Scientometrics (2012) 93:765–785
DOI 10.1007/s11192-012-0810-x

Ten challenges in modeling bibliographic data for bibliometric analysis

Alfio Ferrara · Silvia Salini

Received: 20 October 2011 / Published online: 15 July 2012

Abstract The complexity and variety of bibliographic data is growing, and efforts to define new methodologies and techniques for bibliometric analysis are intensifying. In this complex scenario, one of the most crucial issues is the quality of data and the capability of bibliometric analysis to cope with multiple data dimensions. Although the problem of enforcing a multidimensional approach to the analysis and management of bibliographic data is not new, a reference design pattern and a specific conceptual model for multidimensional analysis of bibliographic data are still missing. In this paper, we discuss ten of the most relevant challenges for bibliometric analysis when dealing with multidimensional data, and we propose a reference data model that, according to different goals, can help analysis designers and bibliographic experts in working with large collections of bibliographic data.

Keywords Dimensional data modeling · Multivariate statistics · Multidimensional data analysis · Topics models

Introduction

One of the most important needs when dealing with bibliographic data is to aggregate and manipulate data according to different goals and requirements by focusing on a variety of different features extracted from the same data. For example, one may be interested in analysing the publications produced by an organization by focusing on the last 3, 5, or 10 years. At the same time, by working the same data, another person could be interested in analyzing the degree of cooperation among authors working at the same institution, or

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

S. Salini (✉)
Dipartimento di Scienze Economiche, Aziendali e Statistiche,
Università degli Studi di Milano, Milan, Italy
e-mail: silvia.salini@unimi.it

they could focus on the impact of the institution's products on the research community, or they might combine the results of these two tasks to understand authors' cooperation. There are many possible examples that have some key requirements and challenges in common:

- the analysis is based on multiple points of view, which need to be used for aggregation and manipulation purposes, such as time, kinds of products, or contributor organizations;
- the granularity of data analysis is not known a priori and may change in time; for example, one should be able to easily scale from a 5-years analysis of bibliographic products to another time interval, or to move from the analysis of the productivity rate of individuals to the productivity rate of institutions; and
- the different points of view should be easily combined in complex goals for analysis, such as the citation rate for a given institution in a given time interval, normalized by the number of authors.

These requirements are usually difficult to satisfy at the analysis level. In fact, such a complex analysis requires an adequate model of data in order to be effective. In the database community, this problem has been addressed by the introduction of the multidimensional data model in the 1990s [42], which has been used for data warehouse and OLAP projects and tools [41]. In the multidimensional model, the idea of having different points of view over the data is represented by the notion of *dimension*, which is a particular descriptive component of an event, which is represented as a *fact* [1]. Using multidimensional models for bibliometric analysis is a first important challenge in bibliographic data modeling and can be stated as follows:

Challenge I (Multidimensional analysis) *In order to achieve meaningful results from bibliometric analysis, bibliographic data features cannot be taken into account separately. Bibliometric analysis must be supported by a multidimensional data model.*

Recently, such a multidimensional approach to the analysis and management of bibliographic data has been proposed [28]. However, a design pattern and a specific conceptual model for multidimensional analysis of bibliographic data are still missing. In particular, we stress how the multidimensional approach should not be limited to the analysis phase, but should be extended to the design of bibliographic databases and data access services. We would like also to give a guide to data organization for those researchers interested in collecting bibliographic data for analysis purposes.

In this paper, we address this requirement by providing a multidimensional model for bibliographic data, by stressing ten challenges that bibliometric analyses must deal with and showing how the model can be used to address these challenges. In particular, in "A conceptual model for bibliographic data", we sketch our proposed multidimensional model. In "Data description: publications, products and time", we discuss the representation of traditional bibliographic metadata and the challenges related to them. In "Indexes and metrics", we present our model for indexes and metrics used in bibliometric analysis. In "Text mining and topic extraction", we discuss the challenges and opportunities of using text mining techniques in bibliometric analysis. In "Putting things together: complex models and trends", we discuss the problem of combining the previously presented dimensions in a unique, comprehensive model. Finally, we give our "Concluding remarks".

A conceptual model for bibliographic data

Bibliographic data are often archived and organized through relational databases. However, the use of relational technology is only part of the solution for the effective organization of data. In order to obtain a complete solution to this problem, we need a *conceptual model* providing a correct and complete description of the bibliographic domain in terms of objects of interest, relationships between objects, and the objects' attributes. The main focus of the relational model as a data organization tool is on supporting a wide variety of transactional operations including data maintenance, creation, deletion and updating, together with searching and retrieving information of interest in a database. However, the relational model was not specifically conceived for supporting the requirements of data analysis. In the 1990s, operations like summarizing, consolidating, viewing, applying formulae to, and synthesizing data according to multiple dimensions [1] were addressed by Codd [13] in proposing the principles of so-called On-Line Analytical Processing (OLAP). For many reasons, including efficiency in performing operations on data and the flexibility of the model, traditional relational database systems have been shown to be inadequate for OLAP applications and, as a consequence, several approaches to a multidimensional model for databases have been proposed. In this section, we present the foundations of multidimensional models of data and discuss our proposal for a multidimensional model specifically tailored for representing bibliographic data. The multidimensional models of data have been introduced to provide a suitable and effective tool for analyzing data and supporting decisions. The main idea in the multidimensional model of data is that objects involved in the analysis are *facts*, i.e. events of interest in a domain. In our case, for example, the publication of an article is a fact, as well as the association of an index value with a publication, or the fact that a publication contains some terms or expressions. The reason why the notion of fact is introduced is that it helps in isolating and describing the different elements that compose the fact itself. These elements are *measures* associated with the fact and *objects* that are involved in the fact. For example, if we take into account the publication of an article as a fact, we can isolate some measures like the number of authors or the relevance of each author in the publication, and we can also isolate some objects composing the fact, such as the product published, the persons authoring the publication, and the year of publication. In general, the objects associated with facts are called *dimensions*, since each dimension can be chosen as a criterion for grouping data and analysing them. For example, one could be interested in counting the publications per author (i.e., grouping publications according to the author dimension), or even in counting the publication of an author per year. In the first case, we use only one dimension as an aggregation criterium over publications, while in the second case we take into account two different dimensions, such as the authors and the time. This capability of flexibly aggregating data according to multiple criteria is typical of data analysis and is supported by introducing dimensions, because they represent fact components that may be easily included in or left out of the analysis operations. Dimensions are then organized as *hierarchies* in order to represent different possible levels of aggregation. For example, the time dimension in case of bibliographic data may be represented as a hierarchy of `year` → `5-year` → `10-year` where the minimum level of aggregation is the single year of publication and the maximum level of aggregation is made of slots of 10 years. The dimension of products can range from a single article, to the journal, passing through the journal issue. Moreover, hierarchies are introduced in order to help to scale up and down along the dimensions to realize different levels of aggregation.

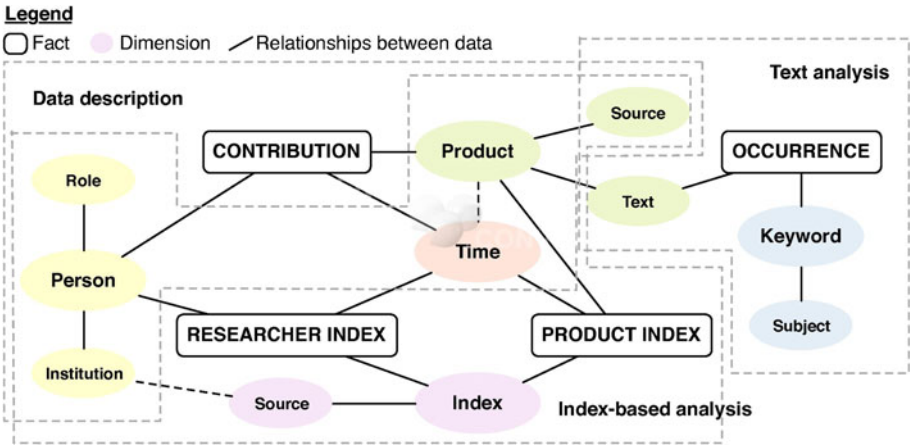

Fig. 1 Overview of the main fact and dimension hierarchies

A multidimensional data model for bibliographic data

In order to provide a conceptual view of the bibliographic data domain, we exploit the *Dimensional Fact Model (DFM)* [23], which is basically an extension of the dimensional model specifically tailored for data warehouse applications. A DFM schema is a collection of *fact schemas*. Each fact schema represents a point of view over data and is composed if the following constructs:

- a *fact*, which models a relevant event in the domain of interest;
- one or more *measures*, which are numerical attributes describing the fact according to different perspectives;
- one or more *dimensions*, which are discrete attributes used to represent facts; and
- one or more *hierarchies*, which aggregate dimensions in different levels of generality.

According to this approach, we referred to the DFM in order to represent bibliographic data by means of four main facts (leading to four different fact schemas) and five hierarchies of dimensional attributes of these facts. Moreover, the five hierarchies contain six sub-hierarchies, as shown in Fig. 1, where we provide an overview of our approach.[1]

The fact schemas are articulated to cover three main categories of goals: data description, index-based analysis, and text analysis. The main fact of data description is the `contribution`. A contribution records the fact that a `person` produces a `product` at a given `time`. This involves three dimensional hierarchies: a `person`, a `product` and a `time`. The `person` hierarchy includes two sub-hierarchies, `institution` and `role`, representing the affiliation and the role/profession of each person, respectively. The `product` contains the sub-hierarchies' `text` and `source`. `Text` represents portions of textual data extracted from the publication, such as the title, the abstract or even the full text. This hierarchy is not included in data description, but is used in text analysis. The `source` hierarchy represents the bibliographic/physical manifestation of a publication. If, for example, the publication is included in a volume or a journal, the volume or journal is taken as `source`.

[1] A detailed description of each fact schema according to DFM is given in the following sections.

Table 1 Inter-relations among dimension hierarchies and facts

	Contribution	Researcher Ind.	Product Ind.	Occurrence
Person	✔	✔	×	×
Institution	✔	✔	×	×
Role	✔	✔	×	×
Product	✔	×	✔	✔
Source	✔	×	✔	✔
Text	×	×	×	✔
Index	×	✔	✔	×
Source	×	✔	✔	×
Keyword	×	×	×	✔
Subject	×	×	×	✔
Time	✔	✔	✔	×

The index-based analysis has the goal of representing data used for all the analysis based on bibliometric indexes. Since we can distinguish between indexes conceived for authors and others conceived for publications or sources, we have two facts: `researcher index` and `product index`. As an example, the H-index is a well-known researcher index, and the Impact Factor is an example of a product index that depends on the journal of publication. In particular, the number of citations is represented in our approach as a product index. Instead, citation relationships are represented as a specific database table connecting the citing product to the cited one. Both researcher and product indexes represent the fact that a specific value of a given index is associated with a person or a product, respectively. Thus, the dimensions involved are `person` and `product`. However, index values for a given person/product may change in time and may be taken from different sources/authorities. In order to represent these two further dimensions, we include also the `time` and `index` hierarchies within this goal.

A third goal is text analysis, where we are interested in describing the contents extracted from a corpus of publications. In this case, the main fact is the `occurrence` of a `keyword` in a `text` extracted from a `product`. Each keyword is then organized in a `subject` hierarchy. A summary of the inter-relations among dimension-hierarchies and facts is shown in Table 1, while a detailed description of each fact schema according to DFM is given in the following sections.

Example In order to provide a running example of how data are organized in our model, and to show how the proposed conceptual model can be implemented in terms of concrete database tables (i.e., a logical model), we refer to a small example of referred journal publications of two hypothetical institutions in Italy.[2] Our example was conceived in order to highlight some of the typical situations of real bibliographic data, where several authors produce papers collaborating with other authors of their same institution or other institutions in other countries. In order to show the data representation in detail, we will focus in particular on five papers that are shown in Fig. 2.

[2] The model has also been tested on a collection of about 8,000 publications in the research area of databases and data modeling.

- **Inst. 1 (IT)**
  - **Paper 1**
    Auth. 1 (Inst. 1, IT), Auth. 2 (Inst. 3, UK), Auth. 3 (Inst. 4, ES)
    Journal 1, Year 2011, IF 6, Cit 20
  - **Paper 2**
    Auth. 4 (Inst. 1, IT), Auth. 5 (Inst. 2, IT)
    Journal 2, Year 2008, IF 3, Cit 100
  - **Paper 3**
    Auth. 6 (Inst. 1, IT), Auth. 7 (Inst. 1, IT)
    Journal 3, Year 2009, IF 7, Cit 10
- **Inst. 2 (IT)**
  - **Paper 4**
    Auth. 8 (Inst. 2, IT), Auth. 9 (Inst. 5, FR)
    Journal 1, Year 2011, IF 6, Cit 15
  - **Paper 2**
    Auth. 4 (Inst. 1, IT), Auth. 5 (Inst. 2, IT)
    Journal 2, Year 2008, IF 3, Cit 100
  - **Paper 5**
    Auth. 10 (Inst. 2, IT), Auth. 11 (Inst. 6, USA), Auth. 12 (Inst. 6, USA), Auth. 13 (Inst. 7, FR), Auth. 14 (Inst. 8, DE), Auth. 15 (Inst. 9, UK)
    Journal 4, Year 2006, IF 12, Cit 80

Fig. 2 Example of hypothetical publications produced by two Italian institutions

Data description: publications, products and time

The first set of data we take into account contains the traditional metadata that are usually available in most of the bibliographic data repositories. These data are intended to describe products/publications in terms of three main subsets of data: (i) product description, (ii) contributors, and (iii) time. Before discussing the modeling of these dimensions, we will focus on some challenges we need to deal with when working on these data.

Challenge II (Data availability and integration) *Data are usually provided by different and heterogeneous data sources and need to be discovered and integrated.*

In a typical scenario where we execute a bibliometric analysis of a collection of products, data are collected by more that one data source, including manual data entry, semi-automatic data collection from existing repositories (like official organization archives or libraries), and automatic data collection from public/generic repositories on the Web. Discovery and acquisition of data from multiple datasources is a classical problem in data integration and leads to a number of technical solutions that are out of the scope of this paper. What is relevant with respect to our goals is the fact that product descriptions acquired from different datasources can be featured by different levels of quality and detail. For example, products acquired from a given data source can be described by the title, the year of publication, the complete record of the journal where it was published, and the publisher of the journal. At the same time, products acquired from a different data source can be described by means of less data, such as the title and year of publication only. To this end, the resulting multidimensional model should be flexible with respect to the data that are strictly required from the analysis and should support partial analysis when only partial data are available. The separation of facts and dimensions make it possible to focus attention on the main components of bibliographic events and address the fact that some dimensional data can be unavailable.

Challenge III (One thing, one record duplicate detection and data normalization) *Bibliographic data often contain multiple references to the same objects. Data must be cleaned, normalized, and disambiguated to the end of bibliometric analysis.*

The presence of duplicates in bibliographic data collections is a frequent occurence. A duplicate in this context is multiple references to the same object, such as the reference to an author name and/or institution. This is due to two main reasons. The first is that data are often collected from multiple data sources, which can contain records referring to the same product. The second reason is that duplication is an intrinsic feature of bibliographic data even when they are collected from the same repository. Sometimes duplication is due to errors in the data repository, but often this is simply due to the fact that the same object is involved multiple times in data, such as an author who contributes to more than one publication. Although duplication is not always taken into account as a major problem in bibliometric analysis, it causes several problems in terms meaningfulness and accountability of the analysis results. In our approach, we address this problem starting from the idea of collecting a record for each distinct real object involved in the dataset at hand. In other words, we have multiple references to an object only in tables representing facts, such as the fact that an author contributes to a product at a given time. But all the data concerning the objects (e.g., authors, products, time units like years) must be recorded only one time in a single record. The achievement of this result requires standard services supporting data transformation, data cleaning, and instance matching techniques in the acquisition and loading of data into the model [1, 11].

Challenge IV (Data aggregation) *Multidimensional data need multivariate data analysis, data analysis models and statistical techniques, in that bibliographic data could be referred to different statistical units.*

In literature [20] the main types of statistical units for analysis have been distinguished at the micro (persons), meso (institutions, disciplines) and macro (regions, nations) levels. From a statistical point of view it is not easy to swich from one level to another. The underlying structure of data is not a typical hierarchical structure [10]. The basic idea of hierarchical modeling (also known as *multilevel modeling*, *empirical Bayes*, *random coefficient modeling*, or *growth curve modeling*) is to think of the lowest-level units (smallest and most numerous) as organized into a hierarchy of successively higher-level units. For example, students are in classes, classes are in schools, schools are in school districts, school districts are in states. We can then describe outcomes for an individual student as a sum of effects for the individual student, for their class, for the school, for the district and for the state. Each of these effects can be often regarded as one of an exchangeable collection of effects (e.g. all school-level effects) drawn from a distribution described by a variance component. There may also be regression coefficients at some or all of the levels [21]. For bibliographic data we cannot assume that a product belongs only to an author and to an institution or a country. Typically, a product belongs to several authors from different institutions and different countries. Multilevel models are extended to the case of an un-nested hierarchical model with a crossed structure. They can be used for the bibliometric data, but the data matrix must be extracted from the database in a timely manner, i.e. without violating any of the assumptions of the models.[3] A synthetic way to stress the core of the problems is that the aggregation of personal production steps are not additive in any to the possible dimensions of aggregation. This is because it si not

[3] A very important contribution about the statistical issues in comparing institutional performance is [22].

possible to allocate a product to a unique unit used for aggregation. Consider, for example, a paper with multiple authors from different departments (and/or subject areas, and/or countries). If we use the full count of the product for each author, in the process of aggregation, the data cannot be simply added together, because this would end up in duplication: the total number of products for the whole department would exceed the total number of actual products. Only in cases where the units of aggregation form completely disjointed sets are the productivity measures additive [20]. With respect to our example, does Paper 2 belong to Institution 1 or to Institution 2? Does Paper 5 belong to Institution 2 as well as Paper 3 to Institution 1? Paper 3 enters in the sum of the papers of Institution 1, both as a paper of Author 6 and Author 7. Are there three or four total papers from Institution 1? In general, aggregated measures of production can not be calculated as simple sums of the person's production. When it comes to obtaining, for example, the production of articles for a given university, there are two possible strategies: to consider the persons and their aggregate measures and then aggregate them in the institutions (in this case the number of papers of Institution 1 is four) or bringing together all the products selected for each author of the institution, eliminating duplicates, and finally aggregating by institution (in this case the number of papers from Institution 1 is four). Both strategies actually lose important information, which are the true characteristics of the research profile of the institution, the relationships between authors/products and the dynamics of the group. We should answer several questions. Are we interested in the distribution of products per persons, who are the unproductive researchers? Are there some excellent researchers who may meet the institution's needs? Does the institution tend to be cohesive (prevalence of internal authors) or is opened to the outside world (prevalence of external co-authors)? Does the institution have a good level of internationalization (prevalence of foreign co-authors)? It is therefore important to produce measures directly related to the institution and not only to aggregate measures of persons.

Challenge V (Comparison and ranking) *Dimensions and data that are compared actually need to be comparable*.

The multidimensional data model presented in the preview sections make possible the comparison between persons, groups of persons, or institutions. It is important to note that this comparison must be made between institutions that are mutually comparable. Institutions, in general, are not homogeneous in terms of subject areas, roles and the seniority of the persons composing them. The work of [29] focuses on the idea of not applying relative indicators. As a consequence, only a comparison of the performance of institutes with similar activities working in fields with similar bibliometric factors is possible. A good analysis must take into account all information. Especially in the comparative approach, it is important to consider all the dimensions of the data for all level of analysis. Comparisons could be done normalizing the bibliometric measures by field, age and the roles of persons. About this point, it is important to note that making a comparative bibliometric analysis of two institutions is different from creating a synthetic measure (rating), based on measurable outcomes, in order to rank very different units. The international rankings typically consider a large number of universities, regardless of their structure and vocation. They are not able to consider different missions, disciplinary compositions, incentives, structures, foundations, and so on. Instead, when we evaluate an institution, even by comparison with others, attention is paid to the choice of benchmarks that must be comparable between the institutions we want to evaluate, according to the preview aspects mentioned. Nevertheless, the rankings are used to make comparisons between institutions in absolute terms. Rankings and evaluations are often confused. This may be due to the widespread

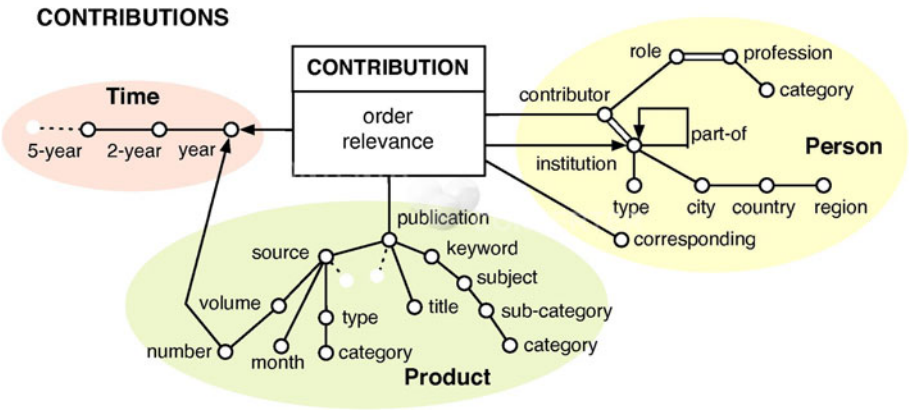

Fig. 3 Fact schema of contributions and publications

dissemination of rankings in the field of higher education in recent years [12, 39]. The need to have comparable data between institutions belonging to very different university systems has been met by an extensive use of bibliometric indicators in the rankings.[4] In fact, specific types of academic institutions (research universities in the USA) are recognised by many international rankings in which research is considered the most relevant dimension.

A fact schema for data description

Challenges III–V involve persons, products, and institutions. These data are organized in our model as shown in Fig. 3. In general terms, graphical representations of facts and dimensions like the one shown in Fig. 3 represent facts as square boxes connecting together trees of attributes, which represent dimensions. The different depth levels of dimension trees represents the different levels of the dimension hierarchy. For example, looking at the dimension of persons, we see how the institution, which represents the person's affiliation, is internally articulated in a `type` and a `city`, representing the city where the affiliated institution is located. The city then is contained in a `country` and the `country` in a `region`. According to this representation, `city` → `country` → `region` is a dimension hierarchy corresponding to different possible levels of data aggregation. This conceptual representation is then translated in relational database tables by building one or more tables for each dimension and a main fact table containing references to the dimensions.

The main fact described by this schema is the `contribution`. A contribution is referred to the single real event of a single person contributing in any role to a *product* at a *time*. Each contribution is featured by the `order` of contribution, often required in case of products with many authors, and a `relevance` of the contribution, which can be used to represent the weight of each author's contribution to a specific product. Optionally, the `contribution` fact can be also associated with information about the role of each contributor (e.g., author, editor). We note that, according to this schema, when a product has more that one contributor, we have a single record for the product, a single record for each contributor, and a single record for each contribution. In such a way, if one of the

[4] For a detailed overview of the international rankings, see [19]; for the quality assessment of composite indicators, see [3].

contribution

product	person	year	inst.	order
Paper 1	Auth. 1	y1	Inst. 1	1
Paper 1	Auth. 2	y1	Inst. 3	2
...	...	...	...	...
Paper 4	Auth. 8	y1	Inst. 2	1
...	...	...	...	...

year

id	year	ten-years	...
y1	2011	2010	...

product

id	title	source	...
Paper 1	Dynamic XML Documents...	s1	...
Paper 4	Static analysis of...	s1	...

source

id	title	type	...
s1	Journal 1	Article	...

person

id	fname	lname	name	...
Auth. 1	Mario	Rossi	M. Rossi	...
Auth. 2	John	Brown	J. Brown	...
Auth. 8	Silvia	Verdi	S. Verdi	...

institution

id	name	region	...
Inst. 1	Department...	IT	...
Inst. 2	Institute...	IT	...
Inst. 3	Department...	UK	...

Fig. 4 Portion of the main tables used for the implementation of the data description fact schema

contributors contributes to another product, this leads to a new fact, but not to a new contributor. A `contributor` is a `person` with a primary affiliation (i.e., `institution`), but who is also associated with a (potentially) different institution in case of a contribution. This allows the model to represent the current affiliation of a contributor, together with the affiliation they had at the time of the contribution. A `product` is featured by a set of attributes including, for example, `title` and `publisher`, and also by two main sub-dimensions. The first one is referred to by keywords extracted from the product and describing its contents, as we will discuss in further detail in "Text mining and topic extraction". The second is referred to the `source` of product, which represents where the product has been published (e.g., journals, proceedings, books). Finally, the `year` has been chosen as the minimal time unit in the model, and is organized in time slots as required by the analysis goals (e.g, 2-, 3-, 5-year slots).

Example In order to show an example of data descriptions, we refer to the publications `Paper 1` and `Paper 4` in Fig. 2. A portion of the main tables used for the implementation of the data description for the example is shown in Fig. 4.

The two papers were published in the same journal (i.e., `Journal 1`). We note that the database only requires the inclusion of one record for the journal that is then associated with the papers through the table `product`. Author contributions are recorded in the relational table `contribution` that implements the main concept of the contribution dimension. Then, each contribution is associated with the corresponding affiliation (through the table `contribution`). Finally, a specific table `year` is used to represent years and their relationships with the year slots in the corresponding decade.

Indexes and metrics

One of the main goals of bibliometric analysis is to discriminate among researchers and products according to their impact on a scientific community of interest. Many metrics and

indexes are available for measuring the impact of researchers and products, leading to some specific challenges for the analysis and representation of these data.

Challenge VI (Aggregation of indexes) *Indexes must be aggregated with respect to different possible dimensions and according to different aggregation functions.*

When we consider the analysis of bibliometric indexes, the main problem is aggregation. Three cases, above anything else, deserve attention and must be singled out:

(a) Aggregation of indexes referred to products per author
(b) Aggregation of indexes referred to products per institution
(c) Aggregation of indexes referred to authors per institution (including the ones aggregated in the first case)

Before describing the problems connected to these three cases, it is important to stress that the results obtained, such as if we want to calculate the average of the IF for an institution, will be different if we take a list of products (case b) without copies—in case there are more authors from the same institution—or if we take the average of the IF calculated per author (case a) and then we aggregate it per institution (case c). In this latter case, if some papers are written by one or more authors from the same institution, their value will be enhanced. As we will eventually see, it is possible to use weights in the aggregations, in order to draw a more accurate portrait of the real situation.

In case (a) of indexes per products (e.g., Citations, Impact Factor, Scimago Journal Rank) to be aggregated per authors, some of the questions that need to be answered are as follows:

- Which function of aggregation is being used: the sum, mean or median? Each of these function has limits. The sum considers the number of works, while the other two functions do not. The mean gets enhanced if at least one paper, possibly made by many, has a high impact factor and is an outlier, and is hence not representative of the real position of the author. Between the mean and the median, the latter is certainly better, even though Web Of Science, comparing the two, shows the medians of IF per Subject Category in Journal Citation Report.
- Should the function of aggregation take into account the number of authors and the position of the authors in the paper? Depending on the field considered, different different system weights can can be enhanced (e.g., first author, last author, corresponding). The work of [43] dedicates an entire chapter to the determination of the contribution of a single author according to their position and the total number of authors.

In case (b), it must be taken into account how many authors in the paper come from the institution, so that just the share of the index concerned should be added to the institution.

In case of indexes per authors (e.g., H-index, G-index) to be aggregated per institution (case c), if we consider, for example, the formula to calculate the H-index [25], it makes little sense to think that the H-index of an institution is the simple mean of the H-index of the authors composing it. The authors of [36] stress the problems connected to the aggregation of the H-index and offer a mathematic model that may help us in creating some institutional rankings; [19], in their review on rankings, used a synthetic measure of an institution, the number of scholars of an institution with an H-index superior to 30, thus avoiding the aggregation of the index.

Example Referring to our example, in the following tables we show how the indexes [number of publications (Npub) and impact factor (IF)] can be aggregated to obtain different results (rankings) for the two institutions, depending on criteria used (Table 2).

Table 2 The first table shows the aggregate (sum) indexes of the two institution in the case b) *aggregation of indexes referred to products per institution* and in the case c) *aggregation of indexes referred to authors per institution*

	Npub		IF		
	b)	c)	b)	c)	c) (Nauth)
Inst 1	3	4	16	23	10.5
Inst 2	3	3	21	21	6.5

	NPub	IF	Nauth	IF/Nauth
Inst 1				
Auth 1	1	6	3	2
Auth 4	1	3	2	1.5
Auth 6	1	7	2	3.5
Auth 7	1	7	2	3.5
Inst 2				
Auth 5	1	3	2	1.5
Auth 8	1	6	2	3
Auth 10	1	12	6	2

The last two tables show the indexes referred to products per institution for each institution. The step b) of the first table can be easily deduced taking into account that Author 6 and Author 7 are co-authors of Paper 3 and both belong to the Institution 1 (see also Fig. 2)

According to step (c) and Npub, **Inst 1** is more productive than Inst 2; however, if we instead use step b) the two institutions would be equivalent. If we look at the IF, **Inst 1** is beter if we take the step (c) and worse if we take (b). **Inst 1** is again better if the IF is normalized using the number of authors per paper.

Challenge VII (Multiple measures) *Indexes are taken from different sources and have different values associated with the same researcher/product.*

Another important aspect to consider regarding the indexes is the multifarious nature of the sources. In several case [14, 17] in which, for example, the scientific productivity of an author is described by the number of publications, citations and H-index, it is important to understand their sources, because the same variables may take different values if they are obtained via Web of Science, Scopus, Google Scholar or other disciplinary databases. In literature, there are also some papers in which the journal coverages of the various fields of these databases are compared [2, 16, 35]. In the greater part of these cases, although exceptions are made for medical sciences and some hard sciences, none of these is thoroughly exhaustive or representative. It is then quite likely to reach a point where we lack a single value for the citations and a single H-index, but one for each examined data source.

Multidimensional representation of indexes

Our approach to challenges VI–VII is to provide a flexible representation of data where switching to cases (a), (b) or (c) is easy. In particular, the multidimensional representation of indexes is achieved in our model by means of two different fact schemas, shown in Figs. 5 and 6, respectively.

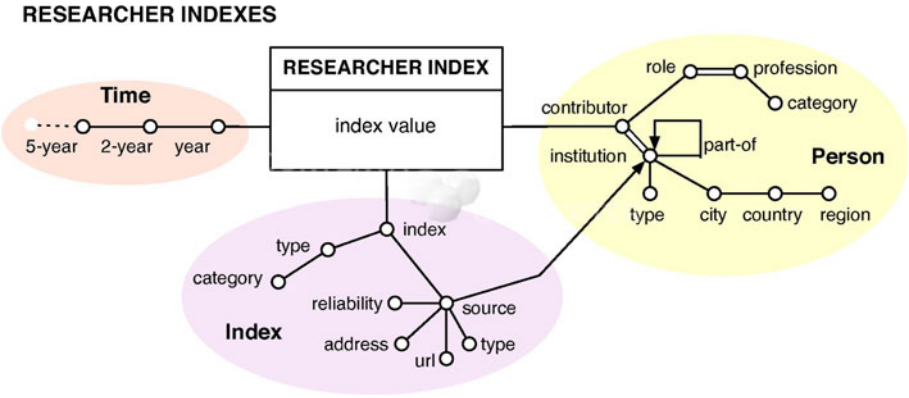

Fig. 5 Fact schema of researcher indexes

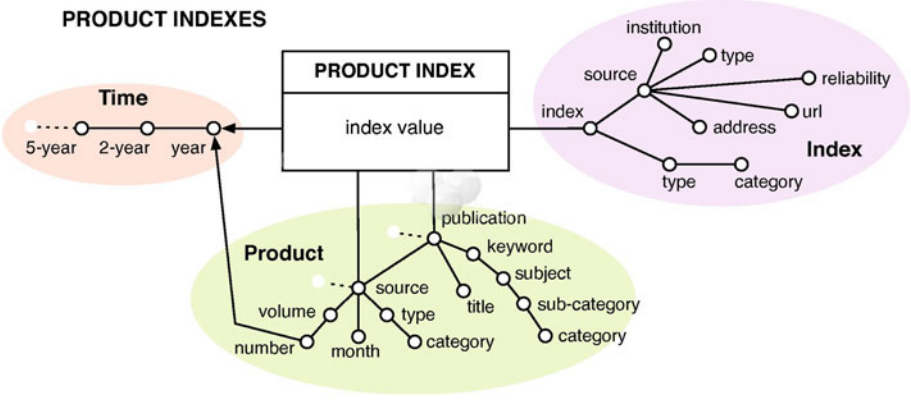

Fig. 6 Fact schema of product indexes

Both schemas share information about the index and the time, while the third dimension is given by person data in the case of the researcher's indexes and by product data in the case of product indexes. The idea is that a fact, in the case of indexes, is the attribution of an `index value` (i.e., a numerical value) associated with an `index` to a `person`/product at a `time`. It is important to note that, in our model, indexing is a time-dependent activity. In such a way, the same person/product can be featured by the same index in different times and with (potentially) different values. This represents the value of indexes changes in time. Another important consideration to capture the challenges discussed above is the fact that the identity of the index depends on the source from which it is acquired. In other words, the same index (e.g., H-Index) is considered a different and autonomous index when derived from different data sources (e.g., the H-Index of Scopus, of Web of Science, of Google Scholar). The relationship among indexes of the same type are represented by the information about `type` and `category` of the specific index.

From an analytical point of view, there are many different paths worth treading. We could be interested in grouping together the authors who boast high citation values and H-indexes according to Scholar but not in case of Web of Science; whereas others boast high values for Scopus but not for Web of Science. A particular research profile should correspond to the identified groups. Conversely, we could be interested in getting some

product_index

product	index	year	value
Paper 1	IF	y1	6
Paper 1	CIT	y1	20
Paper 4	IF	y1	6
Paper 4	CIT	y1	15

index

id	index	type	source
IF	Impact Factor	index	i3
CIT	Citation number	citations	i4

year

id	year	ten-years	...
y1	2011	2010	...

institution

id	name	type	...
i3	Web of Science	cit. database	...
i4	Scopus	cit. database	...

Fig. 7 Portion of the main tables used for the implementation of the product index fact schema

synthetic impact, diffusion and internationalization indexes. In the traditional logic of the statistical techniques applied to the reduction of data, it is possible to try to reduce the dimension of the data matrix composed by the different matrixes obtained from the different databases. These metrics, the manifest variables, are obviously tightly connected because they basically represent the same thing, and may emit some latent variables that should be exactly interpreted as the hypothesized synthetic indexes.

It may also be possible to think of predictive models in which we try to figure out if the bibliometric measures, response variables, depend in any way on factors (e.g., gender, country, field) or covariates (e.g., age of the institution, age of the author, expenses for researches, GDP). In this, case the *generalized linear models* may be useful. These models consider all levels of analysis and also the use of repeated measures, particularly useful when the same measure is used many times, because of the multifariousness of the sources (as stated above), or because it is measured many times. The dimension of time is another remarkable aspect of the analysis. It is certainly useful to understand if there is a temporal progress of the metrics—or of their synthesis, or of the clusters—connected to the institutions and the single authors alike, as well as. And if, and how, the factors and the covariates influence this very progress.

Example As an example of database representation of indexes, we take into account two different indexes, namely the Impact Factor and the number of citations. We refer to this example in order to show the importance of distinguishing the source of information used to determine the index value. In our case, in fact, we ypothesise the number of citations for papers `Paper 1` and `Paper 4` of Fig. 2 from Web of Science and Scopus, respectively. The resulting database instance is shown in Fig. 7.

As a first comment, we note how the same kind of index, i.e., number of citations, can correspond to completely different values according to the data source used to access it. In order to keep these differences, we create an index for each kind of data source. According to this approach, different citation numbers, such as the Web of Science and Scopus citations, would not be mixed but are still comparable in that we keep track of the index type (i.e., citations number).

Text mining and topic extraction

Challenge VIII (Extraction and indexing of textual data) *In order to support text mining activities on a collection of product/publications, textual information must be extracted, represented, and pre-processed.*

The introduction of text mining techniques in the workflow of bibliometric analysis is a promising research direction. However, it introduces a new kind of data that needs to be acquired from bibliographic data sources: the collection of textual data associated with each publication/product. In other words, if metadata such as author names, titles, and years were enough to support a more traditional bibliometric analysis, text mining requires the acquisition of the full-text information about publications (or, at least, of representative abstracts). Thus, full-text harvesting of publications can be a very complex task. A reasonable approach is to collect abstract and keyword lists for each publication, especially considering that many bibliographic data sources provide this information. As soon as abstracts and keywords have been collected, a pre-processing activity is needed before running text mining analysis techniques. During pre-processing, abstracts and keywords are manipulated using standard techniques for natural language processing, including the retrieval of synonymous relationships among terms, the retrieval of compound terms, the transformation of terms through stemming and/or lemmatization, and the deletion of stop-words and common terms. The general goal of this pre-processing activity is to transform textual data in a more standard and comparable collection of words.

Challenge IX (Topic-based analysis of textual data) *A collection of products/publication must be described in terms of the topics it contains in order to understand the research area of interest and the most relevant trends in that field.*

In general, when dealing with text data, we seek to achieve at least one of these goals:

1. We have a query represented by several terms and we like to retrieve the relevant documents.
2. We have a collection of documents and the goal is to get a grasp of different topics discussed in this collection.

In a bibliometric context, the second goal is more realistic and interesting: we are seeking to classify the publications (products/papers) in the collection into distinct classes (topics) and we have a compact and interpretable representation for each class. The work in [34] explains how classical bibliometric analysis can improve by using the topics. The methodological framework is not completely new [27]. In recent years many different approaches have been proposed for modeling text data. These methods can be broadly divided into two categories:

Deterministic models This family of methods is widely used in traditional text mining applications. The common framework used to represent text in this family is the vector space model. The principal component analysis and singular value decomposition methods are used to mitigate some of the problems arising from using this framework, including high dimensionality and inability in modeling some natural language features (synonymy for instance). The most well-known method in this family is Latent Semantic Analysis (LSA) [15]. Despite their wide use, deterministic methods are criticized for their lack of statistical foundations and their difficulty of interpreting results.

Probabilistic Models Probabilistic models are used to model text data by assuming a random process responsible for generating the text. Given the observed data, statistical inference procedures are used to infer the structure of the assumed random process. Unigram models, mixture of unigram models [37], probabilistic latent semantic indexing (PLSI) [26], latent Dirichlet allocation (LDA) [7], hierarchical LDA (hLDA) and

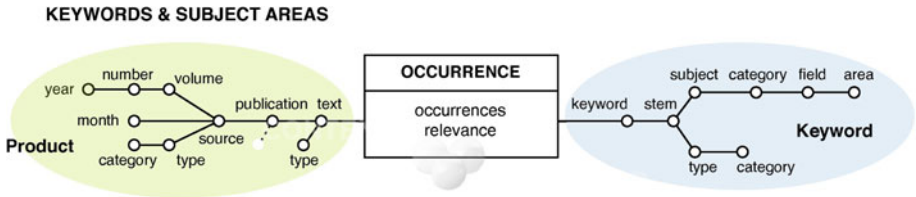

Fig. 8 Fact schema of keyword and subject areas

hierarchical Dirichlet processes (HDP) [40] are among the most well-known methods in this family. These models are often referred to as *topic models* because the latent variables in these models are mostly associated with topics in the text corpus.

Modeling text corpora has received a lot of attention in recent years. Finding compact descriptions of documents in a corpus has been one of the main goals of the research community. The availability of such descriptions will make processing increasingly large collections of text more efficient while preserving the essential statistical properties of the collection. The output will then be useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments. An introduction to topic models is given in [6, 38]. The first generation of topic models was able to capture different topics covered by a collection of documents. LDA is the best known model in this generation. The basic assumption of the LDA model is that each document is a mixture of topics, where each topic is a distribution over words. The LDA model choice of probability distributions for specifying the generative model corresponding to documents makes the resulting topics almost independent. However, it is common to have correlations between these topics. The second generation of topic models tried to capture correlations between topics [5]; moreover, a family of probabilistic time series models was developed to analyze the time evolution of topics in large document collections [4].

Multidimensional representation of textual information and topics

Multidimensional representation of textual information is achieved by two main dimensions, as shown in Fig. 8.

The first dimension describes a product in terms of its textual components, such as the title, the abstract, or even the full-text content, when available. This representation supports the analysis in focusing on the textual segments of interest. The other dimension represents the output of the textual analysis process, which is given in terms of statistical information about terms' occurrences and topics in the product collection.

Example As an example of textual information representation, we present the data obtained by running LDA on the titles and abstracts of a collection of products containing the papers in Fig. 2. As a result, we attempt to obtain ten topics, which are ten clusters of keywords extracted from the papers. Each keyword is associated with a relevant topic, which describes the prominence of the keyword for the topic at hand. Products are also associated with topics with a given relevance, denoting the importance of the topic for the product at hand. An example (limited to papers `Paper 1`, `Paper 2`, and `Paper 4` of Fig. 2) of how this information is represented in our model is shown in Fig. 9.

product_topic

keyword	text	topic	topic rel.	keyword rel.
XML	te1	t1	0.958	0.782
data	te1	t1	0.958	0.432
...	...	...	...	...
web	te2	t2	0.761	0.054
semantic	te3	t2	0.838	0.802
...	...	...	...	...

topic

id	keywords	...
t1	xml query data ...	...
t2	web semantic query ...	...

product

id	title	...
Paper 1	Dynamic XML...	...
Paper 2	Emergent semantics...	...
Paper 4	Static analysis of...	...

text

id	prod.	text	type	...
te1	Paper 4	Nowadays, the...	abstract	...
te2	Paper 2	It is well known...	abstract	...
te3	Paper 1	In this paper...	abstract	...

Fig. 9 Portion of the main tables used for the storage of textual information and topics

Putting things together: complex models and trends

The application of statistical techniques to bibliographic data is almost never immediate and simple, although most of these are known in the literature and widely used in various fields. There are still many possible alternatives and innovative proposals that can arise from bibliographic data, and which lie at the frontier of multivariate statistics and data analysis.

Challenge X (Combining multidimensional information) *The effective combination of different analysis dimensions requires a comprehensive analysis model to retrieve trends and statistical properties of the bibliographic data collection.*

Following the idea of information quality concepts (*infoQ*) [31], different goals need different data and integrated analyses. It is very unlikely, without a set a target, that a single statistical technique is immediately identified and sufficient. It is not possible to provide a simple table for bibliometric data analysis with two columns: a goal and appropriate techniques to achieve it. As evidenced by previous challenges, bibliometric data are complex in many ways. Table 3 highlights the primary and secondary variables corresponding to the different statistical units.

The transition from primary variables to derived variables requires in itself the application more or less sophisticated statistical analysis. We just mentioned the probabilistic topic models, but suitable methods are needed to select co-authorship or co-citation networks [32]. Simple questions, such as *For a specific field, do the topics changes over years and countries?*, require complex data and complex analyses. In the example, two statistical units are involved, the variable `country` is derived from the `affiliation` of the `person`, the variable `year` is related to the `product`, and the `topics` are extracted and associated with the paper through a sophisticated analysis of `abstracts` or `keywords` or full text. Moreover, it is not immediately clear to get, represent and describe the multiple relationships between the variables and their causal dependencies. Table 4 is a non-exhaustive list of possible techniques of statistical analysis that can answer some typical questions of bibliometrics. For a specific goal, the appropriate data preparation and manipulation needed is highlighted and possible techniques of analysis, with general references, are indicated.

Table 3 Statistical units, primary and derivate variables

	Primary variables	Derived variables
Person	Affiliation, role, index	Region, country,…
Product	Authors, year, source (journal, publisher), abstract, keywords, index, citation	Topics, co-authorship networks, co-citation networks,…

For example, if the goal is to understand whether, in a certain scientific community, some topics are mainly related to some countries in some years and are favored by some journals, we could apply the association rules analysis. Collections of item sets used for transaction databases and sets of associations can be represented as binary incidence matrices with columns corresponding to the items (variables levels) and rows corresponding to the papers. The matrix entries represent the presence (1) or absence (0) of an item in a particular paper. An example of a binary incidence matrix for the five papers of the example is shown in Table 5. The institutions and networks could also be considered as items.

Each statistical model assumes a specific data structure. The organization of data should not depend on the planned analysis. This would be inefficient and would lead to duplication of data and potential errors. The best way to organize data for bibliometrics is to centralize and consolidate storage in a single model that is sufficiently flexible to produce the necessary data structures for analysis when they become necessary. The topic of describing requirements and structural design of relational databases for bibliographic data

Table 4 Bibliometric data analysis schema: goals, data preparation and modeling

Goals	Data preparation	Modeling
To identify similar profiles of scholars	Aggregation of data by person, data transformation (to get normality) and data cleaning (outliers detection)	Cluster analysis [18, chap. 14.3]
To identify synthetic indexes of productivity, dissemination, internationalization	Aggregation of data by person, data transformation (to get normality) and data cleaning (outliers detection)	Factor analysis [18, chap. 14.7]
To identify associations between bibliometrics variables	Extraction of topics, extraction of networks, selection of journals, selection of countries, selection of years, and so on	Association rules [18, chap. 14.2]
To represent the associations between bibliometrics variables	Extraction of topics, extraction of networks, selection of journals, selection of countries, selection of years, and so on	Multiple correspondence analysis [24], Multidimensional scaling [8]
To study the dependency/ relationship between bibliometrics variables	Extraction of topics, extraction of networks, selection of journals, selection of countries, selection of years, and so on	Tree-Based Models [18, chap. 9], Bayesian networks [30]
To investigate the factors and variables affecting a measure	Aggregation of measures by person or by institution, data transformation	Multilevel models [21]
To study how the indexes change over time	Aggregation of indices by person or by institution, selection of years	Time series models [9]
To study how the topics change over time	Extraction of topics, selection of years	Dynamic topic models [4]

Table 5 Association rules: binary incidence matrix

	T1	T2	...	IT	UK	...	2008	2009	...	J1	J2	...
Paper 1	0	1		1	1		0	0		1	0	
Paper 2	0	1		1	0		1	0		0	1	
Paper 3	0	0		1	0		0	1		0	0	
Paper 4	1	0		1	0		1	0		0	1	
Paper 5	0	0		1	1		0	0		0	0	

T topic, *J* journal

has been addressed in detail in some previous works [44, 45]. One of the most complete works in this direction is [33], where the author proposes a relational database structure based on a detailed analysis of the main bibliometric indicators and the data required by each indicator in order to be calculated and stored. However, the work is not focused on the variability of the analysis dimensions and on the need of flexibly scaling data aggregation along an analysis dimension according to different aggregation criteria. As a result, the queries needed to extract data for the analysis are quite complex and specifically tailored for each indicator. With respect to this work, our proposal is more focused on multidimensional analysis. In particular, the fundamental distinction between facts and objects makes it possible to easily derive some indicators by referring to a limited number of tables. As an example, in [33], the retrieval of the publications of authors working in the same institutions requires work on three tables, namely `authorship`, `person`, and `affiliation`. In relational terms, this means that two *join* operations are required. Since the relational `join` is a quite complex operation, our model is conceived to avoid it when possible. In case of authors and institutions, for example, we support the query by referring only to the fact table `contribution` without any *join*.

Concluding remarks

The main idea behind this work is that bibliometrics needs multidimensional data and analysis, and that reducing bibliometric analysis in search of one-dimensionality is a stretch. Our first goal is to show how the increased use of bibliometrics to produce comparisons and rankings isonly a very marginal part of the whole world of bibliometric analysis. The paths of science and scientists can be successfully described, understood and dealt with by an appropriate organization of data and appropriate statistical analysis techniques. Remembering the warning that Peter Hall gave thrown in his speech to the *Institute of Mathematical Statistics* community in August 2011,[5] we highlighted some challenges that experts on data and data analysis should try to deal with:

1. Multidimensional analysis
2. Data availability and integration
3. Duplicate detection and data normalization
4. Data aggregation
5. Comparison and ranking
6. Aggregation of indexes

[5] http://bulletin.imstat.org/2011/09/presidential-address-peter-hall/.

7. Multiple measures
8. Extraction and indexing of textual data
9. Topic-based analysis of textual data
10. Combining multidimensional information

In this paper we have tried to describe each of these challenges and outline a possible way forward. We are aware that the list and the proposed solutions are far from exhaustive. We firmly believe that only extreme attention to the data and their organization will allow the application of advanced techniques of analysis. We also believe that only by means of advanced analysis, properly applied, taking into account various points of view, it is possible to really understand how to increase the quality of research and not just its assessment.

Acknowledgments We would like to thank the UNIMIVAL group of the University of Milan (http://www.unimi.it/cataloghi/nucelo_valutazione/ricercatori_in_breve.pdf). In the last year, they worked with us on the subject of bibliometrics; many of our ideas come from our common work and from our many fruitful discussions.

References

1. Agrawal, R., Gupta, A., Sarawagi, S. (1997). Modeling multidimensional databases. In: Proceedings of the Thirteenth International Conference on Data Engineering, ICDE '97, (pp. 232–243). Washington, DC, USA: IEEE Computer Society. http://portal.acm.org/citation.cfm?id=645482.653299.
2. Bakkalbasi, N., Bauer, K., Glover, J., Wang, L. (2006). Three options for citation tracking: Google scholar, scopus and web of science. *Biomedical digital libraries*, *3*(1), 7.
3. Benito, M., Romera, R. (2011). Improving quality assessment of composite indicators in university rankings: A case study of french and german universities of excellence. *Scientometrics*, *89*, 153–176.
4. Blei, D., Lafferty, J. (2006). Dynamic topic models. In: *Proceedings of the 23rd international conference on Machine learning* (pp. 113–120). New York: ACM.
5. Blei, D., Lafferty, J. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17–35.
6. Blei, D., Lafferty, J. (2009). Topic models. Text mining: classification, clustering, and applications, *10*, 71.
7. Blei, D., Ng, A., Jordan, M. (2003) Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.
8. Borg, I., Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications*. Berlin: Springer.
9. Brockwell, P., Davis, R. (2002). *Introduction to time series and forecasting*. Berlin: Springer.
10. Bryk, A., Raudenbush, S. (1992) *Hierarchical linear models: Applications and data analysis methods*. New York: Sage Publications, Inc.
11. Castano, S., Ferrara, A., Lorusso, D., Montanelli, S. (2008). On the Ontology Instance Matching Problem. In: *Proceedings of the 7th DEXA Workshop on Web Semantics (WebS 08)* (pp. 180–184). Turin, Italy
12. Coates, H. (2007). Universities on the catwalk: Models for performance ranking in australia. *Higher Education Management and Policy*, *19*(2), 69.
13. Codd, E., Codd, S., Salley, C. (1993). *Providing olap to user-analysts: An it mandate. Tech. rep.*
14. DeBattisti, F., Salini, S. (2010). Bibliometric indicators for statisticians: critical assessment in the Italian context. Università di Firenze, Firenze. http://air.unimi.it/handle/2434/152106.
15. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.
16. Falagas, M., Pitsouni, E., Malietzis, G., Pappas, G. (2008). Comparison of pubmed, scopus, web of science, and Google scholar: strengths and weaknesses. *The FASEB Journal*, *22*(2), 338.
17. Franceschet, M. (2009). A cluster analysis of scholar and journal bibliometric indicators. *Journal of the American Society for Information Science and Technology*, *60*(10), 1950–1964.

18. Friedman, J., Tibshirani, R., Hastie, T. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
19. Geraci, M., Degli Esposti, M. (2011). Where do italian universities stand? An in-depth statistical analysis of national and international rankings. *Scientometrics*, *87*(3), 667–681.
20. Glänzel, W., Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.
21. Goldstein, H. (2010). *Multilevel statistical models, 4th edn*. New York: Wiley.
22. Goldstein, H., Spiegelhalter, D. (1996) League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 385–443.
23. Golfarelli, M., Rizzi, S. (2009). *Data Warehouse design: Modern principles and methodologies*. Maidenheach: McGraw-Hill.
24. Greenacre, M., Blasius, J. (2006). *Multiple correspondence analysis and related methods*. Boca Raton: Chapman & Hall/CRC.
25. Hirsch, J. (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), 16,569.
26. Hofmann, T. (1999). Probabilistic latent semantic indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50–57). New york: ACM.
27. Hubert, J. (1977). Bibliometric models for journal productivity. *Social Indicators Research*, *4*(1), 441–473.
28. Hudomalj, E., Vidmar, G. (2003). Olap and bibliographic databases. *Scientometrics*, *58*(3), 609–622.
29. Irvine, J., Martin, B. (1984). *Foresight in science: picking the winners*. London.
30. Jensen, F. (1996). *An introduction to Bayesian networks, vol. 210*. London: UCL press.
31. Kenett, R., Salini, S. (2011). Modern analysis of customer satisfaction surveys: comparison of models and integrated analysis. *Applied Stochastic Models in Business and Industry*, *27*(5), 465–475.
32. Kolaczyk, E. (2009). *Statistical analysis of network data: methods and models*. Berlin: Springer.
33. Mallig, N. (2010). A relational database for bibliometric analysis. *Journal of Informetrics*, *4*(4), 564–580.
34. Mann, G., Mimno, D., McCallum, A. (2006). Bibliometric impact measures leveraging topic analysis. In: *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 65–74). New york: ACM.
35. Meho, L., Yang, K. (2007). Impact of data sources on citation counts and rankings of lis faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125.
36. Molinari, J., Molinari, A. (2008). A new methodology for ranking scientific institutions. *Scientometrics*, *75*(1), 163–174.
37. Nigam, K., McCallum, A., Thrun, S., Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning 39*(2), 103–134.
38. Steyvers, M., Griffiths, T. (2007) Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424–440.
39. Tapper, T., Filippakou, O. (2009). The world-class league tables and the sustaining of international reputations in higher education. *Journal of Higher Education Policy and Management*, *31*(1), 55–66.
40. Teh, Y., Jordan, M., Beal, M., Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.
41. Vassiliadis, P. (1998). Modeling multidimensional databases, cubes and cube operations. In: *Scientific and Statistical Database Management, International Conference on*, (p. 53). IEEE Computer Society, Los Alamitos, CA, USA. http://doi.ieeecomputersociety.org/10.1109/SSDM.1998.688111.
42. Vassiliadis, P., Sellis, T. (1999). A survey of logical models for olap databases. *SIGMOD Rec. 28*, 64–69. http://doi.acm.org/10.1145/344816.344869. http://doi.acm.org/10.1145/344816.344869.
43. Vinkler, P. (2010). *The evaluation of research by scientometric indicators*. London: Chandos Publishing.
44. Wolfram, D. (2006). Applications of SQL for informetric frequency distribution processing. *Scientometrics*, *67*(2), 301–313.
45. Yu, H., Davis, M., Wilson, C., Cole, F. (2008). Object-relational data modelling for informetric databases. *Journal of Informetrics*, *2*(3), 240–251.