly for short time windows, it is unlikely that the topics evolve according to a linear model or that they conform to specific "ex post" assumptions regarding their evolution. Rather, our method enables topics to emerge from the data by using a nonparametric approach to topic modelling.

While Mann–Kendall's test and Sen's slope analysis have thus far mainly been used in meteorological studies, the seminal papers that introduce them do not argue for an exclusive applicability. Instead, in introducing Sen's slope, Sen (1968) focuses on providing a (general) alternative to the least squares estimator β, which is vulnerable to gross errors and its confidence interval to non-normality. The proposed analysis is applied to a general set of numbers, without mentioning any particular academic field. While the indicators have since been applied almost exclusively to understanding climate fluctuations, such as temperature and water accessibility, their use by other authors to study different phenomena, from lobster fisheries to bibliometrics, supports a wider applicability. This paper provides exploratory evidence that the indicators provide an efficient way to study whether topic trends are strictly increasing or decreasing. To the best of our knowledge, we are the first to employ Mann–Kendall's test and Sen's slope analysis to understand topic evolution and the first to combine the algorithms with entity linking in the IS literature.

In this paper, we apply our approach to two academic fields: information sciences and accounting. However, the methodology could also be used to identify changes in the topics discussed further afield, including in different types of literature such as patents, grants and practitioner journals. An analysis of the trends in governmental policy papers (for example, policies on information security and management) may also benefit from the application of this approach. TAGME's use of Wikipedia makes the approach amenable to a great number of different topics. Regarding academia, a vast array of academic fields, such as biology, medicine and chemistry may find it useful to use this approach to gain a rapid and thorough understanding of the way a field has progressed over time. For practitioners, the approach could be applied to understanding changes in organisational policies as well as to analyse trends in annual reports, for example, how different information technologies are viewed and where attention has been placed over time.

As with any other research endeavour, there are various limitations to the method and indicators proposed in this article. First, while entity linking only requires limited manual intervention on the part of the researchers, those who wish to use this approach must nonetheless first become familiar with the method. However, while the learning process may take some time, it is an activity that only needs to occur once and can be used to better understand topics in multiple fields. While topic modelling approaches such as LDA may serve a similar purpose, the entity linking approach has a number of benefits, including disambiguating terms using a standardized knowledge base, error reduction, time savings and being an automated and repeatable approach. Secondly, without the evolution of topic frequencies being normally distributed, the process of identifying "hot" and "cold" topics through z-score analysis does not provide statistically significant results. While we focus on the topics with the ten highest z-scores as "hot" and those with the ten most negative z-scores as "cold", this choice is arbitrary and may not be appropriate for every literature collection. Finally, while we use a timeline of 5 years to more closely focus upon the trends during this period, this is a rather small timeline. While a ten-year period would have provided more data, it would have prevented us from identifying those topics that increased or decreased continuously in the recent past.

Conclusion

This paper uses four indicators—Mann–Kendall's test, Sen's slope, z-score analysis and Kleinberg's burst detection algorithm—to examine topic trends and popularity in the IS literature over the past 5 years. The main contribution of our article is to propose a new approach, based on entity linking, to identify research fronts. To the best of our knowledge, the paper is also the first to use Mann–Kendall's test and Sen's slope to understand whether topics monotonically trend over time and innovatively combines several approaches with entity linking to investigate topic popularity with greater complexity than would be afforded with one method only.

While we use the IS and Accounting literature as an example, we believe the methodology and indicators we discuss would also be helpful to researchers who wish to examine topic evolution in other fields. Analysts and other stakeholders require tools that can turn vast amounts of data into clear and constructive messages (Chen 2006). Analysing topic trends and popularity in the way this paper discusses would assist researchers to gain insights into the development phases of topics in their academic fields and where they should concentrate their efforts, guide firms in deciding how to structure their Research and Development and where collaborations could be sought with academia, and support governments in deciding where to target grants.

Appendix: queries used

The query used for Accounting is:

EXACTSRCTITLE ("Accounting Review") OR EXACTSRCTITLE ("Journal of Accounting and Economics") OR EXACTSRCTITLE ("Journal of Accounting Research") OR EXACTSRCTITLE ("Contemporary Accounting Research") OR EXACTSRCTITLE ("Review of Accounting Studies") AND (LIMIT-TO (EXACTSRCTITLE, "Contemporary Accounting Research") OR LIMIT-TO (EXACTSRCTITLE, "Accounting Review") OR LIMIT-TO (EXACTSRCTITLE, "Journal Of Accounting And Economics") OR LIMIT-TO (EXACTSRCTITLE, "Journal Of Accounting Research") OR LIMIT-TO (EXACTSRCTITLE, "Review Of Accounting Studies")) AND (LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014)).

The query used for Information Science is:

EXACTSRCTITLE ("Journal of the Association for Information Science and Technology") OR EXACTSRCTITLE ("Scientometrics") OR EXACTSRCTITLE ("Journal of Informetrics") OR EXACTSRCTITLE ("Journal of Academic Librarianship") OR EXACTSRCTITLE ("Online Information Review") AND (LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015)

OR LIMIT-TO (PUBYEAR, 2014)) AND (LIMIT-TO (EXACTSRCTITLE, "Scientometrics") OR LIMIT-TO (EXACTSRCTITLE, "Journal Of The Association For Information Science And Technology") OR LIMIT-TO (EXACTSRCTITLE, "Journal Of Informetrics") OR LIMIT-TO (EXACTSRCTITLE, "Journal Of Academic Librarianship") OR LIMIT-TO (EXACTSRCTITLE, "Online Information Review")).

References

Adams, C., Buetow, S., Edlin, R., Zdravkovic, N., & Heyligers, J. (2016). A collaborative approach to integrating information and academic literacy into the curricula of research methods courses. *The Journal of Academic Librarianship, 42*(3), 222–231. https://doi.org/10.1016/j.acalib.2016.02.010.

Bayramusta, M., & Nasir, V. A. (2016). A fad or future of IT?: A comprehensive literature review on the cloud computing research. *International Journal of Information Management, 36*(4), 635–644. https://doi.org/10.1016/j.ijinfomgt.2016.04.006.

Bender, A. (2014). Sedgman unearths greater value for cost with ITSM switch. *Computerworld*. Retrieved February 1, 2019, from http://www.computerworld.com.au/article/543129/sedgman_unearths_greater_value_cost_itsm_switch/.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). https://doi.org/10.1145/1143844.1143859.

Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology, 65*(4), 670–685. https://doi.org/10.1002/asi.22990.

Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. II: Dynamical aspects. *Journal of the American Society for Information Science, 42*(4), 252–266. https://doi.org/10.1002/(SICI)1097-4571(199105)42:4%3c252:AID-ASI2%3e3.0.CO;2-G.

Buzydlowski, J. W., White, H. D., & Lin, X. (2002). Term co-occurrence analysis as an interface for digital libraries. In C. Chaomei (Ed.), *Visual interfaces to digital libraries* (pp. 133–144). Berlin: Springer.

Chao, C.-M., & Yu, T.-K. (2018). The moderating effect of technology optimism: How it affects students' weblog learning. *Online Information Review, 43,* 161–180.

Cheadle, C., Vawter, M. P., Freed, W. J., & Becker, K. G. (2003). Analysis of microarray data using Z score transformation. *The Journal of Molecular Diagnostics, 5*(2), 73–81. https://doi.org/10.1016/S1525-1578(10)60455-2.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377. https://doi.org/10.1002/asi.20317.

Chen, K., & Guan, J. (2011). A bibliometric investigation of research performance in emerging nanobiopharmaceuticals. *Journal of Informetrics, 5*(2), 233–247. https://doi.org/10.1016/j.joi.2010.10.007.

Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics, 11*(4), 1175–1189. https://doi.org/10.1016/j.joi.2017.10.003.

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach*. New York: Wiley.

Cornolti, M., Ferragina, P., & Ciaramita, M. (2013). A Framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 249–260). Retrieved February 1, 2019, from http://dl.acm.org/citation.cfm?id=2488388.2488411.

Cox, D. R., & Stuart, A. (1955). Some quick sign tests for trend in location and dispersion. *Biometrika, 42*(1/2), 80–95. https://doi.org/10.2307/2333424.

Cuzzola, J., Jovanović, J., Bagheri, E., & Gašević, D. (2015). Evolutionary fine-tuning of automated semantic annotation systems. *Expert Systems with Applications, 42*(20), 6864–6877. https://doi.org/10.1016/j.eswa.2015.04.054.

da Silva, J. A. T., & Tsigaris, P. (2018). What value do journal whitelists and blacklists have in academia? *The Journal of Academic Librarianship, 44*(6), 781–792.

da Silva, R. M., Santos, C. A. G., Moreira, M., Corte-Real, J., Silva, V. C. L., & Medeiros, I. C. (2015). Rainfall and river flow trends using Mann–Kendall and Sen's slope estimator statistical

tests in the Cobres River basin. *Natural Hazards, 77*(2), 1205–1221. https://doi.org/10.1007/s11069-015-1644-7.
Demir, S. B. (2018). Predatory journals: Who publishes in them and why? *Journal of Informetrics, 12*(4), 1296–1311.
Ding, Y., & Stirling, K. (2016). Data-driven discovery: A new era of exploiting the literature and data. *Journal of Data and Information Science, 1*(4), 1–9.
Doumit, S., & Minai, A. (2012). *Online News Media Bias Analysis using an LDA-NLP Approach*.
Elango, B., Bornmann, L., & Kannan, G. (2016). Detecting the historical roots of tribology research: A bibliometric analysis. *Scientometrics, 107*(1), 305–313.
Fernández, A., Ferrándiz, E., & León, M. D. (2016). Proximity dimensions and scientific collaboration among academic institutions in Europe: The closer, the better? *Scientometrics, 106*(3), 1073–1092.
Ferragina, P., & Scaiella, U. (2010). TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 1625–1628). https://doi.org/10.1145/1871437.1871689.
Fujita, K., Kajikawa, Y., Mori, J., & Sakata, I. (2014). Detecting research fronts using different types of weighted citation networks. *Journal of Engineering and Technology Management, 32,* 129–146. https://doi.org/10.1016/j.jengtecman.2013.07.002.
Garousi, V., & Mäntylä, M. V. (2016). Citations, research topics and active countries in software engineering: A bibliometrics study. *Computer Science Review, 19,* 56–77. https://doi.org/10.1016/j.cosrev.2015.12.002.
Gocic, M., & Trajkovic, S. (2013). Analysis of changes in meteorological variables using Mann–Kendall and Sen's slope estimator statistical tests in Serbia. *Global and Planetary Change, 100,* 172–182.
Gohr, A., Hinneburg, A., Schult, R., & Spiliopoulou, M. (2009). Topic evolution in a stream of documents. In *Proceedings of the 2009 SIAM international conference on data mining* (Vols. 1–0, pp. 859–870). https://doi.org/10.1137/1.9781611972795.74.
Griffith, B. C., Small, H. G., Stonehill, J. A., & Dey, S. (1974). The structure of scientific literatures II: Toward a macro- and microstructure for science. *Science Studies, 4*(4), 339–365. https://doi.org/10.1177/030631277400400402.
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.
Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 363–371). Retrieved February 1, 2019, from http://dl.acm.org/citation.cfm?id=1613715.1613763.
Hamed, K. H., & Rao, A. R. (1998). A modified Mann–Kendall trend test for autocorrelated data. *Journal of Hydrology, 204*(1–4), 182–196.
Han, J.-C., Huang, G.-H., Zhang, H., Li, Z., & Li, Y.-P. (2014). Heterogeneous precipitation and streamflow trends in the Xiangxi River watershed, 1961–2010. *Journal of Hydrologic Engineering, 19*(6), 1247–1258. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000898.
Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1262–1273).
Hopcroft, J., Khan, O., Kulis, B., & Selman, B. (2004). Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5249–5253. https://doi.org/10.1073/pnas.0307750100.
Huang, J., Peng, M., Wang, H., Cao, J., Gao, W., & Zhang, X. (2017). A probabilistic method for emerging topic tracking in Microblog stream. *World Wide Web, 20*(2), 325–350. https://doi.org/10.1007/s11280-016-0390-4.
Ivancheva, L. (2008). Scientometrics today: A methodological overview. *COLLNET Journal of Scientometrics and Information Management, 2*(2), 47–56. https://doi.org/10.1080/09737766.2008.10700853.
Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics, 1*(4), 287–307. https://doi.org/10.1016/j.joi.2007.07.004.
Kaushal, S. S., Likens, G. E., Jaworski, N. A., Pace, M. L., Sides, A. M., Seekell, D., et al. (2010). Rising stream and river temperatures in the United States. *Frontiers in Ecology and the Environment, 8*(9), 461–466. https://doi.org/10.1890/090037.
Khalid, M. A., Jijkoun, V., & de Rijke, M. (2008). The impact of named entity normalization on information retrieval for question answering. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. W. White (Eds.), *Advances in information retrieval* (pp. 705–710). Berlin: Springer. https://doi.org/10.1007/978-3-540-78646-7_83.
Kim, M. C., & Chen, C. (2015). A scientometric review of emerging trends and new developments in recommendation systems. *Scientometrics, 104*(1), 239–263. https://doi.org/10.1007/s11192-015-1595-5.

Kirkwood, B. R., & Sterne, J. A. C. (Eds.). (2003). Chapter 31—Analysis of clustered data. In Essential medical statistics, 2nd edn. Malden, Massachusetts: Blackwell Science Ltd.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery, 7*(4), 373–397. https://doi.org/10.1023/A:1024940629314.

Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing and Management, 42*(1), 56–73. https://doi.org/10.1016/j.ipm.2004.11.007.

Kulkarni, S., Singh, A., Ramakrishnan, G., & Chakrabarti, S. (2009). Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 457–466). https://doi.org/10.1145/1557019.1557073.

Lang, R., Mintz, M., Krentz, H. B., & Gill, M. J. (2018). An approach to conference selection and evaluation: Advice to avoid "predatory" conferences. *Scientometrics, 118,* 687–698.

Lee, H., & Kang, P. (2018). Identifying core topics in technology and innovation management studies: A topic model approach. *The Journal of Technology Transfer, 43*(5), 1291–1317.

Liu, X., Jiang, T., & Ma, F. (2013). Collective dynamics in knowledge networks: Emerging trends analysis. *Journal of Informetrics, 7*(2), 425–438. https://doi.org/10.1016/j.joi.2013.01.003.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica: Journal of the Econometric Society*, *13*(3), 245–259.

Marrone, M., & Hammerle, M. (2017). Relevant research areas in IT service management: An examination of academic and practitioner literatures. *Communications of the Association for Information Systems, 41,* 23.

McDonald, R. J., Neff, K. L., Rethlefsen, M. L., & Kallmes, D. F. (2010). Effects of author contribution disclosures and numeric limitations on authorship trends. *Mayo Clinic Proceedings, 85,* 920–927.

McDonald, R. J., Schwartz, K. M., Eckel, L. J., Diehn, F. E., Hunt, C. H., Bartholmai, B. J., et al. (2015). The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic Radiology, 22*(9), 1191–1198. https://doi.org/10.1016/j.acra.2015.05.007.

Michelson, M., & Macskassy, S. A. (2010). Discovering users' topics of interest on Twitter: A first look. In *Proceedings of the fourth workshop on analytics for noisy unstructured text data* (pp. 73–80). https://doi.org/10.1145/1871840.1871852.

Moselen, C., & Wang, L. (2014). Integrating information literacy into academic curricula: A professional development programme for librarians at the University of Auckland. *The Journal of Academic Librarianship, 40*(2), 116–123.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes, 30*(1), 3–26.

Nederhof, A., & Van Wijk, E. (1997). Mapping the social and behavioral sciences world-wide: Use of maps in portfolio analysis of national research efforts. *Scientometrics, 40*(2), 237–276.

Perlin, M. S., Imasato, T., & Borenstein, D. (2018). Is predatory publishing a real threat? Evidence from a large database study. *Scientometrics, 116*(1), 255–273.

Phillips, M., Van Epps, A., Johnson, N., & Zwicky, D. (2018). Effective engineering information literacy instruction: A systematic literature review. *The Journal of Academic Librarianship, 44*(6), 705–711.

Price, D. J. D. S. (1965). Networks of Scientific Papers. *Science, 149*(3683), 510–515. **Retrieved from JSTOR**.

Rossetto, D. E., Bernardes, R. C., Borini, F. M., & Gattaz, C. C. (2018). Structure and evolution of innovation research in the last 60 years: Review and future trends in the field of business through the citations and co-citations analysis. *Scientometrics, 115*(3), 1329–1363.

Salmerón, L., Macedo-Rouet, M., & Rouet, J.-F. (2016). Multiple viewpoints increase students' attention to source features in social question and answer forum messages. *Journal of the Association for Information Science and Technology, 67*(10), 2404–2419.

Santa Soriano, A., Álvarez, C. L., & Valdés, R. M. T. (2018). Bibliometric analysis to identify an emerging research area: Public Relations Intelligence—a challenge to strengthen technological observatories in the network society. *Scientometrics, 115*(3), 1591–1614.

Schmidt, L., & English, M. (2015). Copyright instruction in LIS programs: Report of a survey of standards in the USA. *The Journal of Academic Librarianship, 41*(6), 736–743.

Schober, A., Kittel, C., Baumgartner, R. J., & Füllsack, M. (2018). Identifying dominant topics appearing in the Journal of Cleaner Production. *Journal of Cleaner Production, 190,* 160–168. https://doi.org/10.1016/j.jclepro.2018.04.124.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association, 63*(324), 1379–1389.

Sharp, W. C., Bertelsen, R. D., & Leeworthy, V. R. (2005). Long-term trends in the recreational lobster fishery of Florida, United States: Landings, effort and implications for management. *New Zealand Journal of Marine and Freshwater Research*, *39*(3). Retrieved February 1, 2019, from https://www-tandfonline-com.simsrad.net.ocs.mq.edu.au/doi/pdf/10.1080/00288330.2005.9517349.

Sinnasamy, J., & Karim, N. H. A. (2014). A correlational study of foreign language anxiety and library anxiety among non-native speakers of English: A case study in a Malaysian Public University. *The Journal of Academic Librarianship, 40*(5), 431–435.
Small, H. (1973). 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. https://doi.org/10.1002/asi.4630240406.
Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics, 68*(3), 595–610. https://doi.org/10.1007/s11192-006-0132-y.
Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450–1467.
Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies, 4*(1), 17–40. https://doi.org/10.1177/030631277400400102.
Small, H., & Upham, P. (2008). Citation structure of an emerging research area on the verge of application. *Scientometrics, 79*(2), 365–375. https://doi.org/10.1007/s11192-009-0424-0.
Soria, K. M., Fransen, J., & Nackerud, S. (2014). Stacks, serials, search engines, and students' success: First-year undergraduate students' library use, academic achievement, and retention. *The Journal of Academic Librarianship, 40*(1), 84–91.
Stojković, M., Ilić, A., Prohaska, S., & Plavšić, J. (2014). Multi-temporal analysis of mean annual and seasonal stream flow trends, including periodicity and multiple non-linear regression. *Water Resources Management, 28*(12), 4319–4335. https://doi.org/10.1007/s11269-014-0753-5.
Sun, J., Sheng, D., Gu, D., Du, J. T., & Min, C. (2017). Understanding link sharing tools continuance behavior in social media. *Online Information Review, 41*(1), 119–133.
Thomas, S. W., Adams, B., Hassan, A. E., & Blostein, D. (2014). Studying software evolution using topic models. *Science of Computer Programming, 80,* 457–479. https://doi.org/10.1016/j.scico.2012.08.003.
Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent Named Entity Recognition. In *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003* (Vol. 4, pp. 142–147). https://doi.org/10.3115/1119176.1119195.
Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., et al. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Journal of Web Semantics, 4*(1), 14–28. https://doi.org/10.1016/j.websem.2005.10.002.
van Eck, N., Waltman, L., Noyons, E., & Buter, R. (2010). Automatic term identification for bibliometric mapping. *Scientometrics, 82*(3), 581–596. https://doi.org/10.1007/s11192-010-0173-0.
Wallace, M. L., Gingras, Y., & Duhon, R. (2009). A new approach for detecting scientific specialties from raw cocitation networks. *Journal of the American Society for Information Science and Technology, 60*(2), 240–246. https://doi.org/10.1002/asi.20987.
Walshe, K. (2009). Pseudoinnovation: The development and spread of healthcare quality improvement methodologies. *International Journal for Quality in Health Care, 21*(3), 153–159. https://doi.org/10.1093/intqhc/mzp012.
Wang, Y., Agichtein, E., & Benzi, M. (2012). TM-LDA: Efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 123–131). https://doi.org/10.1145/2339530.2339552.
Westgate, M. J., Barton, P. S., Pierson, J. C., & Lindenmayer, D. B. (2015). Text analysis tools for identification of emerging topics and research gaps in conservation science. *Conservation Biology, 29*(6), 1606–1614. https://doi.org/10.1111/cobi.12605.
White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163–171. https://doi.org/10.1002/asi.4630320302.
Young, S., & Maley, M. (2018). Using practitioner-engaged evidence synthesis to teach research and information literacy skills: A model and case study. *The Journal of Academic Librarianship, 44*(2), 231–237.
Yue, S., Pilon, P., Phinney, B., & Cavadias, G. (2002). The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrological Processes, 16*(9), 1807–1829.
Zhang, S., & Lu, X. X. (2009). Hydrological responses to precipitation variation and diverse human activities in a mountainous tributary of the lower Xijiang, China. *Catena, 77*(2), 130–142.
Zhou, P., & Lv, X. (2015). Academic publishing and collaboration between China and Germany in physics. *Scientometrics, 105*(3), 1875–1887.
Zhou, H., Yu, H., Hu, R., & Hu, J. (2017). A survey on trends of cross-media topic evolution map. *Knowledge-Based Systems, 124,* 164–175. https://doi.org/10.1016/j.knosys.2017.03.009.