Electronic Journal of Applied Statistical Analysis
EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 3, 353 – 359
e-ISSN 2070-5948, DOI 10.1285/i20705948v5n3p353

BIBLIOGRAPHIC DATA: A DIFFERENT ANALYSIS PERSPECTIVE

Francesca De Battisti*, Silvia Salini

Department of Economics, Management and Quantitative Methods, University of Milan, Italy

Received 19 July 2012; Accepted 07 October 2012
Available online 16 November 2012

***Abstract**: A bibliographic record, related to a product, is composed by different information: authors, year, source, publisher, keywords, abstract, citations and so on. Citations usually have a central role in bibliometric analysis. The study of textual information could be a different analysis perspective. The idea is that documents are mixture of latent topics, where a topic is a probability distribution over words. In this paper we try to show how the scientific productivity of a research group can be described using topic models. Moreover, for the same sample, we test if the other bibliometric measures follow the known distribution laws.*

***Keywords**: Text mining, topic models, bibliometrics, distribution laws*

1. Introduction

A bibliometric database contains a large amount of different information, making possible different types of analysis [8, 7]. The purpose of the study is to present an overview of them, focusing on the analysis of textual information in order to extract the latent topics that characterize the papers.
Bibliographic data are complex, different type of information and objects are involved: measures (counts, indices), networks (co-citations, co-authorships), textual data (title, keywords, abstract, full-text). Bibliometrics, define by the Oxford English Dictionary as '*the branch of library science concerned with the application of mathematical and statistical analysis to bibliography; the statistical analysis of books, articles, or other publications*', could be used with two main aims: evaluation of research and measure of science. In this paper we focus on the second one.
Web of Science database, edited by the Institute for Scientific Information and distributed by Thomson Reuters (http://isiwebofknowledge.com/), is used for this exercise. The database is queried with reference to scientific output of all Researchers in Statistics, SECS/S01 (444

* E-mail: francesca.debattisti@unimi.it

Subjects). We analyse 302 authors and 1309 products.
In Section 2, topic models are presented and applied. In section 3 bibliometrics laws are briefly described and tested in order to verify if they are satisfied by our data. Finally, future developments and conclusions are proposed.

2. Topic Models

Topic models are based upon the idea that documents are mixture of topics, where a topic is a probability distribution over words. The documents are observed, the topics (and their distributions) are considered as hidden structures or latent variables. Topic modelling algorithms are statistical methods that analyse the words of the original texts to discover the themes that run through them, how these themes are connected to each other, and how they change over time [1].
The simplest and most commonly used probabilistic topic approach to document modelling is the generative model Latent Dirichlet Allocation (LDA) [4]. 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: i) randomly choose a distribution over topics (Dirichlet distribution); ii) 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 modelling is to use 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 modelling 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 [19]; 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 transforms the inference problem to a optimisation problem. In 2007 a correlated topic model (CTM), which explicitly models the correlation between the latent topics in the documents, has been introduced [3].
We have fitting topic models using the R Package *Topic models* [10]. To choose the optimal number of topics, perplexity is calculated [4]. The perplexity, used by convention in language modelling, is monotonically decreasing in the likelihood of the test data; a lower perplexity score indicates better generalization performance.

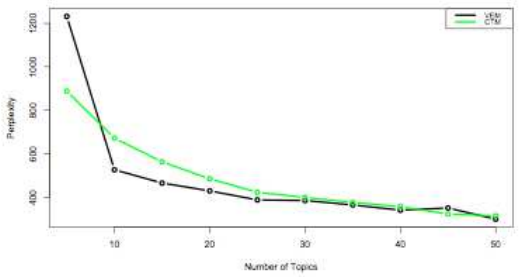

Figure 1. Perplexity by number of topics for VEM and CTM.

We have compared perplexity between VEM and CTM algorithms. The optimal number of topics looking to Figure 1 seems to be n = 30, because after this value the functions become stationary.

By topic identification, papers can be clustered. It is useful to evaluate the probabilities of assignment to the most likely topic for all documents for the estimation model chosen and to calculate the number of papers corresponding to each topic, when the most relevant one is considered (see Figure 2).

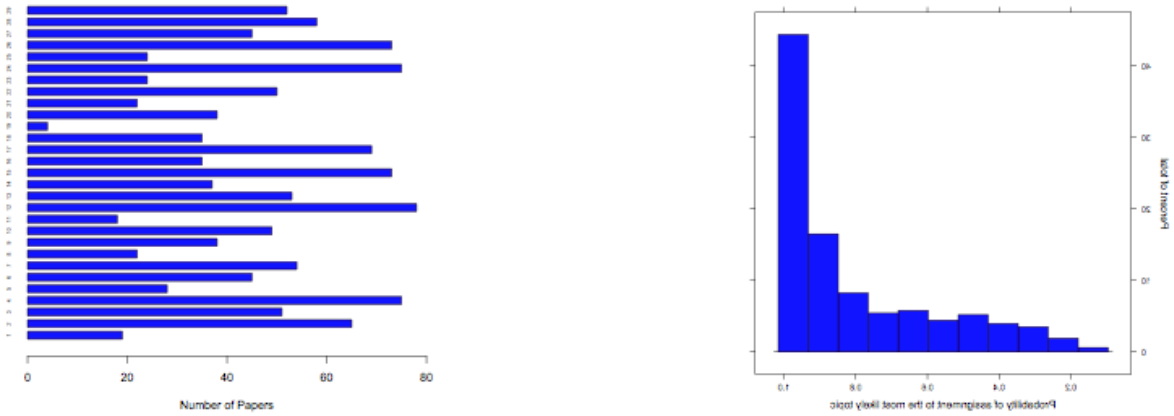

Figure 2. CTM: most likely topic distribution (left) and topic relevance (right).

It is also important to examine the strength of each topic over time, providing quantitative measures of the prevalence of particular kinds of research [9].

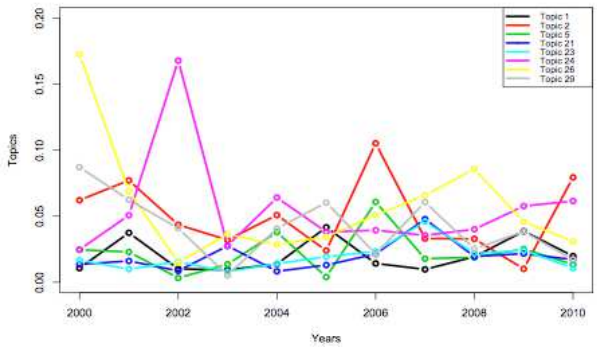

Figure 3. Topic trends according to their relevance by year.

Figure 3 shows only the topics for which there is a significant difference of the relevance means over the years: for example, in 2002 topic 26 was popular. Below the words associated with topics 26, 24 and 2 are listed:

Topic 26: bayes, factor, prior, design, priors, size, sample, evidence, fractional, trials…

Topic 24: estimator, method, simulation, function, integration, estimation, tree, measurement, reliability, risk…

Topic 2: fuzzy, component, principal, approach, clustering, dynamic, time, squares, interval, spatial…

3. Bibliometric Laws

The laws of bibliometrics originated in the first half of the 900 to describe, monitor and model the production, use and dissemination of knowledge. In particular, Lotka's law [12] characterizes the frequency of publications by author in a given field; Bradford's law [6] is useful for librarians in determining the number of core journals in any field; finally, Zipf's law [20] is often used to predict the frequency of words within a text.

The *Lotka* distribution is based on an inverse square law where the number of authors writing *n* papers is $1/n^2$ of the number of authors writing one paper. In order to test the applicability of Lotka's law to our data, for a given number of paper (NP), the number of authors (NA), the observed relative frequencies (Obs) and the expected ones (Exp) are reported in Table 1 and plotted in Figure 4. Moreover, a test based on the distance between the two cumulative quantities can be done [16].

Table 1. Observed and expected frequencies of authors for number of papers.

NP	NA	Obs	CumObs	Exp	CumExp	Dist
1	70	0.23	0.23	0.61	0.61	0.38
2	54	0.18	0.41	0.15	0.76	0.03
3	39	0.13	0.54	0.07	0.83	0.06
4	28	0.09	0.63	0.04	0.87	0.05
5	31	0.10	0.74	0.02	0.89	0.08
6	24	0.08	0.81	0.02	0.91	0.06
7	13	0.04	0.86	0.01	0.92	0.03
8	10	0.03	0.89	0.01	0.93	0.02
9	7	0.02	0.91	0.01	0.94	0.02
10	6	0.02	0.93	0.01	0.94	0.01

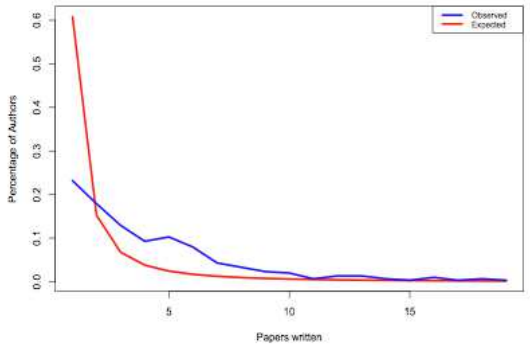

Figure 4. Observed and expected percentage of authors for number of papers.

The Kolmogorov-Smirnov (K-S) test is based on the maximum deviation D = Max |CumExp – CumObs|. At a 0.01 level of significance, the K-S statistic is equal to 0.094. If D is greater than the K-S statistic, then the sample distribution does not fit the theoretical distribution. In our case, D is 0.38, so Lotka's law does not apply to our data. Review of literature [16], different criticisms [15] and re-evaluation of the law [13] were proposed.

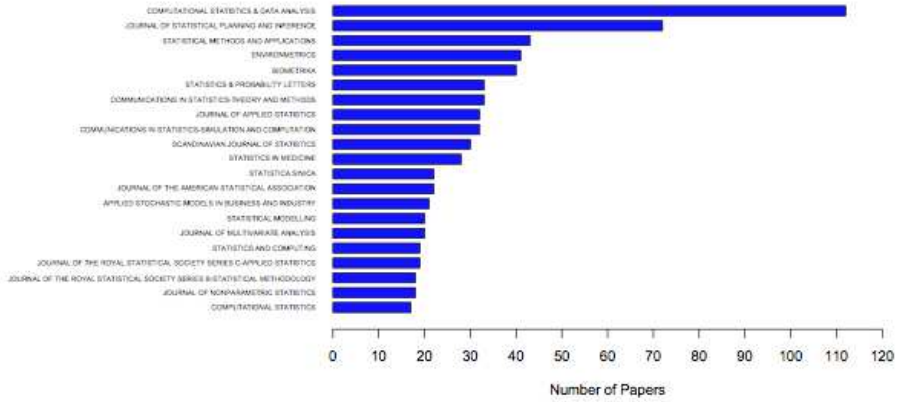

Figure 5. Distribution of the top 20 journals.

The *Bradford* distribution groups journals and articles to identify the number of periodicals relevant to a particular subject. A core of journals is thus identified which could be used to select the essential journals for a special collection. Bradford's distribution was made more general by grouping journals according to the number of citation they receive [11]. In the Figure 5 the most frequent top 20 journals of Italian statisticians are shown.

The citation distribution provides basic insight about the relative popularity of scientific publications. The number of citations received by scientific papers appears to have a power-law distribution [14, 17]. The distribution of citations is a rapidly decreasing function of citation count but does not appear to be described by a single function over the entire range of this variable [18]. *Zipf* plot is well suited for determining the large-x tail of the citation distribution.

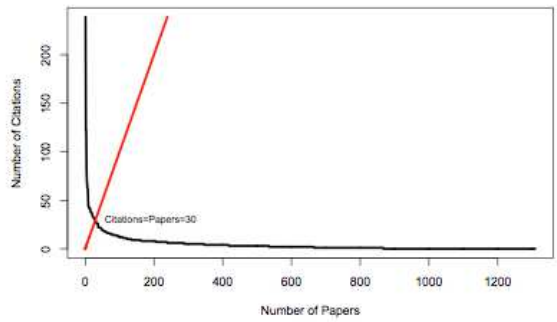

Figure 6. H-index.

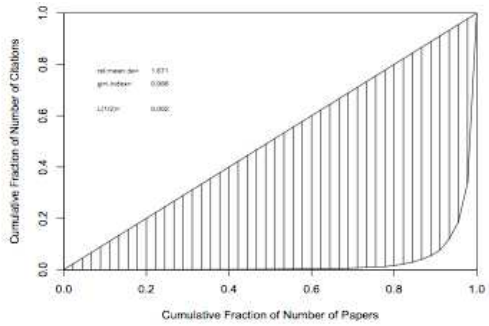

Figure 7. Lorenz Curve.

The Figure 6 shows the distribution of papers ranked by decreasing citations. The intersection between the paper distribution and the diagonal is the H-index of the Italian Statisticians' community.

When the research group is the unit of analysis, some measures of concentration should be computed. In the Lorenz curve, the cumulative proportion of articles (x-axis) is plotted against the cumulative proportion of their total citations on the y-axis. Lorenz curve captures the degree of inequality or concentration. If each article had equal value in its shares of the total citations, it would plot as a straight diagonal line (the perfect equality line); if the observed curve deviates from the perfect equality line, the articles do not contribute equally strongly to the total number of citations [5]. In our case, as confirmed by Gini's index equal to 0,956, there is a very high

degree of concentration; indeed, the 67% of papers correspond to 0 citations (see Figure 7).

4. Conclusions and Future Perspectives

Concerning to topic models, LDA and CTM assume that documents are exchangeable within the corpus and, for many corpora, this assumption is inappropriate. The topics of a document collection evolve over time. The dynamic topic model (DTM) captures the evolution of topics in a sequentially organized corpus of documents [2]. In the future we will study the evolution of topics over time and the similarity between them. Furthermore, it will be interesting to evaluate, maybe by association rules or map of science, if there are significant associations among topics, journals, country, author/ citation networks, time.
Concerning to distribution laws, a simulation study will be implemented to identify the factors that could influence them: field or area, time period, type of publication and so on.

References

[1]. Blei, D. M. (2011). *Introduction to Probabilistic Topic Models*. Princeton University.
[2]. Blei, D.M., Lafferty, J.D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113-120.
[3]. Blei, D.M., Lafferty, J.D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*.
[4]. Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*.
[5]. Bornmann, L., Mutz, R., Neuhaus, C., Daniel, H. (2008). Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*.
[6]. Bradford, S.C. (1934). Sources of information on specific subjects. *Engineering*, 137, 85-6.
[7]. De Battisti, F., Salini, S. (2012). Robust analysis of bibliometric data. *Statistical Methods & Applications*. In press. DOI 10.1007/s10260-012-0217-0.
[8]. Ferrara, A., Salini, S. (2012). Ten challenges in bibliographic data for bibliometrics analysis. *Scientometrics*, 93-3, 765-785.
[9]. Griffiths, T., Steyvers, M. (2004). Finding scientific topics. *Proceeding of the National Academy of Sciences*.
[10]. Grün, B., Hornik, K. (2011). topicsmodels: An R Package for fitting topic models. *Journal of Statistical Software*.
[11]. Hubert, J.J. (1977). Bibliometric Models for Journal Productivity. *Social Indicators Research*.
[12]. Lotka, A.J. (1926). The frequency of distribution of scientific productivity. *Journal of the Washington Academy of Science*.
[13]. McRoberts, M.H., McRoberts, B.R. (1982). A Re-Evaluation of Lotka's Law of Scientific Productivity. *Social Studies of Science*.
[14]. Newman, M.E.J. (2006). Power laws, Pareto distribution and Zipf's law. *arXiv:cond-mat/0412004v3*.

[15]. O'Connor, D.O., Voos, H. (1981). Empirical Laws, Theory Construction and Bibliometrics. *Library Trends*.
[16]. Potter, W.G. (1981). Lotka's Law Revisited. *Library Trends*.
[17]. Price, D.J. De S. (1965). Networks of scientific papers. *Science*, 149, 510-515.
[18]. Render, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B*.
[19]. Steyvers, M., Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*.
[20]. Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge.