Google Scholar as a Citation Database for Non-bibliometric Areas: The EVA Project Results

Alfio Ferrara, Stefano Montanelli, and Stefano Verzillo

1 Introduction

The use of bibliometric approaches for scholar and research assessment is widely enforced in STEM disciplines and it has been recently investigated also for social sciences and humanities where the use of quantitative methods is not a consolidated practice (Garfield 1980; Linmans 2010; Nederhof 2006). A number of commercial databases exists, like for example NCBI-PubMed, ISI-WoS, and Elsevier-Scopus, and they are recognized as reliable data sources for supporting calculation of citation indexes (Ferrara and Salini 2012). These databases provide "trustworthy" citation calculation, in that authors, titles, and venues of indexed publications are checked and verified. On the other side, issues about limitation and partial coverage of these databases with respect to the overall scientist production are also recognized (Archambault et al. 2006). For social sciences and humanities, open issues are even more challenging due to the fact that reference citation databases are still missing for both reliability and coverage.

In this chapter, we present the EVA (Extraction, Validation, and Analysis) project and related results about the use of Google Scholar as web database for calculation of citation indexes in non-bibliometric scientific areas, such as social sciences and humanities. The core research issue of EVA is to investigate the quality of publication records that can be retrieved from Google Scholar. In particular, the focus of this chapter is on the problem of properly disambiguating author names in retrieved records, with the aim at assigning scholars to the set of publications they actually

A. Ferrara (✉) • S. Montanelli
Department of Computer Science, Università degli Studi di Milano, Milan, Italy
e-mail: alfio.ferrara@unimi.it; stefano.montanelli@unimi.it

S. Verzillo
Department of Economics, Management and Quantitative Methods,
Università degli Studi di Milano, Milan, Italy
e-mail: stefano.verzillo@unimi.it

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_12

authored. To this end, we present the *author disambiguation techniques* developed in the EVA project based on two different strategies called *similarity-oriented* and *specificity-oriented* characterized by the use of latent semantic indexing and text normalization techniques, respectively. In EVA, similarity and specificity strategies can be exploited as alternative options as well as in a combined way for enabling a flexible configuration of the author disambiguation process. The results of the EVA project are presented on a case-study about the publication records retrieved from Google Scholar for a dataset of Italian academic researchers belonging to non-bibliometric scientific areas. The goal of the EVA case-study is twice. First, we provide evaluation results about the effectiveness of the EVA techniques for author disambiguation. Second, we provide a descriptive analysis of the obtained results. As a further contribution of EVA about the coverage of Google Scholar, a comparative evaluation of the case-study results against the Elsevier-Scopus database is also provided.

The chapter is organized as follows. In Sect. 2, we provide the literature review. In Sect. 3, we present the EVA approach and related phase-organization to author disambiguation and bibliometric data analysis. The author-based disambiguation techniques developed in EVA are illustrated in Sect. 4. In Sects. 5 and 6, we discuss the results obtained by applying the EVA approach to an Italian case-study and the results obtained by evaluating EVA against the Elsevier-Scopus commercial database, respectively. Concluding remarks are finally provided in Sect. 7.

2 Literature Review

In the recent years, the idea to exploit Google Scholar as a citation database has been proposed as a possible alternative to commercial solutions like for example ISI-WoS and Elsevier-Scopus (Aguillo 2012; Archambault et al. 2006; Falagas et al. 2008; Kousha and Thelwall 2007). In particular, in the related work, the focus is on coverage aspects. On the one side, it is recognized that Google Scholar outperforms commercial databases on the number of indexed publications (also in non-English language), as well as on the number of considered scientific areas, especially in the field of humanities and social sciences. On the other side, drawbacks about accuracy and quality of the retrieved results are widely recognized as well. In most of the existing approaches where Google Scholar is considered for citation extraction, the discussion is on bibliometrics aspects with focus on index selection for addressing a given analysis/evaluation problem (Ferrara and Salini 2012). As a matter of fact, issues about retrieval techniques and quality of extracted metadata are only marginally discussed. This is the case of systems and tools for large-scale bibliographic-data analysis, such as for example Academic analytics, Global Research Benchmarking (GRB), InCites – Thomson Reuters, Scival – Elsevier, Evaluation Support System – Research Value2, in which the supported procedure for data acquisition and validation are only marginally described (Biolcati-Rinaldi et al. 2012).

Other related works are about author disambiguation over a set of publication records, with the aim at evaluating the authorship relations between scholars and publications. The goal is to distinguish those relations that are correct (and need to be confirmed) from those relations that are incorrect mainly due to homonyms (and need to be discarded). Some interesting work in this field are (D'Angelo et al. 2011; Ferreira et al. 2010; Han et al. 2004, 2005a, b; On and Lee 2007; Smalheiser and Torvik 2009; Tang et al. 2012; Torvik et al. 2005; Yang et al. 2008). In the literature, a popular disambiguation method exploits co-authorship relations and it is based on the idea that a co-author of a scholar *sn* usually appears in multiple publications of *sn*. Thus, when a set of homonyms can be authors of a publication *p*, disambiguation is enforced by considering the other authors of *p* and by selecting the homonym-scholar that has the largest overlap between her/his sets of co-authors in past publications and the authors of the publication *p*. Disambiguation approaches based on co-authors are ineffective in those scientific areas where single-author publications are a frequent practice, such as in humanities and social sciences. As an alternative, the use of disambiguation solutions based on keyword and linguistic analysis techniques has been also proposed (Han et al. 2005b; Tang et al. 2012). However, we stress that this kind of solutions are poorly effective on scholars that are used to publish in different languages. Disambiguation can be also characterized by the execution of a preliminary clustering step with the aim at generating groups of publications with similar authors. In (Han et al. 2005b; Torvik et al. 2005), the use of an unsupervised clustering procedure is proposed. As an alternative solution, in (Ferreira et al. 2010; Tang et al. 2012), a learning approach based on a gold dataset of disambiguated publication is enforced. In these related work, the main focus is on how to perform disambiguation from a technical point of view, but the specification of a comprehensive framework is mostly missing where all the relevant aspects from dataset acquisition to disambiguation configuration and result release are adequately addressed. In particular, what is really missing is the capability to dynamically select the appropriate disambiguation strategy in which different techniques are effectively combined on the basis of the considered disambiguation case-study.

Original Contribution of the Chapter With respect to the literature, the EVA contribution is the specification of a single, comprehensive approach for author-based disambiguation. In EVA, two different disambiguation strategies have been developed to be invoked alone or in combination according to a dynamic setup mechanism that can be configured according to the dataset to disambiguate. The EVA approach has been conceived to work with publication datasets acquired from a web repository with possible unclean and duplicate metadata such as Google Scholar. In this respect, an experimental evaluation is provided to assess the effectiveness of the overall EVA approach by comparing the results of the EVA techniques applied to a dataset retrieved from Google Scholar against the results extracted from a commercial citation database.

3 The EVA Approach

The EVA approach has been conceived to enforce data analysis functionalities on a dataset of Google Scholar publications related to a selection of scholars/researchers. The approach receives in input (i) the scholars *SD* to consider, and (ii) the publications *GS* available on Google Scholar for the researchers belonging to *SD*. The publications in *GS* are retrieved by searching on Google Scholar the full name of each researcher in *SD*. Given a scholar name $s(n) \in SD$, $GS_s \subseteq GS$ represents the set of *GS* publications retrieved for *s(n)* from Google Scholar. As a matter of fact, it is possible that "mismatching" publications, namely publications not authored by the scholar *s(n)*, are included in GS_s. For this reason, the EVA approach is characterized by the use of the following disambiguation stages (see Fig. 1).

Syntax-Based Disambiguation This step works on syntax-level ambiguities of scholar names and it is concerned with the retrieval mechanism of Google Scholar. When a "full-name query" is submitted, the set of publications returned by Google Scholar is selected according to an approximate, surname-based matching procedure. As an example, it is possible that publications with author S. Ferrara and A. Ferrara are both included in the publications returned by Google Scholar for the query Alfio Ferrara. Given a scholar name *s(n)*, the EVA step of syntax-based disambiguation has the goal to detect and remove from GS_s the retrieved publications that are mismatching on author names, even when (first and/or second) names are shortened to the initial letter. Conventional cleaning techniques based on regular expressions are employed to this end.

Author-Based Disambiguation This step works on identity-level ambiguities. Given a scholar name *s(n)*, the EVA step of author-based disambiguation has the goal to detect and remove from GS_s the retrieved publications that are authored by a homonym of *s(n)*. To this end, two different author-disambiguation strategies called

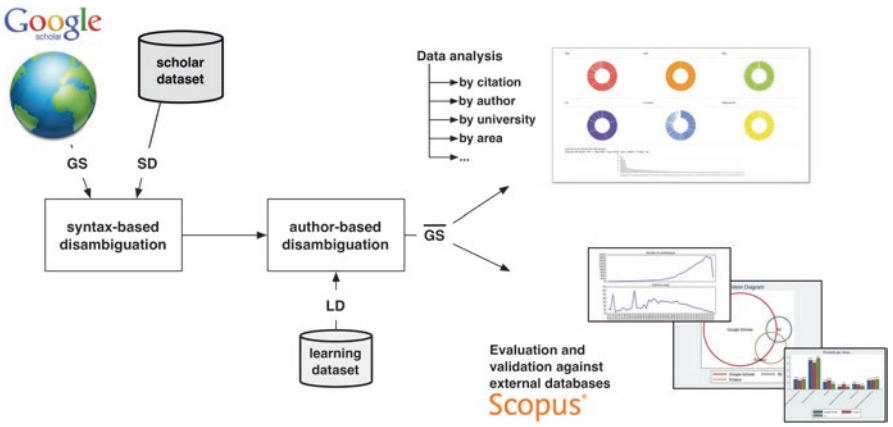

Fig. 1 The EVA approach

Table 1 The EVA case-study on Italian university researchers

Non-bibliometric area	Number of scholars
Civil engineering and architecture	1856
Historical artistic sciences	4942
History, philosophy, education and psychology	3392
Law	4706
Economics and statistics	4749
Political science	1697
Total	21,342

similarity-oriented and *specificity-oriented* have been specifically developed in EVA. These strategies represent a distinguishing feature of the EVA approach and a detailed description is provided in Sect. 4.

Syntax- and author-based disambiguation steps produce the set of publications $\overline{GS}$ as a result. In EVA, the set $\overline{GS}$ is exploited for (i) analyzing the publications authored by the scholars in *SD* (Sect. 5), and (ii) evaluating the EVA results against the Scopus citation database (Sect. 6).

Running Example The EVA project is focused on a case-study about the Italian university researchers belonging to non-bibliometric scientific areas. It is important to note that the Italian scenario is characterized by two main peculiar features. First, in Italy, the distinction between bibliometric and non-bibliometric scientific areas is defined by the Italian Ministry of University and Research (MIUR). Second, each Italian university researcher is associated with a reference scientific area. As a result, in the EVA case-study, the dataset of considered scholars *SD* has been retrieved from the MIUR by selecting the names of scholars associated with non-bibliometric scientific areas. A summary table of the scholar dataset used in the EVA case-study is shown in Table 1. For the above scholar dataset, a set *GS* containing 887,514 publications has been retrieved from Google Scholar and they have been submitted to author-based disambiguation.

As a running example throughout the chapter, we consider the following publications retrieved from Google Scholar for an Italian researcher associated with the Historical Artistic Sciences area.[1]

p_1: Renaissance Siena: Art for a City
p_2: Ludic maps and capitalist spectacle in Rio de Janeiro
p_3: Guidelines for the Management of Atrial Fibrillation

[1] The complete name of the area is Antiquities, Philological-Literacy and Historical Artistic Sciences. It has been shortened only for readability. The name of the Italian researcher is irrelevant for the clarity of the example and it is omitted.

The considered researcher has homonyms in different scientific areas, such as for example Political or Medical science. As a result, the publications retrieved from Google Scholar contain various mismatching records and the use of author disambiguation techniques is actually required. In particular, for the considered publications, only p_1 is authored by the considered Italian researcher, while p_2 and p_3 are authored by homonyms. In the following, we apply the author-based disambiguation techniques of EVA and we discuss the obtained results on the considered publications p_1, p_2, and p_3.

4 Author-Based Disambiguation Techniques

Consider a publication $p \in GS_S$ retrieved from Google Scholar in reply to a full-name query $s(n) \in SD$.[2] In EVA, the goal of author-based disambiguation techniques is to enforce an automatic decision-making mechanism for determining whether to confirm or to discard the proposed authorship relation $s(n) \longleftrightarrow p$. In other words, we aim at detecting and removing incorrect authorship relations due to homonyms on scholar names. To this end, the author-based disambiguation techniques of EVA are based on the following two requirements:

– *Scholars are associated with a reference scientific areas of research*. In EVA, for a scholar $s(n) \in SD$, we call *s(sa)* the scientific area in which *s(n)* places most of her/his publications. The specification of *s(sa)* can be exploited in different ways. As an option, the scholar $s(n) \in SD$ can manually specify her/his reference scientific area *s(sa)* within a set of available alternatives. As another option, the area *s(sa)* can be assigned to a scholar by a trusted authority (e.g., a research and education ministry, a research centre administration).
– *Scientific areas of research are associated with a dictionary of featuring keywords*. In EVA, a set of supported scientific areas $SA = \{sa_1, \ldots, sa_k\}$ is predefined and each area $sa \in SA$ is represented as a dictionary of featuring keywords.

In EVA, author disambiguation is based on the idea that an authorship relation $s(n) \longleftrightarrow p$ is confirmed when the title of the publication p is "coherent" with the terminology (i.e., the dictionary of featuring keywords) of the scientific area *s(sa)* of the scholar *s(n)*. To this end, two different disambiguation strategies called *similarity-oriented* and *specificity-oriented* have been developed in EVA. Moreover, a two-phase process articulated in *preparation* and *execution* has been specified as described in the following (see Fig. 2).

[2] In this chapter, we call publication p the bibliographic record returned by Google Scholar in response to a query.

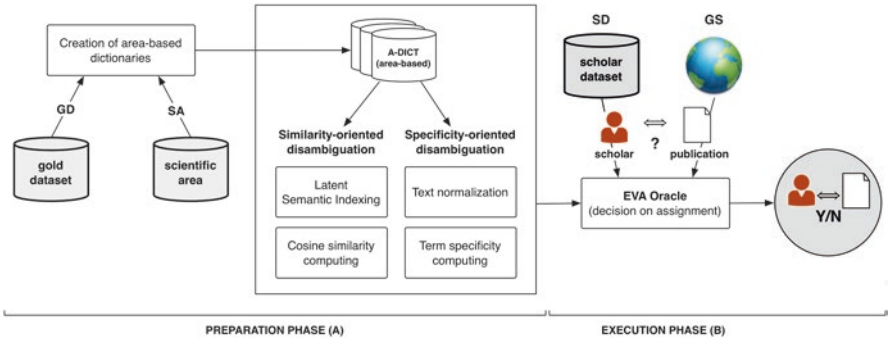

Fig. 2 The author-based disambiguation techniques of EVA

4.1 The Preparation Phase

This is a batch, preparatory phase for enabling to invoke similarity- and specificity-oriented disambiguation. Both strategies require the preliminary *creation of area-based dictionaries* in which a scientific area $sa \in SA$ is associated with a set of featuring keywords. These keywords are extracted from a *gold dataset* containing a set of publications that are expert-assigned and validated with respect to the scientific areas SA. A publication $p \in LD$ is defined as $p = < p_t, p_a >$, where p_t is the publication title and p_a is the scientific area that has been expert-assigned. Given a set of scientific areas $SA = \{sa_1, \ldots, sa_k\}$, the creation of area-based dictionaries generates a set $ADICT = \{AD_1, \ldots, AD_k\}$, where $AD_j \in ADICT$ contains the featuring keywords associated with the scientific area sa_j. The area-based dictionaries ADICT are exploited by both similarity-oriented and specificity-oriented disambiguation strategies.

Similarity-Oriented Disambiguation This strategy relies on *latent semantic indexing* (LSI) techniques, that are based on the idea to represent a potentially large set of documents over a relatively small number of considered dimensions (e.g., 400 dimensions in our case-study) (Dumais 2004). Dimensions are also called *topics* and they have the goal to make explicit the latent variables that can be inferred by observing the term distribution and co-occurrence over the considered documents. A document is created for each scientific area $sa_j \in SA$ and it corresponds to the set of featuring keywords $AD_j \in ADICT$. The set AD_j contains the titles of the publications belonging to the scientific area sa_j in the gold dataset (i.e., $AD_j = \{p_t | < p_t, p_a > \in LD \bigwedge p_a = sa_j\}$). In similarity-oriented disambiguation, the execution of LSI techniques over area-based documents produces a set of document-vectors $\overrightarrow{DOCLSI}$ as a result, where a document-vector $\overrightarrow{AD_j} \in \overrightarrow{DOCLSI}$ provides the representation of the document AD_j (and thus of the scientific area sa_j) with respect to the LSI dimensions.

Consider an authorship relation $s(n) \longleftrightarrow p$ to confirm or to discard. In similarly-oriented disambiguation, the idea is that the more the title of the publication p is similar to the document-vector $\overrightarrow{AD_j} \in \overline{DOCLSI}$ associated with the scientific area $s(sa_j)$ of the scholar $s(n)$, the higher is the likelihood that the authorship relation is correct and can be confirmed. To this end, the set of document-vectors *DOCLSI* produced as result by LSI is exploited as input for *cosine similarity computing*. Given the publication $p \in GS_s$, we call $\overrightarrow{p_t}$ the vector representation of p with respect to the document-vectors *DOCLSI*. The cosine similarity between the publication $p \in GS_s$ and the scientific area $s(sa_j)$ is calculated as follows:

$$sim\left(p,s\left(sa_j\right)\right) = \frac{\overrightarrow{p_t} \cdot \overrightarrow{AD_j}}{\overrightarrow{p_t} \quad \overrightarrow{AD_j}}$$

where $\overrightarrow{p_t} \cdot \overrightarrow{AD_j}$ is the scalar product between the two vectors $\overrightarrow{p_t}$, $\overrightarrow{AD_j}$, and $\overrightarrow{p_t}$, $\overrightarrow{AD_j}$ represent the Euclidean norm of the vectors $\overrightarrow{p_t}$, $\overrightarrow{AD_j}$, respectively.

Specificity-Oriented Disambiguation This strategy relies on *text normalization* techniques for transforming an input text into a set of string tokens (or simply tokens in the remaining of the chapter) that represent the relevant lexical elements of the input text. Normalization is enforced by applying a sequence of *natural language processing* (NLP) functions, such as for example tokenization, upper-case and elision removal, and stemming that are employed in the Italian case-study.[3] Consider the set AD_j containing the titles of the publications belonging to the scientific area sa_j in the gold dataset. In specificity-oriented disambiguation, the execution of normalization techniques over area-based documents produces a set of documents $\overline{DOCNORM}$ as a result, where a document $\overrightarrow{AD_j} \in \overline{DOCNORM}$ contains the tokens extracted from the publication titles of AD_j through the execution of normalization techniques.

Consider an authorship relation $s(n) \longleftrightarrow p$ to confirm or to discard. In specificity-oriented disambiguation, the idea is that the more the terms/tokens in the title of the publication p are "specific" of the scientific area $s(sa_j)$ (i.e., the tokens appear frequently in $\overrightarrow{AD_j}$ and rarely in other areas), the higher is the likelihood that the authorship relation is correct and can be confirmed. To this end, the set of documents $\overline{DOCNORM}$ produced as result by normalization is exploited as input for *term specificity computing*. Given the publication $p \in GS_s$, we call $\overline{p_t}$ the tokens extracted from the title of p. The specificity between the publication $p \in GS_s$ and the scientific area $s(sa_j)$ is calculated as follows:

[3] A detailed description of normalization techniques and related NLP functions is provided in (Manning et al. 2008).

$$spec\left(p,s\left(sa_j\right)\right)=\left(\sum_{i=1}^{i=r}\frac{f\left(t_i,\overline{AD_j}\right)-f\left(t_i,\overline{DOCNORM}\right)}{\sqrt{f\left(t_i,\overline{DOCNORM}\right)}}\right)^3$$

where $t_i \in \overline{p_t}$ is a token belonging to the normalized title of the publication $p \in GS_s$, r is the number of tokens in $\overline{p_t}$, $f\left(t_i,\overline{AD_j}\right)$ is the number of occurrences of the token t_i in $\overline{s\left(sa_j\right)}$, and $f\left(t_i,\overline{DOCNORM}\right)$ is the number of occurrences of t_i in all the documents of $\overline{DOCNORM}$. Negative values of $spec(p, s(sa_j))$ of means that the tokens of $\overline{p_t}$ rarely appear in $\overline{s\left(sa_j\right)}$, thus the title of the publication p is not coherent with the scientific area $s(sa_j)$. Conversely, positive values of $spec(p, s(sa_j))$ means that the tokens of $\overline{p_t}$ frequently appear in $s\left(sa_j\right)$, thus the publication p is coherent with the scientific area $s(sa_j)$.

4.2 The Execution Phase

This is an executive phase where disambiguation is invoked. This phase is based on the so-called *EVA Oracle* which is in charge of disambiguation resolution. Given a scholar name $s(n) \in SD$ and a publication $p \in GS_s$, the EVA Oracle exploits similarity- and specificity-oriented strategies to decide whether to confirm or to discard the authorship relation $s(n) \longleftrightarrow p$. Three disambiguation modalities are enforced by the EVA Oracle for taking the decision:

- *Modality-by-similarity* (*modsim*). This modality exploits the similarity-oriented disambiguation strategy. According to *modsim*, the authorship relation $s(n) \longleftrightarrow p$ is confirmed when the result of cosine similarity $sim(p, s(sa_j))$ is higher than a similarity threshold th_{sim} denoting the minimum degree of similarity required for authorship-relation confirmation. In the EVA case-study, the similarity threshold is experientially set to $th_{sim} = 0.8$.
- *Modality-by-specificity* (*modspec*). This modality exploits the specificity-oriented disambiguation strategy. According to *modspec*, the authorship relation $s(n) \longleftrightarrow p$ is confirmed when the result of the specificity function $spec(p, s(sa_j))$ is higher than a specificity threshold th_{spec} denoting the minimum degree of specificity required for authorship-relation confirmation. In the EVA case-study, the specificity threshold is experientially set to $th_{spec} = 0$.
- *Modality-by-similarity-specificity* (*modsimspec*). This modality exploits both the similarity- and the specificity-oriented disambiguation strategies. According to *modsimspec*, the authorship relation $s(n) \longleftrightarrow p$ is confirmed when both the criteria of *modsim* and *modspec* are satisfied. The relation is discarded otherwise.

4.2.1 Running Example

Consider the scholar *s(n)* associated with the scientific area s(sa) = Historical Artistic Sciences and the publications p_1, p_2, and p_3 presented in Sect. 3. According to similarity- and specificity-oriented strategies, the results of cosine similarity computing (*sim*) and term specificity computing (*spec*) are the following:

$sim(p_1,s(sa)) = 0.855$; $spec(p_1,s(sa)) = 0.0237$
$sim(p_2,s(sa)) = 0.564$; $spec(p_2,s(sa)) = 0.0125$
$sim(p_3,s(sa)) = 0.485$; $spec(p_3,s(sa)) = -0.0011$

In the EVA case-study, the modality-by-similarity-specificity is exploited for author-based disambiguation. This choice is the result of an extensive experimentation and it has been selected for enabling to obtain the best performance in terms of effectiveness. In this respect, only the authorship relation with the publication p_1 is confirmed (i.e., $sim(p_1,s(sa)) > th_{sim}$ and $spec(p_1,s(sa)) > th_{spec}$), while the relations with p_2 and p_3 are discarded.

4.3 Assessment of the EVA Disambiguation Techniques

The disambiguation techniques exploited in EVA require to properly set up a set of parameters, that are the number of dimensions for latent semantic indexing, the similarity threshold for the similarity disambiguation modality, and the specificity threshold for the specificity disambiguation modality. In order both of experimentally tuning the parameters driving disambiguation and of assessing the EVA disambiguation techniques, we build a dataset of about 350,000 publications extracted from the curriculum vitae of Italian scholars working in all the scientific areas, either bibliometric or non-bibliometric. Since those publications are taken from the curricula, the authorship relation between each publication and the corresponding scholar is correct and, as a consequence, also the relation holding between a publication and the scientific area of the author is correct.

The dataset has been then split in a *training* and a *testing* set. The training set has been used for tuning the EVA parameters, while the testing set has been used for evaluating the quality of the EVA disambiguation process.

In particular, we measured the quality of disambiguation by computing the *precision* and *recall* of the disambiguation process. The precision is the fraction of authorship relations that are confirmed by EVA that are actually correct. The recall is the fraction of the authorship relations of an author that are actually confirmed by EVA. In order to set up the parameters, we executed several tests with different combinations of parameter values over the training set of authorship relations and we measured precision and recall. As a result, we empirically set up to 400 the number of dimensions of latent semantic indexing, to 0.8 the similarity threshold,

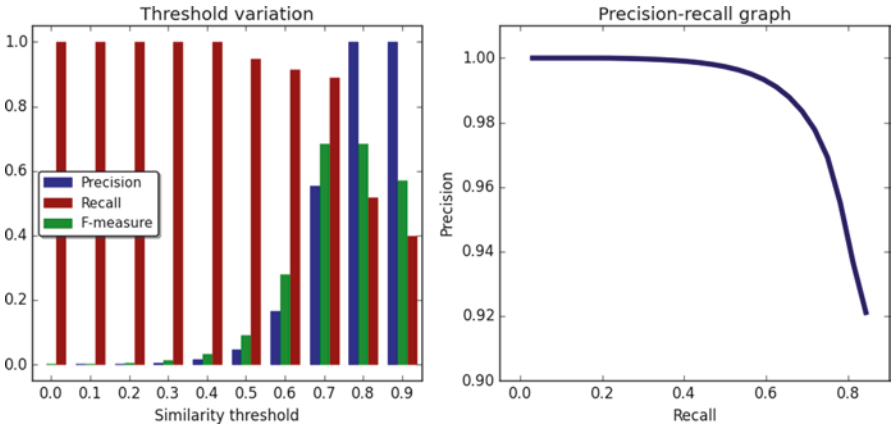

Fig. 3 Precision and recall of the EVA disambiguation techniques

and to 0 the specificity threshold. In particular, we observed that parameter that mainly affects the quality of disambiguation is the similarity threshold. We report the values of precision, recall and F-measure (i.e., the harmonic mean of precision and recall) collected at different levels of similarity in Fig. 3. We note that the balance of precision and recall, represented by the F-measure, reaches the highest values with similarity threshold values of 0.7 and 0.8. We choose 0.8 because with that value of similarity threshold, precision is maximal. This choice is motivated by the fact that we preferred to maximize the correctness of the authorship relation (i.e., precision) rather than the coverage of all the author publications. However, using the testing set we correlated the levels of precision and recall achieved by the EVA disambiguation techniques by means of the Precision-Recall graph (see Fig. 3). The graph report the behavior of the disambiguation precision at different levels of recall. In particular, we note how the EVA disambiguation techniques are capable of correctly validate more than 90% of authorship relations (0.9 precision) by covering correctly more than 80% of author publications (0.8 recall).

5 Evidence from Italian Academia

The approach of adopting Google Scholar to evaluate non-bibliometric disciplines proposed in this chapter may also contribute to the description of the overall Italian research activity in these fields in the last decades and to testing indirectly the validity of a bibliometric-style approach to analysis in fields in which the international scientific journals are not the main vector to disseminate research findings. The main focus of this section is to provide a synthetic description of recent research trends in non-bibliometric disciplines in Italian academia over the last few years.

Table 2 Descriptive statistics

Discipline	Stats	Publication (Sum)	Citation (Avg)	Citation (Sum)	H Index	Academic Seniority
Civil engineering and architecture	Mean	6.69	4.8	18.64	1.25	14.17
	Std	7.71	24.56	65.05	1.58	16.89
	N. Obs.	1353	1353	1353	1353	1286
Antiquities, philological-literacy and historical artistic sciences	Mean	8.41	4.83	36.55	1.59	17.8
	Std	12.79	20.35	332.6	2.14	18.32
	N. Obs.	3836	3836	3836	3836	3.742
History, philosophy, education and psychology	Mean	10.72	5.75	63.19	2.23	18.1
	Std	13.19	10.83	310.41	3.13	20.23
	N. Obs.	2756	2756	2756	2756	2756
Law	Mean	7.73	4.46	22.29	1.49	14.76
	Std	11.61	13.92	81.33	1.7	19.64
	N. Obs.	3071	3071	3071	3071	2991
Economics and statistics	Mean	22.31	13.88	281.97	5.26	16.51
	Std	25.82	31.34	996.39	5.17	17.02
	N. Obs.	4224	4224	4224	4224	4192
Political science	Mean	12.6	8.73	99.92	2.9	16.99
	Std	15.4	18.33	345.01	3.3	17.34
	N. Obs.	1427	1427	1427	1427	1404
Total	Mean	12.41	7.54	104.5	2.69	16.6
	Std	17.92	22.18	562.02	3.67	18.44
	N. Obs.	16,667	16,667	16,667	16,667	16,306

We collapsed the publication records obtained from GS at the researcher level to describe the current status of the research in non-bibliometric disciplines published by academics who were active in the Italian higher education system on 31 December 2014. The descriptive statistics (means, standard deviations and numbers of observations) are reported in Table 2 for the selected disciplines. A first look at the data provides a general idea of great heterogeneity across disciplines, with, on the one hand, the average researcher in economics and statistics publishing 22 research papers or products (e.g. conference proceedings, book chapters or journal articles) indexed by Google Scholar with on average 14 citations each and having an H-Index of 5 after 16 years of research activity and, on the other hand, the representative researcher in law publishing fewer than 8 papers with 4 citations (on average) and having an H-Index of 1:5 after 14 years of activity. However, considering the fact that the data that we used are essentially count data resulting from the collapse of repeated events at a certain point in time (e.g. at the lower level the number of citations of an article while at the aggregate level the cumulative or mean citation counts of individual publications, etc.) analysed by groups of researchers (e.g. researchers within a specific discipline), examining the simple descriptive statistics of a selected set of indexes could clearly be misleading (Bornmann et al. 2008). In fact, the assumption of normally distributed data usually required by common statistical

tests is clearly violated (count data are better represented by Poisson or negative binomial distribution) and a simple means comparison may lead to a distorted picture when comparing different research subsets of individuals. For these reasons simple box plot diagrams (Fig. 3) may be a more convenient way of visually summarizing the differences across discipline distributions. Indeed, outliers usually provide important information on very productive (inactive) academics or highly (lowly) cited scholars, such as academic stars or inactive individuals, even if they cannot depict the entire output of a research group.

Figure 3 shows the box plots of the H-Index and of the total number of EVA publications by discipline. The smallest observations have zero citations and the H-Index for all the disciplines in both cases as well as the lower quartiles (the maximum values of the intervals containing the less productive and less cited 25% of all academics) are close to zero except for economics and statistics and political sciences. In addition, the medians (the maximum value of the interval containing 50% of all observations) and upper quartiles (containing 75% of all observations) are quite different for economics and statistics and political sciences if compared with the other non-bibliometric disciplines conveying the impression of a research attitude that is clearly situated in the middle between hard sciences and arts and humanities (Checchi et al. 2014). Finally, any observed value outside the ends of the whiskers is considered unusual or an academic outlier.

When the academic discipline is our main unit of analysis, some measures of concentration (e.g. Lorenz curves and Gini indexes) may also be calculated to distinguish between research disciplines exhibiting some sort of "collective strength" and groups with more "individual strength" (Daniel and Fisch 1990; Burrell 2006). To describe these pattern of research better, a set of Lorenz curves representing the cumulative number of papers published against the cumulative number of citations of researchers belonging to each of the selected disciplines is provided in Fig. 4. In this way, the Lorenz curves capture the degree of concentration with the implicit assumption that each individual researcher in a specific discipline contributes to an equal share of the total number of citations. If each researcher had an equal value in the shares of the total citations of the discipline or of the H-Index, the Lorenz curve would result as a straight line (the diagonal) reflecting the pattern of perfect equality.

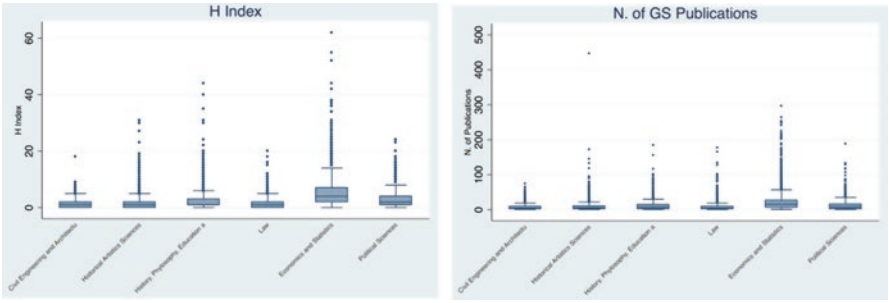

Fig. 4 Boxplots of the H-Index and number of EVA products

Otherwise, if the observed Lorenz curve deviates from the diagonal line, the researchers do not contribute equally to the total number of citations or the H-Index. Figure 4 clearly show that, in the non-bibliometric disciplines of Italian academia, this is not the case.

For example, for civil engineering and architecture and for economics and statistics, the 20% most productive scientists account for about 35% of the total citations in their fields, while for the other disciplines the inequality is a little higher, with 20% of the most productive academics receiving around 45% of the EVA citations in their respective field. The research impact, as measured by the total citations, is more concentrated in a relatively small group of researchers in arts and humanities with respect to economics and statistics or engineering. Overall it seems that in non-bibliometric disciplines "collective strength" is the common pattern instead of situations of "individual strength", with large fractions of citations relating to very small fractions of researchers.

6 External Validity Assessment of the EVA Project

A comparative evaluation of the EVA project data using established bibliometric data sources is also useful for assessing the external validity of the project and discussing the empirical evidence regarding the coverage and reliability of the bibliometric indexes computed using the EVA data set. The most reliable benchmark data source should be the ANPREPS database "National Archive of Professors and Researchers scientific publications" containing the entire academic production of Italian professors as prescribed by the Law 1 = 2009 (art. 3 bis-2), but unfortunately this database has never been realized in Italy. Hence, we focus on international publications only as collected by the two commercial databases ISI-Thomson and Scopus-Elsevier (Fig. 5).

We collected publications of the 16,667 Italian academics active at the end of 2014 in non-bibliometric fields as they were defined by MIUR. The substantive difference between these resources is mainly related to the difference in the extent of coverage among research disciplines. On the one hand, a better representation of publications in arts and humanities is guaranteed by Scopus-Elsevier, while, on the other hand, ISI-Thomson is preferable for more scientific sub-disciplines, such as statistics or psychometrics. For these reasons each database is peculiar and it is necessary to take into account its own characteristics when it is used as reference point. In general, the comparisons with EVA provided in this section must therefore be considered as relative comparisons between benchmark databases instead of absolute comparisons with respect to a true reference point.

Table 3 shows the relative composition of the three data sets by discipline. A similar pattern can be identified between EVA, Scopus and ISI publications collected for Italian academics. About 45% of the overall number of published papers is authored by researchers in the economics and statistics fields, around 15% by authors in history, philosophy, education and psychology, 10% in law and 15% in

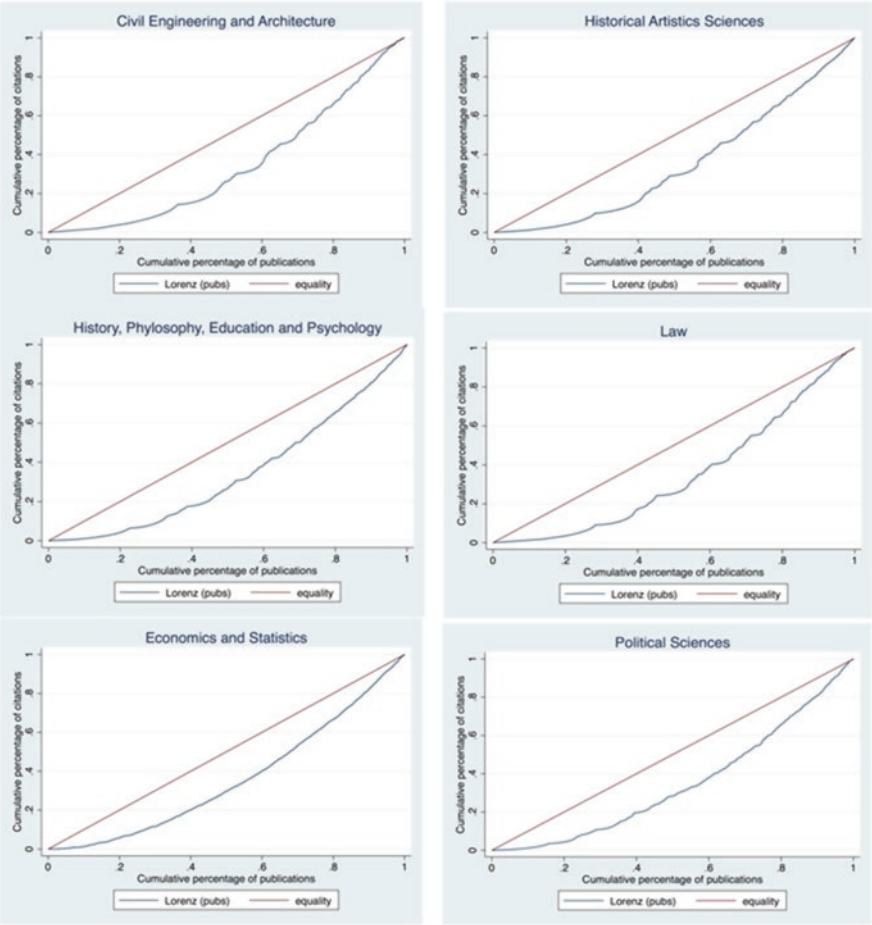

Fig. 5 Lorenz Curves by Discipline

Table 3 Relative frequencies of papers by Source

Discipline	EVA	Scopus	ISI
Antiquities, philological-literacy and historical arts sciences	15.68	13.87	15.98
Economics and statistics	45.72	41.85	49.07
Law	11.56	14.09	8.87
Civil engineering and architecture	4.32	7.92	4.42
Political science	8.62	7.51	5.69
History, philosophy, education and psychology	14.1	14.76	15.97
Total	100.0	100.0	100.0

antiquities, philological-literacy and historical artistic sciences. ISI-Thomson seems to be more biased towards disciplines such as economics and statistics (49% vs 45% of EVA and 41% of Scopus) and history, philosophy, education and psychology (16% versus 14%), while it is less representative of the arts and humanities (e.g. 8% in law versus 11% of EVA and 14% of Scopus). In addition, Scopus is more oriented towards arts and humanities, with a larger fraction of papers in law (14% versus 11% of EVA and 9% of ISI) and in civil engineering and architecture (8% versus 4% of the others). To this end, EVA can be considered as the most balanced source, with a good degree of coverage of all the non-bibliometric disciplines.

In the previous section, we described the main features of the research produced by Italian academics in non-bibliometric disciplines that emerged when adopting the EVA approach. However, the external validation of EVA requires us to compare each paper collected by following the suggested approach with the whole sets of papers in the ISI and Scopus repositories to evaluate their degree of completeness. To this end, some matching algorithms were employed to check whether an EVA record was also collected by the other sources or not. Indeed, differences in titles' syntaxes or in authors' names between the three sources may alter the results of standard matching procedures. Table 4 shows that a conflict arises when the same bibliographic record published in 2014 by "Amendola A." and co-authors is collected with two similar "but not equal" strings for the title in EVA (Google Scholar) and Scopus. In this case, exact matching would fail to recognize the same paper and a less precise matching criterion is required.

In contrast, Table 5 suggests that calibrating the matching algorithm of titles' strings is a considerable challenge given the fact that less stringent criteria may be misleading as well. Table 5 below reports four papers, two collected by EVA and two by Scopus, with the same title in three cases out of four and different publication years (2012, 1997 and 2006), for which it is very difficult to calibrate the matching algorithm (whereby stringent and relaxed criteria produce very different results).

Table 4 Example of matching conflict between different data sources

Id paper	Title	Year	Author	Source
34,889	CFE network: The annals of computational and financial econometrics	2014	Amendola, A. Et al.	EVA
178	CFE network: The annals of computational and financial econometrics: 2nd issue	2014	Amendola, A. Et al.	Scopus

Table 5 Example of matching conflict within the same data source

Id Paper	Title	Year	Source
28,635	15 comparative law and economics	2012	EVA
28,667	Comparative law and economics	1997	EVA
51,885	Comparative law and economics	2006	Scopus
51,886	Comparative law and economics	2012	Scopus

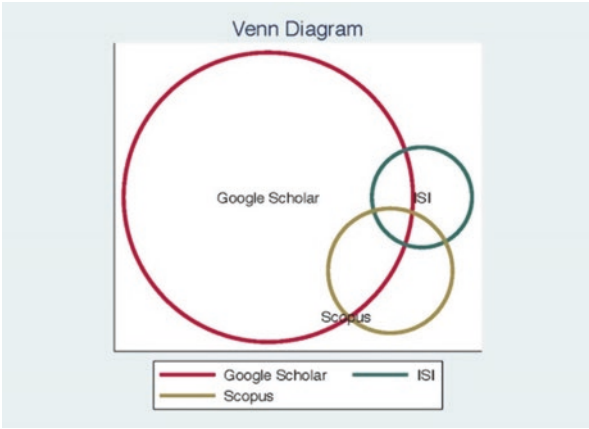

Fig. 6 Venn diagram

Table 6 Intersection of data sources by field

Discipline	EVA in Scopus	EVA in ISI	EVA in ISI and Scopus
Civil engineering and architecture	5.6	1.26	0.87
Antiquities, philological-literacy and historical arts sciences	1.69	0.76	0.32
History, philosophy, education and psychology	4.86	2.24	0.87
Law	1.06	0.36	0.15
Economics e statistics	15.8	7.89	5.21
Political sciences	4.33	1.56	0.75
Total	8.91	4.27	2.67

For the external validity assessment exercise of EVA, we adopted a conservative approach with a highly selective heuristic matching algorithm that guarantees a high degree of reliability of the corresponding associations of records between different data sources. The selected procedure, for example, discarded all the possible associations reported in Table 5 requiring a higher level of correspondence between the title strings. The association of two records in different data sources requires two conditions: a maximum one-year lag between their publication years and an edit distance lower than 2.[4]

A Venn diagram of the matched papers following the described procedure between the three sources is proposed in Fig. 6. It represents the universe of all publications (before 2012) collected by the three data sets and their set representation derived by the adoption of the matching algorithm. It shows that the fraction of papers shared by EVA and Scopus is conditioning by the number of papers in EVA larger than the one shared with ISI (9% versus 4%), but both are subsets of limited size with a very small

[4]The edit distance between two titles' strings (S_1 and S_2) is the minimum number of operations (inclusion, substitution or deletion) on single characters needed to transform S_1 into S_2.

intersection (2:6% of the EVA's size). Surprisingly, the intersection between ISI and Scopus accounts for around 30% of ISI and 15% of Scopus products.

If we consider the intersections of EVA with both ISI and Scopus separately, we observe some peculiar differences. Table 6 shows the existence of large heterogeneity across the disciplines, with a relatively larger fraction of EVA publications shared with Scopus (15%), ISI (8%) and both (5%) for individuals in the field of economics and statistics.

In addition, Fig. 7 confirms for the EVA data set the validity of one of the most common empirical laws in bibliometric disciplines: Lotka's law. It describes the frequency of publication by authors, showing that a large fraction of papers is authored by a small number of researchers. The distribution of the number of authors against the number of contributions made by the authors is highly asymmetric, with a higher concentration of articles among a few authors (great producers), while the remaining articles are distributed among several authors. The empirical estimate of Lotka's law between Scopus and EVA is similar, while it is steeper in ISI, identifying a substantially robust picture of the non-bibliometric research produced in Italy in the observed period. Moreover, Fig. 8 shows that the citation distributions are quite similar, because they do not particularly differ among the sources analysed.

Finally, we analyse the citation counts of each paper as they emerged from different data sources. First, we notice a strongly positive set of correlations between the citation counts, which are all higher than 0.85 and statistically significant (Tables 7 and 8).

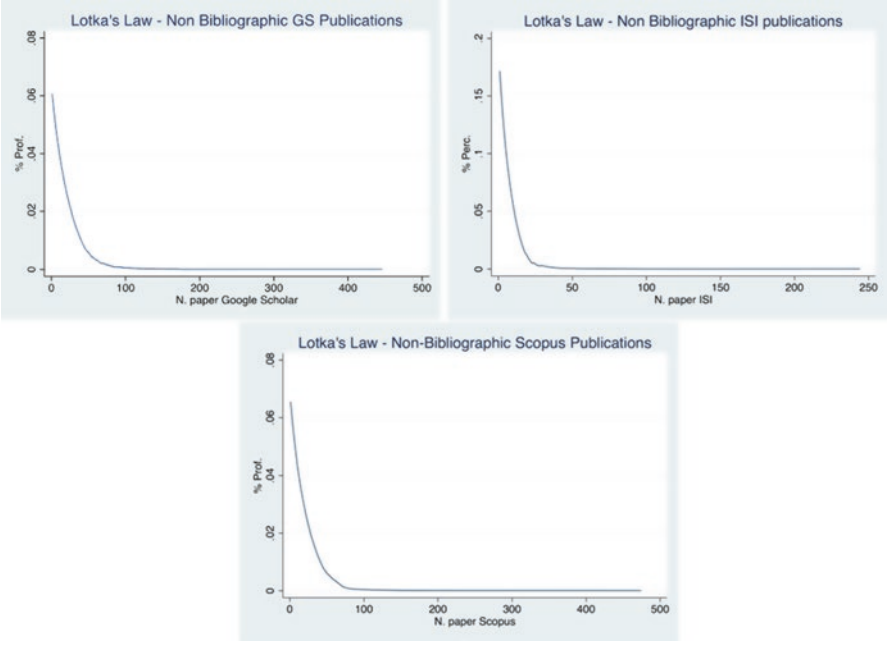

Fig. 7 Lotka Law by bibliometric source

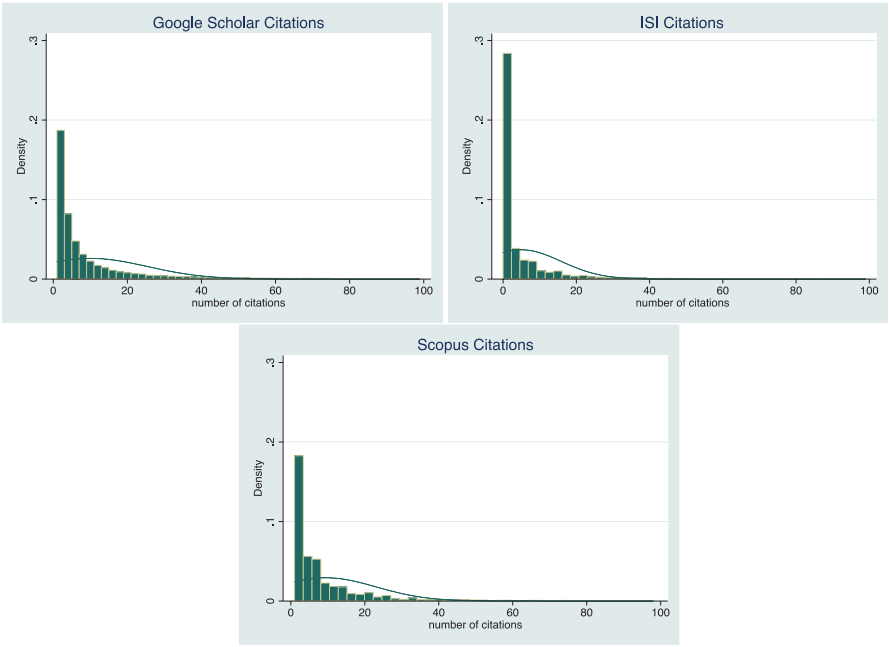

Fig. 8 Distribution of citations by bibliometric source

Table 7 Correlation matrix of citations

Correlation matrix	EVA	Scopus	ISI
EVA	1		
Scopus	0.8749*	1	
ISI	0.8510*	0.9171*	1

In general, it is very interesting to notice that the greater correlation between paper citations is associated with the publications that are simultaneously indexed in both ISI and Scopus (0.91), which account for a small fraction of the EVA products, but both refer to commercial resources. However, the correlations between citations received by papers collected simultaneously in Scopus and EVA as well as in ISI and EVA are both larger than 0.85. In particular, this applies to the economics and statistics disciplines. Finally, the slopes of the three lines in Fig. 9 evidence that the ratio between the citations received in Scholar and in ISI is lower (1/3) than the ratio of the citations collected from Scholar and Scopus (1/2).

Table 8 Correlation matrix of citations by field

Correlation Matrix	EVA	Scopus	ISI
Civil engineering and architecture			
EVA	1		
Scopus	0.8986*	1	
ISI	0.7139*	0.7280*	1
Antiquities, philological-literacy and historical arts sciences			
EVA	1		
Scopus	0.8533*	1	
ISI	0.9090*	0.7680*	1
History, philosophy, education and psychology			
EVA	1		
Scopus	0.9490*	1	
ISI	0.8526*	0.8312*	1
Law			
EVA	1		
Scopus	0.5713*	1	
ISI	0.7257*	0.6294	1
Economics and statistics			
EVA	1		
Scopus	0.8790*	1	
ISI	0.8603*	0.9206*	1
Political sciences			
EVA	1		
Scopus	0.9128*	1	
ISI	0.7207*	0.7747*	1

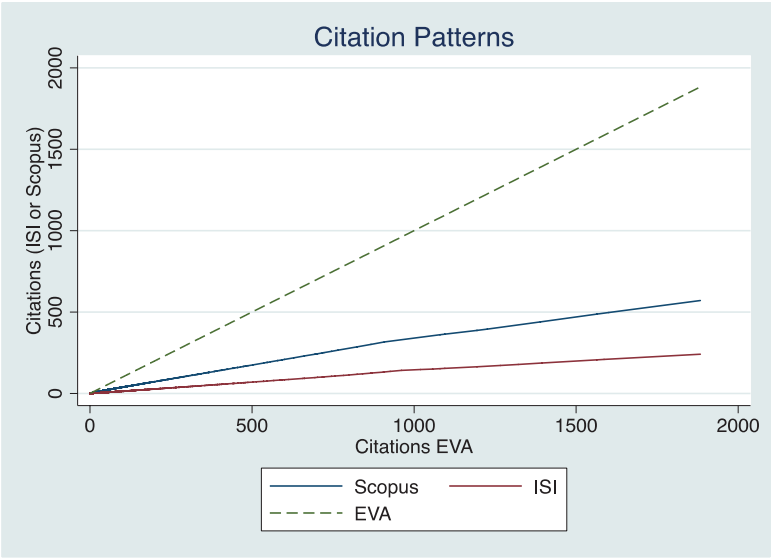

Fig. 9 Lotka Law by Bibliometric Source

7 Concluding Remarks

The main research question motivating the EVA project concerns the use of Google Scholar as a reference bibliographic database for those research areas, mainly humanities and social sciences, where the use of citation-based bibliometric indexes is not a common practice. The use of bibliometrics in general and the use of citations in particular, either for research assessment or just as a mean for achieving a better insight of the disciplines, is a main issue of discussion among social science and humanities researchers. Several authors raised epistemological objections motivated by the nature of the scientific communications in these areas and the kind of publications practices, such as the low proportion of journal articles, the literature post-publication citation rate and the local relevance of social sciences and humanities knowledge (Archambault and Larivière 2010; Prins et al. 2016). However, EVA focused the attention on a second set of objections, based on the idea that the coverage of these areas by the databases commonly used for hard sciences is largely inadequate to represent the scientific production of researchers operating in social sciences and humanities. The EVA results show that Google Scholar can be an alternative for assessing the scientific research in non-bibliometric areas, but only by accurately using suitable techniques for data analysis and quality verification. A first remarkable limit in using Google Scholar is due to the Google terms of service, which clearly state that Scholar is a service for searching data and it is not intended nor usable as a database for downloading data. This means that analysis must be limited to search results provided by Google Scholar online. As a consequence, Google Scholar is mainly usable for analysis of research at the individual level or when dealing with small groups of researchers such as Departments of small institutions. Larger collections of products can be analyzed as well, but this requires time for accessing records online and collecting complete analysis results. Moreover, the data access limitations require also to get as much information as possible from the search answer page provided by Scholar rather than by the complete publication record. A consequence of this limitation is that the information concerning the publication venue and type is very hard to achieve and it is not reliable is several cases. Despite these limits, however, it is indisputable that Google Scholar is an invaluable source of information for what concerns citations. Our results show that less than 10% of publications retrieved from Scholar are available on Scopus and less than 5% of ISI, with some remarkable differences among scientific areas. Proportionally, the number of citations from Scholar is also higher due to the fact that the number of citing publications indexed by the Scholar database is definitively larger than Scopus and ISI. From the statistical point of view, we observe that on large groups of publications there is a good level of correlation between citation distribution in the three databases. However, there are differences at the individual level. As a consequence, the picture we take of the scientific production of individuals and small groups from Google Scholar is completely different from Scopus or ISI in many cases and provides a more realistic insight of the scientific production of researchers, especially for humanities, law and some fields of political sciences.

Finally, we believe that in dealing with bibliometric analysis the crucial issue, even more decisive than having correct measures or indexes, is the quality of data. The main issue about data quality with Google Scholar is to correctly disambiguate the authorship of publication records. When searching for author names on Google Scholar, only the 23% of records retrieved can be correctly attributed to the author. Less than 10% of these can be easily disambiguated by relying on the author name format, but the majorities are due to real ambiguities due to homonymy, which makes the disambiguation process difficult. When working with tools for research assessment based on Google Scholar, this should be the main issue about the trustworthiness of the results. Solutions proposed in literature are mainly based on co-authorship relations which are rare and often untrustworthy when dealing with social sciences and humanities, where it is a common practice to publish work with no more than one or two authors. In EVA, we proposed a new solution based on the analysis of terminology, which takes into account both latent semantics and the specificity of terms with respect to scientific areas. The EVA techniques have been proved to be effective in disambiguating authorship in different areas independently from the publication language. Moreover the EVA system can be easily tuned for the purpose of achieving the required level of precision and recall. Our future work will be devoted to further improve disambiguation. The idea is to start from known publications of a given author in order to model language specificity at the level of single authors and not only of scientific areas. Such an improvement will be used to achieve disambiguation also for that authors working in different areas but on interdisciplinary fields.

References

Aguillo, I. F. (2012). Is Google scholar useful for bibliometrics? A webometric analysis. *Scientometrics, 91*(2), 343–351.

Archambault, É., & Larivière, V. (2010). The limits of bibliometrics for the analysis of the social sciences and humanities literature. *World Social Science Report*, 251–254.

Archambault, É., Vignola-Gagne, E., Côtè, G., Larivire, V., & Gingrasb, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics, 68*(3), 329–342.

Biolcati-Rinaldi, F., Ferrara, A., Pinotti, L., & Salini, S. (2012). Lesson learned by Unimival researchers during the comparative bibliometric analysis project (ABC). in *Rassegna Italiana di Valutazione*, v. XVI, n. 52, pp. 81-100, ISSN 1826-0713.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H. D. (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, 8(1), 93–102.

Burrell, L. Q. (2006). Measuring concentration within and co-concentration between informetric distributions: An empirical study. *Scientometrics, 68*(3), 441–456. https://doi.org/10.1007/s11192-006-0122-0.

Checchi, D., De Fraja, G., & Verzillo, S. (2014). Publish or Perish: An Analysis of the Academic Job Market in Italy. Tech. Rep. Discussion Paper 10084, CEPR Discussion Paper.

D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology, 62*(2), 257–269.

Daniel, H. D., & Fisch, R. (1990). Research performance evaluation in the german university sector. *Scientometrics, 19*(5), 349–361. https://doi.org/10.1007/BF02020698.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1).

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, web of science, and Google scholar: Strengths and weaknesses. *The FASEB Journal, 22*(2), 338–342.

Ferrara, A., & Salini, S. (2012). Ten challenges in modeling bibliographic data for Bibliometric analysis. *Scientometrics, 93*(3), 765–785.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. (2010). Effective self-training author name disambiguation in scholarly digital libraries. In *Proceedings of the 10th annual ACM joint conference on digital libraries* (pp. 39–48). Aarhus.

Garfield, E. (1980). Is information retrieval in the arts and humanities inherently different from that in science? The effect that ISI©'S citation index for the arts and humanities is expected to have on future scholarship. *The Library Quarterly*, 40–57.

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th joint ACM/IEEE conference on digital libraries (JCDL 2004)* (pp. 296–305). Tucson.

Han, H., Xu, W., Zha, H., & Giles, L. (2005a). A hierarchical naive Bayes mixture model for name disambiguation in author citations. In *Proceedings of the ACM symposium on applied computing* (pp. 1065–1069). Santa Fe.

Han, H., Zha, H., & Giles, L. (2005b). Name disambiguation in author citations using a Kway spectral clustering method. In *Proceedings of the 5th joint ACM/IEEE conference on digital libraries (JCDL 2005)* (pp. 334–343). Denver.

Kousha, K., & Thelwall, M. (2007). Google scholar citations and Google web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology, 58*(7), 1055–1065.

Linmans, A. J. (2010). Why with bibliometrics the humanities does not need to be the weakest link. *Scientometrics, 83*(2), 337–354.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge: Cambridge University Press.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66*(1), 81–100.

On, B. W., & Lee, D. (2007). Scalable name disambiguation using multi-level graph partition. In *Proceedings of the SIAM International conference on data mining* (pp. 575–580). Minneapolis, Minnesota.

Prins, A. A., Costas, R., van Leeuwen, T. N., & Wouters, P. F. (2016). Using google scholar in research evaluation of humanities and social science programs: A comparison with web of science data. *Research Evaluation, 25*(3), 264–270.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology, 43*(1), 1–43.

Tang, J., Fong, A. C. M., Wang, B., & Zhang, J. (2012). A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering, 24*(6), 975–987.

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology, 56*(2), 140–158.

Yang, K. H., Peng, H. T., Jiang, J. Y., Lee, H. M., & Ho, J. M. (2008). Author name disambiguation for citations using topic and web correlation. In *Proceedings of the 12th European conference on digital libraries* (pp. 185–196). Aarhus, Denmark.