Scientometrics (2015) 103:413–433
DOI 10.1007/s11192-015-1554-1

A decade of research in statistics: a topic model approach

Francesca De Battisti · Alfio Ferrara · Silvia Salini

Received: 19 May 2014 / Published online: 12 March 2015

Abstract Topic models are a well known clustering approach for textual data, which provides promising applications in the bibliometric context for the purpose of discovering scientific topics and trends in a corpus of scientific publications. However, topic models per se provide poorly descriptive metadata featuring the discovered clusters of publications and they are not related to the other important metadata usually available with publications, such as authors affiliation, publication venue, and publication year. In this paper, we propose a methodological approach to topic modeling and post-processing of topic models results to the end of describing in depth a field of research over time. In particular, we work on a selection of publications from the international statistical literature, we propose an approach that allows us to identify sophisticated topic descriptors, and we analyze the links between topics and their temporal evolution.

Keywords Probabilistic topic models · Scientometrics · Clustering · Text mining

Introduction

The statistical literature has remarkably changed in recent years. Many authors studied this evolution, with different approaches. For example, (Genest 1997, 1999) analyzed respectively sixteen international journals publishing statistical theories during the period 1985–1995 and eighteen international journals, half of which are specialized in probability theory and the other half in statistics during the period 1986–1995. Papers, authors, and

F. De Battisti · S. Salini (✉)
DEMM, Università degli Studi di Milano, Milan, Italy
e-mail: silvia.salini@unimi.it

F. De Battisti
e-mail: francesca.debattisti@unimi.it

A. Ferrara
DI, Università degli Studi di Milano, Milan, Italy
e-mail: alfio.ferrara@unimi.it

adjusted page counts yield measures of productivity for institutions and countries that contributed to fundamental research in statistics and probability during that period. Genest (2002) updated the previous works on world research output in probability and statistics collecting data until 2000. The data provide valuable information on the evolution of publication habits, in terms of the volume of research, the length of papers, co-authorship practices.

Schell (2010) suggested that dissemination of ideas from theory to practice is a significant challenge in statistics; therefore quick identification of articles useful to practitioners would greatly assist in this dissemination, thereby improving science. For this purpose, he studied and used the citation count history of articles to identify key papers for applied biostatisticians that appeared between 1985 and 1992 in 12 statistics journals. Stigler (1994) studied the use of citation data to investigate the role that statistics journals play in communication within that field and between statistics and other fields. The study looks at citations as import-export statistics reflecting intellectual influence. Ryan and Woodall (2005) attempted to identify the 25 most-cited statistical papers, providing some brief commentary on each paper on their list. This list consists, to a great extent, of papers that are on non-parametric methods, have applications in the life sciences, or deal with the multiple comparisons problem. They also briefly discussed some of the issues involved in the use of citation counts.

In this paper we investigate a decade of research in statistics with a topic model approach. Through the topic model algorithms, applied to our data, with a suitably preliminary cleaning, we identify the most relevant topics in statistical literature between 2000 and 2010, obviously according to the three journals considered, and we describe them, associating keywords and publication venue, authors affiliation countries, and year. We also study the citation distribution by topic. Lastly, we show the topic evolution, an innovative approach to investigate the issue, and the mutual relations among them.

Specifically, we have implemented two different approaches: (1) in order to study topic nowadays, we have identified topics by working on the whole corpus of papers; (2) in order to study topic evolution, we have analyzed topics that can be found by taking into account only the papers produced year by year; in the latter case we were dealing with 11 different corpora and we have generated a set of topics for each corpus independently.

In addition, in this context, we will analyze a posteriori the topic characteristics assigning to them bibliometric indices, based on citations as well as the typical descriptors of the papers that compose them. Thus, within the same subject category, it is possible to identify topics with a different impact.

The paper is organized as follows: in “The corpus of publications data” section , we describe the collected data. “Probabilistic topic models as bibliographic descriptors” section is devoted to describe probabilistic topic models. In “Current topics” and “Topic evolution” sections we present the results obtained with the “topic nowadays” and with the “topic evolution” approaches, respectively. In “Conclusions” section we give our concluding remarks.

The corpus of publications data

In this work, we consider papers published between 2000 and 2010 in one of the following journals:

- *The Annals of Statistics* (Ann. Stat.)
- *Journal of the American Statistical Association* (JASA)

– *Journal of the Royal Statistical Society. Series B* (JRSS(B))

Data were collected in June 2013 and stored in our bibliometric database, which is designed according to the model presented in Ferrara and Salini (2012). Since our goal is to identify methodological innovations in statistics that have marked the history of statistical research in the last years, we have chosen three journals that are universally known to publish contributions of methodological and foundational innovation in statistics. The three journals are all in the top decile of Statistics & Probability according to the various bibliometric indices (IF, 5year-IF, Eigenfactor score, etc.). Moreover they have a long tradition of publishing works that are at the leading edge of methodological development, with a strong emphasis on relevance to statistical practice. For these three journals, we create a reference corpus which is used for all the analysis activities that are described in the rest of the paper. The corpus is created by collecting, for each paper in the journals, the metadata concerning authors and their countries of affiliation, title and abstract, year of publication, and the number of citations received by the paper. The paper titles and abstracts, in particular, will be used in order to extract relevant keywords and linguistic information from the corpus papers.

In Table 1, we show the distribution of papers and citations per year of the three selected journals resulting from Web of Science. It can be noted that JASA is the journal with the highest number of paper published. The number of papers published in time increases for *The Annals of Statistics* but remains constant for the other two journals. We note that in the table the citations refer to papers and not to authors. For example, 3536 represents the number of citations in June 2013 from Web of Science to papers published in *The Annals of Statistics* in the year 2000. Obviously the average number of citations is decreasing with time. Indeed, the citations for more recent papers are lower than for older papers, as we expect.

In Table 2, we show the distribution of papers by country of affiliation and journal. In this case, if two authors of the same paper come from two different countries, the citations are counted two times. Therefore the total number of citations of the previous table is lower than the total number of citations in this table. An author from the United States of

Table 1 Distribution of papers per year and journal

Year	Ann. Stat.			JASA			JRSS(B)		
	Papers	Cit.	Mean	Papers	Cit.	Mean	Papers	Cit.	Mean
2000	72	3536	49.11	136	4509	33.15	52	2558	49.19
2001	67	4134	61.70	116	6577	56.70	48	2655	55.31
2002	68	1879	27.63	120	5210	43.42	52	5928	114.00
2003	84	2155	25.65	99	2570	25.96	56	1740	31.07
2004	99	4309	43.53	107	4131	38.61	57	1973	34.61
2005	98	2091	21.34	116	3378	29.12	41	1907	46.51
2006	109	2274	20.86	137	3658	26.70	42	1616	38.48
2007	112	2183	19.49	141	2323	16.48	45	1122	24.93
2008	107	2070	19.35	166	2033	12.25	51	1080	21.18
2009	147	2112	14.37	168	1172	6.98	47	639	13.60
2010	126	920	7.30	149	649	4.36	25	321	12.84
Total	1089	27663	25.40	1455	36210	24.89	516	21539	41.74

Table 2 Distribution of papers per country and journal

Ann. Stat.		JASA		JRSS(B)	
Country	Papers	Country	Papers	Country	Papers
United States	913	United States	2158	United States	446
France	174	United Kingdom	150	United Kingdom	219
Germany	110	Canada	100	Australia	52
Canada	88	China	64	Canada	38
United Kingdom	75	Germany	59	Germany	36
Netherlands	68	Australia	47	France	27
Australia	62	France	45	Netherlands	26
Israel	40	Spain	43	Italy	24
China	34	Hong Kong	40	China	22
Hong Kong	33	Singapore	35	Switzerland	22
Switzerland	31	Italy	34	Belgium	21
Belgium	30	Belgium	32	Taiwan	20
Italy	30	Switzerland	32	Denmark	18
Spain	22	Taiwan	31	Spain	18
South Korea	19	Netherlands	25	Norway	17
Denmark	18	South Korea	18	Japan	13
Japan	17	Israel	17	Hong Kong	12
Singapore	16	Norway	12	Singapore	11
Taiwan	14	Finland	10	Finland	8
India	12	Austria	8	Israel	7

America is present in at least one paper in two in case of JASA and The Annals of Statistics. On JRSS (B), even if the United States authors are the most numerous, there is a significant presence of British authors with respect to the other countries.

Table 3 shows the papers that are cited more than 500 times. The most attractive concepts and/or techniques, according to the three considered journals and their Editors' decisions, seem to be *boosting, false discovery rate, clustering, microarray* and *gene expression data.*

A typical intuition looking at a collection of papers is that papers can be grouped into "topics". The latter could then be described and connected to each other. Moreover, a topic should generate other topics over the years. In the next sections we will develop these ideas.

Probabilistic topic models as bibliographic descriptors

Topic models are based on the idea that documents are a combination of topics, where a topic is defined as a probability distribution over words. Documents are observed, while topics (and their distributions) are considered as hidden structures or latent variables. Topic modeling algorithms are statistical methods that analyze the words of the original documents to discover the topics that run through them, how these topics are connected to each other, and how they change over time (Blei 2012). The simplest and most commonly used

Table 3 Papers with more than 500 citations

Id	Title	Year	Journal	Authors	Citations
3994	Bayesian measures of model complexity and fit	2002	JRSS(B)	Carlin B.P.; Spiegelhalter D.J.; Best N.G.; Van Der Linde A.	2438
6378	Least angle regression	2004	Ann. Stat.	Efron B.; Hastie T.; Johnstone I.; Tibshirani R.	1598
6675	Additive logistic regression: a statistical view of boosting	2000	Ann. Stat.	Tibshirani R.; Hastie T.; Friedman J.	1559
4015	A direct approach to false discovery rates	2002	JRSS(B)	Storey J.D.	1526
6578	The control of the false discovery rate in multiple testing under dependency	2001	Ann. Stat.	Benjamini Y.; Yekutieli D.	1403
5323	Comparison of discrimination methods for the classification of tumors using gene expression data	2002	JASA	Speed T.P.; Dudoit S.; Fridlyand J.	1050
3859	Regularization and variable selection via the elastic net	2005	JRSS(B)	Zou H.; Hastie T.	922
6561	Greedy function approximation: a gradient boosting machine	2001	Ann. Stat.	Friedman J.H.	808
5349	Variable selection via nonconcave penalized likelihood and its oracle properties	2001	JASA	Li R.; Fan J.	776
5020	A model-based background adjustment for oligonucleotide expression arrays	2004	JASA	Wu Z.; Irizarry R.A.; Gentleman R.; Martinez-Murillo F.; Spencer F.	736
5295	Model-based clustering, discriminant analysis, and density estimation	2002	JASA	Raftery A.E.; Fraley C.	733
5365	Empirical Bayes analysis of a microarray experiment	2001	JASA	Tibshirani R.; Storey J.D.; Efron B.; Tusher V.	678
3822	Model selection and estimation in regression with grouped variables	2006	JRSS(B)	Yuan M.; Lin Y.	643
4779	The adaptive lasso and its oracle properties	2006	JASA	Zou H.	585
5994	The Dantzig selector: statistical estimation when p is much larger than n	2007	Ann. Stat.	Tao T.; Candes E.	538
4070	Estimating the number of clusters in a data set via the gap statistic	2001	JRSS(B)	Tibshirani R.; Hastie T.; Walther G.	526
4764	Hierarchical Dirichlet processes	2006	JASA	Teh Y.W.; Jordan M.I.; Beal M.J.; Blei D.M.	510
6405	The positive false discovery rate: a Bayesian interpretation and the q-value	2003	Ann. Stat.	Storey J.D.	506

probabilistic topic approach to document modeling is the generative model Latent Dirichlet Allocation (LDA) (Blei et al. 2003). The idea behind LDA is that documents blend multiple topics.

A topic is defined to be a distribution over a fixed vocabulary. For example the statistics topic has words about statistics with high probability. The model assumes that the topics are generated before the documents. For each document, the words are generated in a two-stage process: (1) randomly choose a distribution over topics (Dirichlet distribution); (2) for each word first randomly choose a topic from the distribution over topics and then randomly choose a word from the corresponding distribution over the vocabulary. The central problem for topic modeling is the use of the observed documents to infer the latent variables. Topic models are probabilistic models in which data are treated as arising from a generative process that includes hidden (or latent) variables. This process defines a joint probability distribution over both the observed and hidden random variables.

The conditional distribution of the hidden variables given the observed variables, also called posterior distribution, is computed. The numerator of the conditional distribution is the joint distribution of all the random variables, which can be easily computed; the denominator is the marginal probability of the observations, or the probability of seeing the observed corpus under any topic model. Theoretically, it can be computed by summing the joint distribution over every possible instantiation of the hidden topic structure; practically, because the number of possible topic structures is exponentially large, this sum is difficult to compute.

Topic modeling algorithms fall into two categories, which propose different alternative distributions to approximate the true posterior: sampling-based algorithms, as Gibbs sampling, and variational algorithms. The first group considers a Markov chain, a sequence of random variables, each dependent of the previous, whose limiting distribution is the posterior (Steyvers and Griffiths 2007); the second group of algorithms, instead, represents a deterministic alternative to sampling-based algorithms (VEM). Rather than approximating the posterior with samples, variational methods posit a parameterized family of distributions over the hidden structure and find the member of the family that is closest to the posterior; in this way, they transform the inference problem into an optimization problem. In 2007, a correlated topic model (CTM) was proposed, which explicitly models the correlation between the latent topics in the documents (Blei and Lafferty 2007). In this paper, we rely on VEM algorithms instead of CTM because VEM algorithms provide more than a topic explanation for each paper. This overlapping between topics is important to find topic correlations, which is one of the main goals of our work.

Choosing the number of topics

As discussed before, topic models are latent variable models of documents that exploit the correlations among the words and latent semantic themes in a collection of papers (Blei and Lafferty 2007). An important consequence of this definition is that the expected number of topics (i.e., the latent variables) is supposed to be set before the computation of the model itself. Thus, since the number of topics has to be set a priori, choosing the best number of topics for a given collection of papers is not trivial. In the literature (Hall et al. 2008; Blei 2012) this problem has been addressed in different ways, but always looking for a compromise between the need for a high number of topics to cover all the themes in the document collection and the need for a limited number of topics, which can be more easily understood and verified by experts in the domain of data collected.

In order to help in choosing the number of topics of interest, a measure of *perplexity* has been introduced (Grün and Hornik 2011). The idea is that model selection with respect to the number of topics is possible by splitting the data into training and test datasets. The likelihood for the test data is then approximated using the lower bound for VEM estimation. In particular, perplexity is a measure of the ability of a model to generalize to unseen data. It is defined as the reciprocal geometric mean of the likelihood of a test corpus given the model, as follows:

$$\text{Perplexity}(w) = \exp\left\{ -\frac{\log(p(w))}{\sum_{d=1}^{D} \sum_{j=1}^{V} n^{(jd)}} \right\}$$

where $n^{(jd)}$ denotes how often the jth term occurred in the dth document.

The common method to evaluate perplexity in topic models is to hold out test data from the corpus to be trained and then test the estimated model on the held-out data. Higher values of perplexity indicate a higher misrepresentation of the words of the test documents by the trained topics. Perplexity is a measure of the quality of the model learned by LDA in predicting future data from the same distribution as the data used to train the model. In doing so, it measures an interesting characteristic of an inference algorithm: given that the model is the same, the best algorithm (in terms of quality of the learned result) will have better perplexity than the others. Perplexity is usually the first or second metric used to judge statistical model quality (other popular methods being test-set likelihood or even marginal probability of the data given the model), but it is too coarse, hence recently the topic modeling community has been moving towards more accurate metrics (Mimno and Blei 2011). Even though these more refined metrics carry a lot more weight and show you all sorts of interesting information, bear in mind that test-set perplexity is probably correlated with all of them. In this paper, we run a set of pre-tests on the whole collection of paper at hand, by executing the VEM algorithm with a variable number of topics, ranging from 10 to 400, and by collecting the perplexity values of each execution. The resulting perplexity plot is shown in Fig. 1.

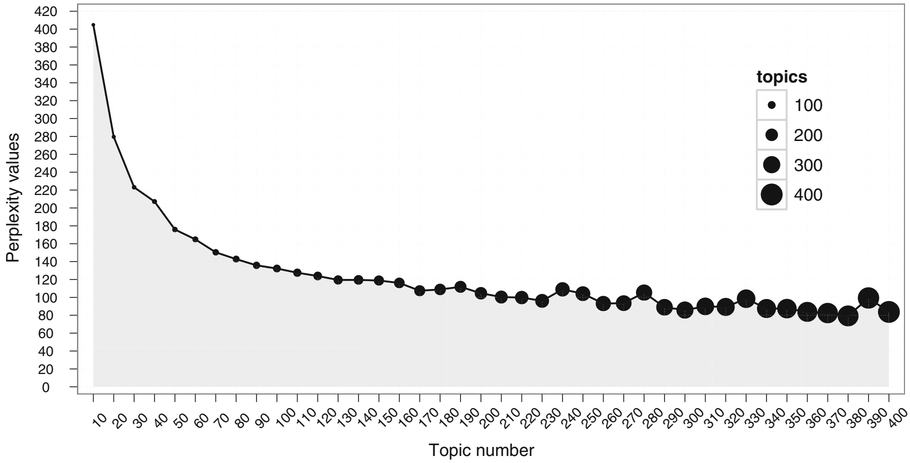

Fig. 1 Perplexity plot for a number of topics ranging from 10 to 400

Looking at the perplexity plot, we can see how the perplexity dramatically decreases from a level of 400–200, just moving from 10 to 30 topics. Then, the decrement continues reaching a more or less stable value of 100/120 with 120 or more topics. Since we are interested in having low levels of perplexity but also in keeping as low as possible the numbers of topics, we decided to limit ourselves to 120 topics.

We briefly recall that each topic is associated as an explanation with each paper together with a measure of relevance (average quality) of that paper for the topic. Thus, a first measure of the ability of topics to explain papers is the average level of paper relevance per topic. Moreover, we are also interested in counting the number of papers explained by each topic. The idea is that a solution where topics explain more papers is better than a solution where topics are capable of explaining less papers, in that the former one provides a better synthesis of the paper corpus. But of course, the two situations are comparable only if the average level of relevance per topic is the same or similar. A relevant issue working with topic models is to determine when a topic has to be considered as a good explanation for a paper at hand. In particular, given a topic T, a paper p, and the relevance $\rho(p, T)$ of T with respect to p, we say that p is *explained* by T if $\rho(p, T) \geq th_e$.[1] Therefore, we call th_e *explanation threshold.* In order to determine the explanation threshold th_e, we start from two main requirements: (1) *corpus coverage:* we are interested in finding a threshold value such that the fraction of papers in the corpus that are explained by at least one topic is high; (2) *explanation quality:* we are interested in finding a threshold value such that the average quality (i.e., relevance) of topic explanations is high. More formally, the corpus coverage $C_P^{th_e}$ for a corpus of papers P and an explanation threshold th_e is defined as:

$$C_P^{th_e} = \frac{\mid \{p_i \mid p_i \in P \wedge \rho(p_i, T) \geq th_e\} \mid}{\mid P \mid}$$

where $\mid \{p_i \mid p_i \in P \wedge \rho(p_i, T) \geq th_e\} \mid$ is the cardinality of the set of corpus papers which are explained by at least one topic with a value of relevance higher than, or equal to, a given explanation threshold th_e. The explanation quality $Q_P^{th_e}$ is defined as:

$$Q_P^{th_e} = \frac{\sum_{i=1}^{|P|} \rho(p_i, T) \mid p_i \in P \wedge \rho(p_i, T) \geq th_e}{K}$$

where $K = \mid \{p_i \mid p_i \in P \wedge \rho(p_i, T) \geq th_e\} \mid$ is the number of papers having at least one explanation topic T such that $\rho(p_i, T) \geq th_e$. We can observe that, as expected, when the corpus coverage increases, the explanation quality decreases. In fact, topics with high quality explanations are more focused on a limited number of papers. On the opposite, when we set a low level of the explanation threshold, the number of papers that are explained by at least a topic is higher, but the average quality of accepted explanations is lower. This situation is illustrated in Fig. 2, where we report the values of corpus coverage and explanation quality for different levels of the explanation threshold th_e when we consider 120 topics.

By taking into account the results shown in Fig. 2, we use as value of the explanation threshold the point where the distance between corpus coverage and explanation quality is minimal, that is $th_e = 0.4$. Thus, working on 120 topics and using an explanation threshold

[1] In our approach, the relevance $\rho(p, T)$ of a topic T for a paper p is the log-likelihood returned for each paper by the VEM algorithm implementation as provided in the R `topicmodels` package (Grün and Hornik 2011).

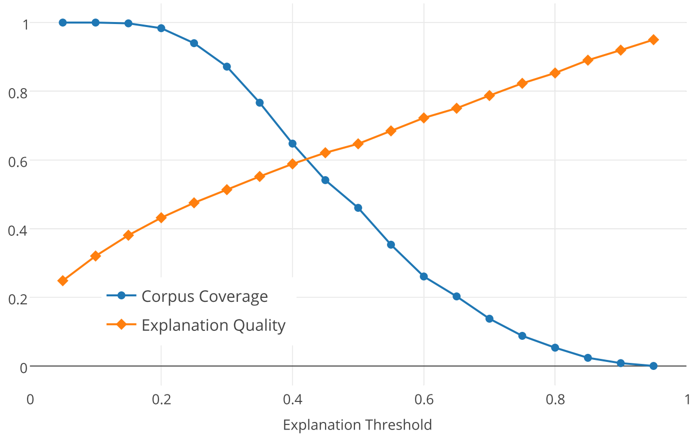

Fig. 2 Corpus coverage and explanation quality at different values of explanation threshold

of 0.4, the coverage of corpus is of 1974 papers out of 3048 (64.76 %).[2] The average explanation quality is 0.59.

Current topics

As reported in the previous section, in our first experiment with topic models we calculate 120 topics by working on the whole corpus of papers considering for each paper title and abstract. Then, we select the top-30 prominent topics (i.e., topics explaining the highest number of papers), in order to have a limited and easily understandable collection of topics which explain the paper corpus with a good level of relevance. The complete list of 30 topics[3] that have been presented in this paper is reported in Table 4, where we anticipate a selection of keywords that we extract from each topic in order to help the reader in understanding the contents of the topics. The procedure of keyword selection, that is also capable of extracting compound keywords, is detailed in "Finding the most relevant topic keywords" subsection.

Topic description

One of the most interesting uses of topics is to provide a synthetic map of the corpus of papers that have been collected, in order, in general, to give a high level view of the most relevant concepts, keywords, and themes that have been addressed in a time period for a specific field of research. In our case, limited to the three journals considered, we identify the relevant themes in the last 10 year in statistics. In order to obtain this result, we first need to associate each topic with a descriptor providing information about the most relevant keywords describing the topic and some useful statistical information concerning the publication venue distribution of papers per topic (useful to understand if there are

[2] We note that the total number of papers is lower than the total number of papers in the corpus, because we excluded from the topic analysis those papers with incomplete metadata as well as those containing editorial material but not a proper scientific contribution.

[3] T960 with 124 papers was not considered here because it groups Discussion papers. Similarly T906 with 20 papers was not considered here because it groups all the Erratum papers.

Table 4 Top-30 topics per number of papers

ID	# of papers	Most rel. keywords
929	43	False, discovery, false discoveries
938	41	Clustered, models, data
875	36	Spatial, models, processes
958	33	Bootstrap, estimator, model
936	28	Designs, optimal, optimal designs
864	27	Quantile, quantile regression, estimation
918	26	Regressions, dimension, dimension reduction
941	26	Wavelet, estimation, thresholding
893	25	Models, estimator, penalized
862	23	Spectral, periodogram, spectral density
869	23	Volatility, jump, discretized
881	23	Spline, models, regression
891	23	Extremes, extreme value, value
909	23	Memory, long, long memory
911	23	Diffusions, estimators, deformable
925	22	Hazards, model, hazards models
897	21	Stepup, procedure, controlling
922	21	Rankings, signed, models
930	21	Filters, particle filtering, model
943	21	Recurrent, event, recurrent events
952	21	Graphs, models, tree
955	21	Bandwidths, bandwidth selection, regression
861	20	Design, aberrant, fractional factorial
903	20	Sequential, sequential analysis, design
927	20	Depth, projection, multivariate
950	20	Regression, estimates, robust
873	19	Optimal designs, designs, statistical
879	19	Disclosure, frailty, model
923	19	Deconvolution, models, wavelet
935	19	Brownian motion, shape constraints, density estimation

journals that are more specialized in some topics) and a distribution of the topic papers per year, country, and number of citations.

More formally, we can define a descriptor $\mathcal{D}_T$ of a topic T as a five-tuple of the form $\mathcal{D}_T = \langle K_T, J_T, Y_T, C_T, R_T \rangle$, where $K_T = \langle (k_1, r_1), (k_2, r_2), \ldots, (k_j, r_j) \rangle$ denotes a list of the j most relevant keywords describing T together with the relevance r_i of the keyword k_i for T; $J_T = \langle (j_1, n_1), (j_2, n_2), \ldots, (j_k, n_k) \rangle$ denotes the list of the k most relevant journals for T together with the fraction n_i of papers explained by T that have been published by the journal j_i; $Y_T = \langle (y_1, n_1), (y_2, n_2), \ldots, (y_l, n_l) \rangle$ denotes the list of the l most relevant years for T together with the fraction n_i of papers explained by T that have been published in y_i; $C_T = \langle (c_1, n_1), (c_2, n_2), \ldots, (c_m, n_m) \rangle$ denotes the list of the m most relevant countries for T together with the fraction n_i of authors of papers explained by T that have been affiliated in an institution located in the country c_i; finally R_T denotes the distribution of citations per papers that have been explained by T.

Finding the most relevant topic keywords

In order to determine the keyword descriptor K_T for the topic T, we first select the set $P_T = \{p_i \mid \rho(p, T) \geq th_e\}$ of papers explained by the topic T, that is the set of papers with relevance $\rho(p, T)$ higher than, or equal to, the explanation threshold with respect to the topic T. In such a way, we create a textual corpus that is then pre-processed by using some standard natural language processing techniques (NLP) in order to create a vector of terms for each paper. The NLP techniques used are elision removal (*log-likelihood* → *log likelihood*), lower case normalization (*New York* → *new york*), and stop words removal (*the next step* → *next step*). During the pre-processing step, we decided not to use some other common techniques like stemming, in order to provide a more human readable descriptor. Instead, we introduced bi-gram research. In our approach, a bi-gram is a pair of terms that often occur together in the corpus and, thus, can be interpreted as a single compound term (e.g., *false discovery*). In order to determine relevant bi-grams, we associate a measure of *mutual information* $I(t_i, t_j)$ with any pair of adjacent terms t_i and t_j. $I(t_i, t_j)$ is defined as[4]:

$$I(t_i, t_j) = \log\left(\frac{p(t_i, t_j)}{p(t_i) \cdot p(t_j)}\right)$$

where $p(t_i, t_j)$ is the ratio between the number of occurrences of the pair (t_i, t_j) to the number of occurrences of all the pairs of adjacent terms in the corpus; $p(t_i)$ is the ratio between the number of occurrences of t_i to the total number of occurrences of any term in the corpus. $I(t_i, t_j)$ measures the relevance of the occurrences of the pair (t_i, t_j) with respect to the relevance of the occurrences of the terms t_i and t_j separately. We select the most relevant pairs as those pairs that appear more than twice in the corpus and have a value of mutual information higher than a fixed threshold equal to 1. Then we substitute the relevant pairs to the single terms in the vector. As soon as all the papers explained by the topic T (i.e., papers in P_T) are associated with a vector of terms, we calculate the most relevant keywords describing T as the list of the most relevant vector terms that are the terms with the highest relevance according to a TF-IDF like measure. The term frequency $\mathrm{TF(t_i)}$ of a term t_i is calculated as the number of occurrences of t_i in all the term vectors associated with papers in P_T. The inverse document frequency $\mathrm{IDF(t_i)}$ of a term t_i is calculated as follows:

$$\mathrm{IDF(t_i)} = \log\left(\frac{\mid \mathrm{P} \mid}{1 + \mid \{\mathrm{p} \in \mathrm{P} : \mathrm{t_i} \in \mathrm{p}\} \mid}\right)$$

where $\mid P \mid$ is the total number of papers in the corpus (not only those explained by the topic T), while $\mid \{p \in P : t_i \in p\} \mid$ is the number of paper vectors containing t_i. Finally, the relevance r_i of a keyword k_i, given the corresponding term t_i, is calculated as $r_i = \mathrm{TF(t_i)} \cdot \mathrm{IDF(t_i)}$. In order to select the most relevant keywords, we experimentally observed that we can take the terms that have a cumulative relevance higher than 10 % of the sum of all the terms' relevances.

Finding publication venue, year, country, and citation descriptors

The approach used for determining the other descriptors J_T, Y_T, and C_T is basically the same. We simply count the number of papers in P_T aggregated by the dimension of

[4] In the subsequent formula and in all the other formulae in the rest of the paper, the log symbol refers to the base-10 logarithm.

interest, that can be the journal, the year, or the country. For what concerns countries, we associate a paper with each country of affiliation of its author. Finally, we take the whole list of journals, years and countries together with the fraction of papers associated with each source, year or country over the total number of papers (authors in case of countries) in P_T.

Another descriptor that can be associated with a topic is its impact, based on citations. Since citations are associated to every paper, it is possible to aggregate them to obtain bibliometric measures related to the topic. In addition to the mean and/or the median of the citations, and the classical *h-index* (Hirsch 2005), a graph of the citation distribution in the topic can be produced to highlight if all the papers have a similar number of citations or not. In the first case it would mean that the topic, and not the paper or the author, is able to produce itself a certain level of citations.

Example

In order to provide an example of topic description, in Table 5 we take into account a sample of paper titles that are explained by the topic T929 related to *false discovery rate*.

According to the approach described in the previous sections, we calculate the descriptors of the topic T929 by determining the most relevant keywords and the publication source, year, and country distribution, which are reported in Table 6.

The descriptors can also be used to provide a graphical representation of a topic, according to the following approach. A topic is represented as a circle, whose diameter is proportional to the number of papers explained by the topic at hand (this is useful when more than one topic is depicted in the same map). The topic circle contains the keywords that are printed in a tag-cloud fashion, where the dimension of each keyword is proportional to its relevance for the topic. Then, the other descriptors are associated with the circle as plots reporting the relevance of each element of the descriptor. An example of such a graphical representation is shown in Fig. 3.

The T929 is characterized by the keywords *false discovery, discovery rate, multiple testing*. The three journals addressed this topic, more particularly *The Annals of Statistics* and JASA. The spread has been fluctuating over the years and tended to increase over time, with a peak in 2009. The most represented country is the United States. The citations are

Table 5 Example of papers explained by the T929

Title	Year	Journal	Authors	Cit	Rel
False discovery and false nondiscovery rates in single-step multiple testing procedures	2006	Ann. Stat.	Sarkar S.K.	30	0.92
The positive false discovery rate: A Bayesian interpretation and the q-value	2003	Ann. Stat.	Storey J.D.	506	0.84
Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach	2004	JRSS(B)	Storey J.D.; Taylor J.E.; Siegmund D.	379	0.79
Operating characteristics and extensions of the false discovery rate procedure	2002	JRSS(B)	Wasserman L.; Genovese C.	205	0.76
A direct approach to false discovery rates	2002	JRSS(B)	Storey J.D.	1526	0.75
The control of the false discovery rate in multiple testing under dependency	2001	Ann. Stat.	Benjamini Y.; Yekutieli D.	1403	0.72

Table 6 Descriptors for topic T929

Keywords		Years		Countries	
False	0.49	2010	3	United States	51
Discovery	0.48	2009	8	Israel	8
False discoveries	0.45	2008	5	Germany	7
Discovery rates	0.36	2007	7	United Kingdom	1
Rates	0.29	2006	4	Italy	1
Procedure	0.20	2005	2	Singapore	1
Multiple	0.19	2004	6	France	1
Testing	0.19	2003	2	Austria	1
Control	0.19	2002	5	Switzerland	1
Multiple testing	0.15	2001	1	South Korea	1
…	…	2000	1	…	…

Journals		Citations	
Ann. Stat.	17	Total	5530
JASA	16	Mean	128.6
JRSS(B)	10	Median	20
		H-index	21

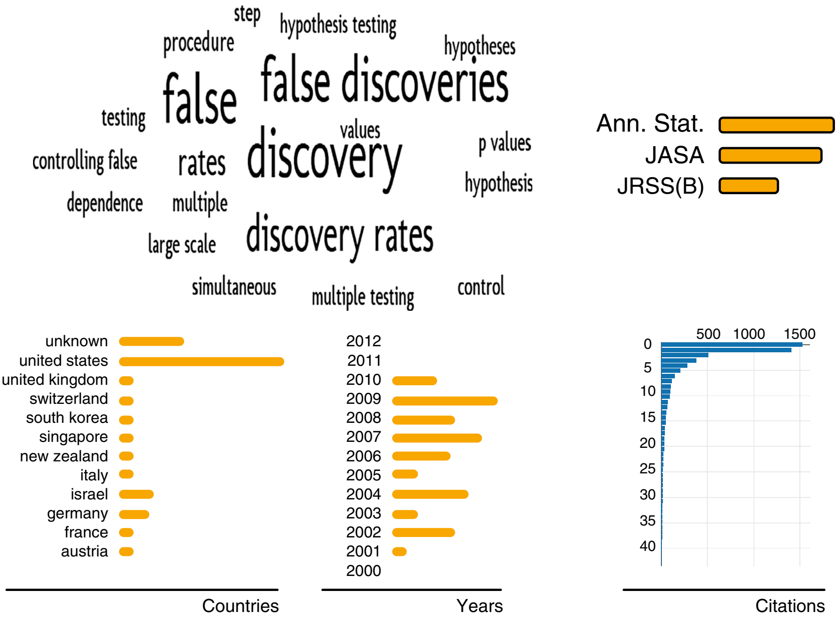

Fig. 3 Graphical representation of topic T929

not distributed evenly among the paper, but, as it can be observed in Table 5, there are papers with a very high number of citations. The median equal to 20 is much lower than the mean which is equal to 128.6. This shape is in accordance with the well known empirical laws governing the distribution of citations. In practice, the number of citations received by scientific papers appears to have a power-law distribution (Newman 2006). The distribution of citations is a rapidly decreasing function of citation count. Zipf plot is well suited for determining the large-x tail of the citation distribution (Gupta et al. 2005). For other topics, another pattern could be observed. In "Citation distribution per topics" subsection a more detailed analysis of citations intra topics will be presented.

Topic mutual relations

An interesting feature of topics is their mutual relations. In particular, we propose a similarity relation $\sigma(T_i, T_j)$ between two topics, which is based on their terminology, as follows. Let the dictionary $\mathcal{D}$ be the set of all the terms used in the corpus, i.e., all the terms appearing in the title or abstract of at least one paper of the entire collection. The topic similarity $\sigma(T_i, T_j)$ between two topics T_i and T_j is based on a weight w_l associated with each term $t_l \in \mathcal{D}$ with respect to T_i and T_j, respectively. In particular, given the topics T_i and T_j, we define two vectors of terms V_i and V_j, which have the following form:

$$V_i = \langle w_1, w_2, \ldots, w_n \rangle, \;\; V_j = \langle w_1, w_2, \ldots, w_m \rangle$$

where n and m are the number of terms extracted from T_i and T_j, respectively. In particular, given a term $t_k \in \mathcal{D}$, its corresponding weight w_{ki} with respect to T_i is equal to 0 if t_k does not appear in T_i (i.e., is not used either in a title or in a abstract of any paper explained by T_i); otherwise, w_{ki} is calculated using the TF-IDF method discussed in "Topic description" subsection. Analogously, we calculate the weight w_{kj} for the topic T_j. On this basis, we evaluate the similarity $\sigma(T_i, T_j)$ between T_i and T_j as the correlation between their corresponding vectors of terms V_i and V_j, as follows:

$$\sigma(T_i, T_j) = \frac{\sum w_{ik} w_{il}}{\sqrt{\sum w_{ik}^2} \sqrt{\sum w_{jk}^2}}$$

where w_{ik} denotes the weight attributed to keyword t_k for topic T_i.

Example

In order to clarify the evaluation of term-based topic similarity, we introduce a very simple example, by taking into account the topics T929 *(false, discovery, false discoveries, discovery rates, rates)* and T878 *(models, endpoints, partitioning, decision, procedures)*, that have the following keyword descriptors:

- **T929:** false (0.49), discovery (0.48), false discoveries (0.45), discovery rates (0.36), rates (0.29), procedure (0.2), multiple (0.19), testing (0.19), control (0.19), multiple testing (0.15), values (0.11), large scale (0.09), controlling false (0.08), simultaneous (0.07), hypothesis (0.07)
- **T878:** models (0.2), endpoints (0.15), partitioning (0.14), decision (0.12), procedures (0.11), testing (0.11), lq (0.11), bayes (0.11), primary (0.1), multiple (0.1), decision theory (0.1), principle (0.09), discovery (0.09), equivalence (0.08), best (0.08)

By considering the two term vectors corresponding to T929 and T878, we retrieve several terms in common (e.g., procedure, testing, discovery) that are used to determine the product vector of common terms, which is equal to 0.1051. The sum of all the weights in the two vectors is equal to 1.0458 for T929 and 0.4515 for T878. This leads to a similarity equal to 0.2225 that is calculated as follows:

$$\sigma(T929, T878) = \frac{0.1051}{1.0458 \cdot 0.4515} = 0.2225$$

In Table 7 the list of topics, with the relative keywords, ordered by descending similarity with T929 is reported.

Topic map

Since a topic is associated with a set of papers, it can be seen as the collection of papers that are explained by the topic at hand according to the explanation threshold th_e. As we have seen, a topic can have a variable degree of similarity to other topics and can be described by the most occurrent terms in the papers therein contained. Moreover, each paper is associated with the paper contributors, usually the authors, with the venue of publication, with the publication year, and with the number of citations received. In order to provide a synthetic and comprehensive view of the topics addressed by a corpus of papers, we introduce the notion of *topic map*. A topic map is a graph where nodes represent topics and edges represent similarity relations between topics. Moreover, each node of the topic map can be graphically represented by a circle containing the most relevant k terms extracted from the papers explained by the corresponding topic. A term's font size is proportional to the number of occurrences of the term in the topic, using conventional tag clouds. The circle area is proportional to the number of papers explained by the topic. Also the graphical disposition of nodes is relevant, since it is determined in order to display similar topics as close as possible one to each other. The edge width is proportional to the strength of the similarity relation. Finally, a node/topic could be also labelled with country, publication source, and year descriptors. As an example of very simple topic map, we show in Fig. 4 a portion of the topic map for the topics extracted around T929 presented in the

Table 7 List of topics ordered by descending similarity with T929

T	σ	Contents
878	0.23	Step-up/down procedure, multiple endpoints, multiple comparisons, finite action problem, Dirichlet process
897	0.23	Multiple testing, step-up/down procedure, family wise error rate, bootstrap, statistical process control
935	0.18	Estimating, distribution, Brownian motion, concave, densities
874	0.10	Oracle inequalities, aggregation, model selection, order-restricted inference, lipoprotein lipase
917	0.10	Horvitz–Thompson estimator, calibration, logistic regression, generalized linear model, population size
908	0.09	Racial profiling, nonstationary random process, SLEX library, autoregressive model, criminology
890	0.09	Statistical process control, average run length, model selection, LISREL, shift function
893	0.09	SCAD,Variable selection, LASSO, resampling, model selection, penalized likelihood
...	...	...

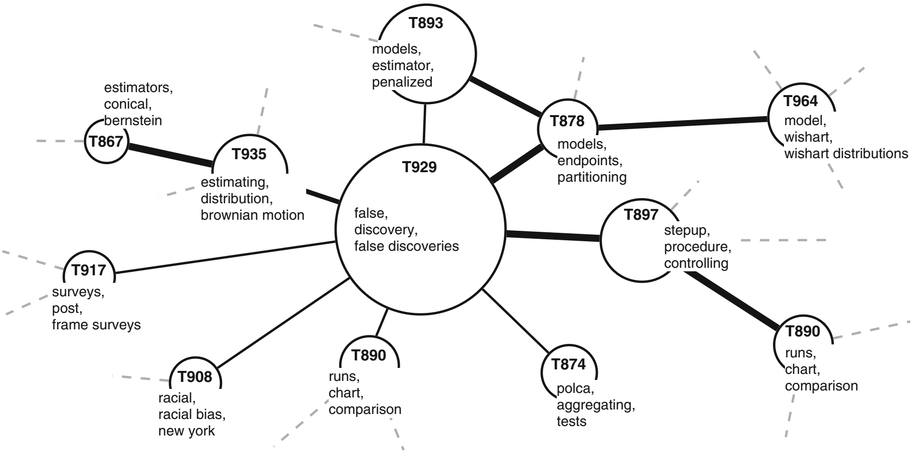

Fig. 4 Topic map representing topics extracted around T929

previous example. In our example, we show in each topic only the topic ID and the three most relevant keywords.

Citation distribution per topics

As just mentioned before, it is not obvious that, within the same subject category, each topic has the same level of citations and then the same benchmark values for bibliometric indicators. Moreover, also the citation distribution could differ from topic to topic. In order to check this hypothesis, we select five topics with different citation patterns. In Table 8, for each topic we show the number of papers, the *h-index*[5], the number of papers with more than 500 citations, the number of papers with more than 100 citations. Moreover for the citations the mean, the standard deviation, the median, the interquartile range (IR) and the Gini coefficient[6] are reported. We can also represent the citation distribution through the Lorenz curve, that is the graphical representation developed by the American economist Max Lorenz in 1905 for the wealth distribution. In our application (see Fig. 5), the horizontal axis is the proportion of papers and the vertical axis is the proportion of citations. A straight diagonal line represents perfect equal distribution of citations per paper; the Lorenz curve lies beneath it, showing the real distribution of the citations. The difference between the straight line and the curved line is the amount of concentration, this area represents the Gini coefficient.

Fox example, T929, related to *false discovery rate* and previously discussed in this section, has a citation mean lower than T878 related to *multiple comparisons*, but its citation median and its h-index are much higher. This means that in T929 the citations are less concentrated than in T878; in fact the Gini coefficient is smaller.

This is evident also looking at the standard deviation and the IR that are higher for topic T878 with respect to T929.

[5] We recall that topic with an index of h includes h papers each of which has been cited in other papers at least h times.

[6] Corrado Gini's concentration index; the value 0 indicates equality or uniform distribution, the value 1 indicates maximum concentration.

Table 8 Citation pattern per topics

Topic	Papers	*h-index*	500+	100+	Mean	Std	Median	IR	Gini
960	124	10	0	0	4	11	0	1	0.88
929	43	21	3	8	129	316	20	70	0.78
967	18	13	0	3	47	48	26	64	0.53
878	15	7	1	1	178	625	7	31	0.89
904	9	7	0	0	10	5	9	6	0.26

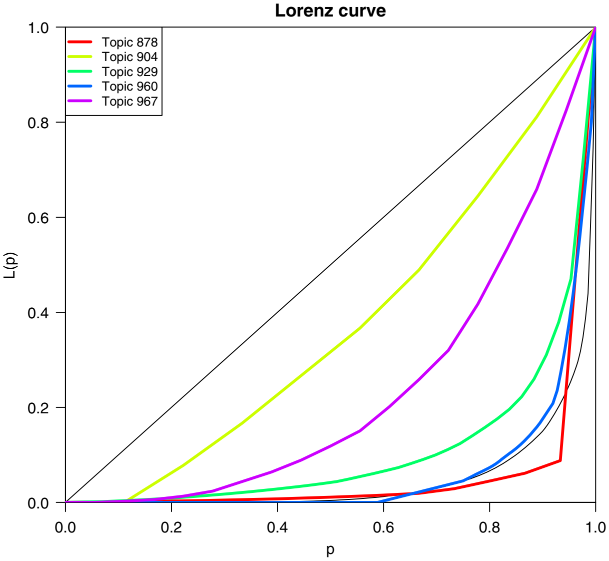

Fig. 5 Lorenz curve of citations per topic

Topic evolution

In order to study the evolution of a scientific field in time, the idea of focusing on the years associated with topics is the most natural approach but it is also affected by a structural problem: in fact, topics are statistically discovered over the whole corpus, which includes papers that have been published in different years. This means that the number of papers published in a given year affects the whole composition of topics and, potentially, may lead to a situation where topics that were popular in years featured by a limited number of publications are not discovered at all. Our idea is that, instead of focusing on the whole corpus of papers, we are now interested in studying the topics that can be found by taking into account only the papers produced year by year. According to this approach, we are now dealing with 11 different corpora (i.e., one per year) and we generate a set of topics for each corpus independently. Then, we study the similarity relations existing between the topics associated with one year and the topics associated with the subsequent year. Our hypothesis is that a similarity relation between a topic $T_{i(y)}$, deriving from the corpus of papers published in the year y, and a topic $T_{j(y+1)}$, derived from the year $y+1$, is a useful index of a possible evolution of the topic T_i into the topic T_j.

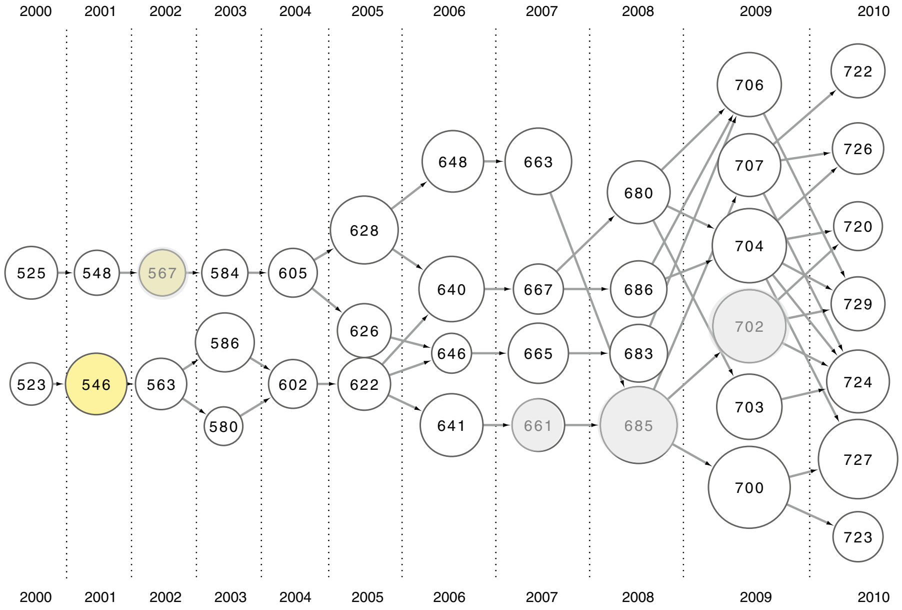

Fig. 6 Topic evolution map

As an example, we discuss the case of two papers:

a. Storey J.D. (2002), *A direct approach to false discovery rates,* JRSS(B)
b. Ronchetti E.; Cantoni E. (2001), *Robust Inference for Generalized Linear Models*, JASA

In Fig. 6, we report a *topic evolution map* in which the topics that contain papers (a and b) are highlighted. The topics here are extracted year by year. In the evolution map, topics, represented as circles, are ordered according to years and are linked one to each other by arrows which represent similarity relations among topics. The similarity has been calculated as discussed in "Current topics" section. Three topics include the false discovery rate (i.e., T567, T661, and T685), and another one includes "multiple testing" (i.e., T702), which is a related broader topic. These have been reported as gray, shadowed circles in the map. The topic chain is created as follows: given a topic $T_{i(y)}$ derived from year y, we calculate the similarity between $T_{i(y)}$ and all the topics $T_{j(y+1)}$ derived from the subsequent year $y+1$; then, we set up a similarity threshold th_s and we create a link between $T_{i(y)}$ and all the topics $T_{j(y+1)}$ such that $\sigma(T_{i(y)}, T_{j(y+1))}) \geq th_s$.

In such a way, the evolution map suggests possible evolution paths connecting topics extracted from papers published in the early period (i.e., 2000–2003) and topics extracted from papers of the late period (i.e., 2008–2010).

Looking at the map, it is possible to understand which topics have chained over the years, leading to the current research topics. For example T640 (*microarray, not parametric, semi-parametric regression, functional data*) is connected to both previous routes that include topics such as *false dicovery rate* (T567), *robust methods* (T546), *functional data analysis* (T628), *robustness issues in multivariate data analysis* (T622). T663 (*risk factors, multicenter survival studies, hazard function*) is connected with the topic T685

Table 9 Most relevant keywords for topics appearing in Fig. 6

ID	# of p.	Most rel. keywords
2000		
523	19	Monotonicity regression, testing monotonicity, sequential testing
525	27	Toxicology, research, functional linear models
2001		
546	34	Robust methods, penalized likelihood, conditional heteroscedastic model
548	21	Serial correlation, coverage, doses
2002		
563	26	Censoring, frailty model, cox regression
567	22	False discovery rate, publication bias, earthquake
2003		
580	16	Differentially expressed, differentiability, paired
586	32	Stochastic optimization, moderate deviations, sequential analysis
584	22	Cusum, leukemia, extremes
2004		
602	24	Hazard, isotropy, shared frailty models
605	24	Prediction error, modes, predictive
2005		
622	27	Robustness multivariate data, random graphs, generalized linear models
626	28	Crossover designs, fused lasso, meta analysis
628	39	Functional data analysis, long memory, observation times
2006		
641	35	Fit tests, testing order, robust estimates generalized
646	17	Causal, causal inference, neighborhood
640	37	Locally stationary processes, semi-parametric regression, functional data
648	34	Binary regression, run, bayesian wavelet
2007		
661	27	Discovery rates, multistage, controlling
665	33	Outcome, discretely sampled, estimation treatment
667	25	Support vector, vector machines, support vector machines
663	38	Risk factors, survival studies, hazard function, comment
2008		
685	46	Randomized experiment, clinical trials, false discovery rate, comment, rejoinder
683	31	Capture recapture, test positives, smoothly clipped
686	30	Lasso, competing risks, data competing
680	35	Clustering, gibbs, gibbs samplers
2009		
700	50	High-dimensional regression, large-scale prediction problems, comment
703	37	Panel count, panel count data, mises distribution
702	43	Multiple testing, testing, testing dependence
704	44	Lasso, high dimensional, large covariance
707	35	Tail index, tails, transformations
706	36	Log linear models, log linear, singular value decompositions
2010		

Table 9 continued

ID	# of p.	Most rel. keywords
723	25	Bootstrap, sided confidence interval, smallest one
727	48	Missing at random, variable selection, comment, letters editor, rejoinder
724	35	Tests high, test, association
729	28	Array, adaptive nonparametric, orthogonal arrays
720	24	Preference, lévy process
726	26	Optimal rates, optimal rates convergence, confidentiality
722	28	Ordinary differential, ordinary differential equation, high frequency

(*randomized experiment, clinical trials, false discovery rate*) preceding T700 (*high-dimensional regression, large-scale prediction problems*) and T727 (*weighted distance estimation, missing at random, variable selection*).

The list of most relevant keywords for topics appearing in Fig. 6 is reported in Table 9.

Looking at Table 9 it is possible to see that some arguments, the biggest ones in Fig. 6, identified by keywords as *comment, letters editor, rejoinder,* etc., have generated a literature debate besides a lot of papers. The evolutionary map highlights the new challenges, according to the three considered journals and their Editors' decisions, that the big data has generated in statistics, due to both high-dimensionality and large sample size. It should however be noted that the map in Fig. 6, as already said, includes only the evolution over the years generated by T525 and T523. It is a capture of the larger map, difficult to represent graphically in whole, which contains the 'hot' topics for each year. This map definitely needs to be explored in future.

Conclusions

Working on a selection of publications from the international statistical literature, we have applied the topic model approach and post-processed the results to the end of describing in depth the corresponding segment of the field over the years. The aim of our analysis was to verify the existence of predominant topics, explained by different descriptors, and to determine whether these topics generate patterns of citations. Our results seem to confirm the expectations. Accordingly, the common evaluation approaches, based on normalization with respect to a field, lose significance; a normalization with respect to the topic would seem more appropriate. Our approach raises a critical situation: with high heterogeneity of data, the identified topics exhibit a problem of robustness; in literature some other methods to cluster textual data exist. Moreover, our contribution doesn't claim to be exhaustive; it presents a case study that raises matter for debate. Taking into account our recommendations, comparisons between the topic model approach and other methods of clustering would be possible, with the purpose to normalize the bibliometric data with respect to the topics.

References

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Ferrara, A., & Salini, S. (2012). Ten challenges in modeling bibliographic data for bibliometric analysis. *Scientometrics*, *93*, 765–787.

Genest, C. (1997). Statistics on statistics: Measuring research productivity by journal publications between 1985 and 1995. *The Canadian Journal of Statistics*, *25*(4), 427–433.

Genest, C. (1999). Probability and statistics: A tale of two worlds? *The Canadian Journal of Statistics*, *27*(2), 421–444.

Genest, C. (2002). Worldwide research output in probability and statistics: An update. *The Canadian Journal of Statistics*, *30*(2), 329–342.

Grün, B., & Hornik, K. (2011). Topicsmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30.

Gupta, H. M., Campahna, J. R., & Pesce, R. A. G. (2005). Power-law distributions for the citation index of scientific publications and scientists. *Brazilian Journal of Physics*, *35*(4A), 981–986.

Hall, D., Jurafsky, D., & Manning, C. (2008). Studying the history of ideas using topic models. In *proceedings of the conference on empirical methods in natural language processing* (pp. 363–371). Honolulu, Hawaii: Association for Computational Linguistics.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Mimno, D., & Blei, D. (2011). Bayesian checking for topic models. In *proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 227–237.

Newman, M. E. J. (2006). Power laws, Pareto distribution and Zipf's law. In arXiv:cond-mat/0412004v3.

Ryan, T. P., & Woodall, W. H. (2005). The most-cited statistical papers. *Journal of Applied Statistics*, *32*(5), 461–474.

Schell, M. J. (2010). Identifying key statistical papers from 1985 to 2002 using citation data for applied biostatisticians. *The American Statistician*, *64*(4), 310–317.

Steyvers, M., T. Griffiths, T. (2007). Probabilistic topic models. In *Handbook of latent semantic analysis*, chapter 21.

Stigler, S. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science*, *9*(1), 94–108.