Which conference is that? A case study in computer science*

CAMIL DEMETRESCU, Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Italy
IRENE FINOCCHI, Department of Business and Management, Luiss Guido Carli University, Italy
ANDREA RIBICHINI, Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Italy
MARCO SCHAERF, Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Italy

Conferences play a major role in some disciplines such as computer science and are often used in research quality evaluation exercises. Differently from journals and books, for which ISSN and ISBN codes provide unambiguous keys, recognizing the conference series in which a paper was published is a rather complex endeavor: there is no unique code assigned to conferences and the way their names are written may greatly vary across years and catalogs. In this article, we propose a technique for the entity resolution of conferences based on the analysis of different semantic parts of their names. We present the results of an investigation of our technique on a dataset of 42395 distinct computer science conference names excerpted from the DBLP computer science repository[1], which we automatically link to different authority files. With suitable data cleaning, the precision of our record linkage algorithm can be as high as 94%. A comparison with results obtainable using state-of-the-art general-purpose record linkage algorithms rounds off the paper, showing that our *ad hoc* solution largely outperforms them in terms of the quality of the results.

Just Accepted

1 INTRODUCTION

In recent years, bibliometric evaluation exercises of the published scientific corpus of universities, research institutes, departments, and individuals have been adopted by policy makers to inform funding allocation, hiring, and promotions. To do so, evaluators base their assessment on bibliographic data excerpted from major catalogs such as Scopus and Web of Science. Since they include widely adopted evaluation indicators such as the number of citations, evaluation committees often base their evaluation on the quality of the publication venues, which is maintained in external authority files. This requires matching a publication from the bibliographic catalogs to the corresponding publication venue in the authority file. While this is straightforward for journal articles, books, and book chapters due to the presence of unambiguous ISSN and ISBN keys, the situation for conferences is instead rather complex due to the lack of standardized keys. Hence, given the DOI of a publication, catalog lookup can provide ISSN or ISBN codes for journal articles and books, but just a string with the free-text name of the conference edition in which it appeared. For instance, a certain conference may appear in the catalog as

> STOC 2019: Theory of Computing, Proceedings of the 51st Annual ACM SIGACT
> Symposium on

and in the authority file as:

*Work partially supported by MIUR, the Italian Ministry of Education, University and Research, under PRIN Project n. 20174LF3T8 AHeAD (Efficient Algorithms for HArnessing Networked Data).

[1] https://dblp.org/

Authors' addresses: Camil Demetrescu, demetres@diag.uniroma1.it, Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Via Ariosto 25, 00185, Rome, Italy; Irene Finocchi, finocchi@luiss.it, Department of Business and Management, Luiss Guido Carli University, Viale Romania 32, 00197, Rome, Italy; Andrea Ribichini, ribichini@diag.uniroma1.it, Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Via Ariosto 25, 00185, Rome, Italy; Marco Schaerf, marco.schaerf@uniroma1.it, Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Via Ariosto 25, 00185, Rome, Italy.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.

1936-1955/2022/3-ART $15.00
https://doi.org/10.1145/3519031

Annual ACM Symposium on Theory of Computing (STOC)

We note that usual string similarity metrics such as edit distance or Jaccard distance are likely to provide poor results due to the large syntactic gap between the two forms.

Contributions of the article. In this article, we propose a method for automatically linking conference names as they appear in prominent catalogs, such as DBLP, to an authority file, such as those derived from DBLP itself and to the GII-GRIN-SCIE list adopted by the Italian Ministry of Education, University, and Research for a number of publication assessment exercises. The GII-GRIN-SCIE authority file was derived by the Microsoft Academic, LiveSHINE, and the Australian CORE conference classifications. Differently from DBLP, the GII-GRIN-SCIE authority file features conference rankings that can be used for research quality assessments.

Our key idea is to convert each conference in the authority file into a vector of features that represent semantic properties of the conference such as sponsor (e.g., ACM), region (e.g., Asian), conference type (i.e., workshop, symposium, conference, etc.), and core (i.e. Theory Computing). We show that our semantic decomposition of the authority file allows for a structured matching of key attributes, ignoring irrelevant features in the conference name.

We perform an extensive evaluation of our approach, showing that the proposed technique is highly sensitive to the quality of the authority file. We show that by manually cleaning the DBLP authority file we can obtain an increase of the precision from 67% to 94%. We also present the results of an evaluation where we link the GII-GRIN-SCIE authority file with the DBLP authority file allowing the classification of conference publications. Finally, we compare our method with state-of-the-art entity resolution toolkits such as Dedupe [7], showing that we can outperform them in terms of the quality of the obtained linking.

Structure of the article. The remainder of this article is organized as follows. Section 2 addresses related work and in Section 3 we discuss our bibliometric datasets. Section 4 focuses on our approach to record linkage of conference names, Section 5 present the results of our evaluation, and Section 6 concludes the article.

2 RELATED WORK

The increased interest in recent years in bibliometric assessments relies on the availability of high-quality bibliometric data as a basis of the analysis (e. g. Ferrara and Salini [6]). Many research-assessment exercises, such as the Italian VQR ("Valutazione della Qualità della Ricerca") are based on both article and publication-venue ranking. In some research areas, most notably Computer Science (CS), conferences are highly-considered publication venues. Hence, it is critical to be able to rank conferences and to uniquely identify the publication venue of papers presented at conferences.

Several recent works have stressed the importance of considering conferences in the assessment of the scientific production in CS. In particular, Vrettas and Sanderson [15] investigate the relative importance of journals and conferences in CS, Almendra et al. [1] propose a methodology to cluster conferences in scientific areas and refine their ranking, while Lee and Brusilovsky [8] analyze the impact of conference publications on overall citation counts.

The need for clean and polished data has always been an issue in many studies in bibliometrics. As an example, the accuracy of names in bibliographic data sources is investigated by Demetrescu et al. [5]. Similarly, in order to study the use of bibliometrics in the CS field in Italy, in [4] names of publication venues extracted from Scopus had to be matched against conferences and journals in a variety of lists. This required to face a matching problem akin to the one addressed in this work. The heuristic procedure used in [4] was based on a multi-pass greedy approach: entries that were matched incorrectly at each phase were reinserted into the pool of unmatched ones for further processing in later phases. This required, however, significant manual intervention.

2.1 Entity Resolution

Entity Resolution (ER) is the task of identifying which entities represent the same object. It is a widely studied field in the scientific literature and many techniques have been developed to solve large real-world instances of this problem. As an example, Papadakis et al. [11] review the most relevant literature on blocking and filtering techniques applied to ER. Entity resolution comes in two main forms, that are:

- Deduplication: when we have one single set of entities and we want to identify which ones are duplicate ones (i. e. they represent the same object)

Just Accepted

- Record Linkage (RL): when we have two separate lists of entities and we want to find out the correspondences between elements of the two sets. In this article, we will be mostly concerned with RL problems.

ER is inherently a quadratic problem, since it requires to compare each pair of possible solutions, it does not scale gracefully when the size of the problem grows. In order to contain the computational complexity, it is important to reduce the number of possible solutions and improve the efficiency of the resolution step. The blocking techniques aim at clustering the entities so that only entities belonging to the same cluster are compared, thus reducing the number of potential solutions, while filtering techniques analyze the elements in the same cluster to identify the true equivalences. Advanced blocking techniques have been introduced by Papadakis et al. [10] and Tao and Kong [14].

There are several previous attempts at using ER techniques in the context of bibliographic data, such as the work of de Jesus and Pereira [3], Paraizo and Pereira [12], Pereira et al. [13], and the work of Wu et al. [17]. Pereira et al. [13] set the foundation to define an authority file for publication venues in the style of the Virtual International Authority File (VIAF www.viaf.org), while [3] extend the framework by enriching it with data acquired on the web. Finally, Paraizo and Pereira [12] present the PVAF System that is “A publication venue authority file stores variants of the names of journals and conferences that publish scientific articles”.

This work has been used by Wu et al. [17] where they work on the entity resolution of articles coming from different datasets, in their case DBLP and CiteSeerX, where IEEE Explore is used as ground truth (for a subset of articles).

3 BIBLIOMETRIC DATASETS

In this section we introduce the terminology and describe the datasets used in our experimental study. Throughout the paper, we will use the general term *conference* to refer to scientific meetings that take place regularly (e.g., yearly or every two years), including workshops, symposia, colloquia, congresses. Each conference has its own *name*, which might change occasionally over time, and possibly many *editions*.

We used two lists of conferences, called *authority files* in the remainder of this article:

- The 2018 GII-GRIN-SCIE conference list has been developed as a joint initiative sponsored by GII (Group of Italian Professors of Computer Engineering), GRIN (Group of Italian Professors of Computer Science), and SCIE (Spanish Computer Science Society). The list includes 2831 conferences, rated according to different classes, and is available at http://www.consorzio-cini.it/gii-grin-scie-rating.html. Each item in this list contains the conference title, its acronym, and different ratings either computed by GII-GRIN-SCIE or obtained from other public sources.
- The DBLP conference list, gently provided as a personal communication by Florian Reitz. This list contains 6533 conference records. Each record is typically associated with a unique key, an acronym, and a title. Other information may be possibly available, but not always, such as the former name or a different name for the same conference, a URL, a publisher for the proceedings, and a reference to a larger event that the conference is part of.

Conference editions. The list of conference editions used in our study has been obtained from the DBLP Computer Science Bibliography, which provides open bibliographic information on major computer science journals and proceedings. The list has been extracted from a dataset available at the location http://dblp.uni-trier.de/xml/ and contains 42395 entries. Each entry includes a distinct conference edition name (proceedings name) along with the year of the edition and a key in DBLP. The key was assigned via a manual entity resolution process by the DBLP maintainers as reported by Ley [9].

In the following, we provide a few examples of the contents of the different lists, also highlighting some common issues arising in the linking process.

Example 1: OOPSLA. The conference title in the DBLP authority file is “ACM SIGPLAN International Conference on Object-Oriented Programming Systems, Languages, and Applications” and its DBLP key is *oopsla*. We notice, however, that the same DBLP key is associated with 11 other satellites events. Moreover, the same conference has a slightly different name in other authority files. For instance, in GII-GRIN-SCIE, OOPSLA appears as “ACM Conference on Object Oriented Programming Systems Languages and Applications.”

Just Accepted

The DBLP conference edition list contains 96 items with key *oopsla*. Out of them, only 31 correspond to the main conference, the others being satellites events, whose number increased especially in the last few years. Titles associated with different OOPSLA editions can be rather different within the list. For instance:

- The 1994 edition appears in the list as

  OOPSLA'94, Proceedings of the Ninth Annual Conference on Object-Oriented Programming Systems, Languages, and Applications, Portland, Oregon, USA, October 23-27, 1994

  Most notably, the acronym, followed by an abbreviation of the year, appears at the beginning, sponsors are not specified, the ordinal number of the conference edition is written in letters, the conference location appears before the date.
- The 2009 edition appears in the list as

  Proceedings of the 24th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications, OOPSLA 2009, October 25-29, 2009, Orlando, Florida, USA

  In this case the ordinal number of the edition is provided using digits followed by suffix "th", the acronym is in the middle, the sponsor is provided, the conference location appears after the date.
- The 2015 edition appears in the list as

  Proceedings of the 2015 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications, OOPSLA 2015, part of SPLASH 2015, Pittsburgh, PA, USA, October 25-30, 2015

  In this case, the edition is not given through an ordinal number but is specified by its year, sponsors are provided, the acronym is in the middle, and the name of a larger event that OOPSLA is part of is also given.

Example 2: WSE. The DBLP authority file contains two different conference titles associated with key *wse*, because the conference name and the type of the event changed over time: the base title is "Symposium on Web Systems Evolution" and the former name is "Workshop on Web Site Evolution". Moreover, starting from the 7^{th} edition in 2005, the event became IEEE-sponsored. The event was also upgraded from workshop to symposium at its 9^{th} edition in 2009, where the keyword "Web site" was replaced by "Web systems". As a result, titles in the conference editions can be rather different. For instance, 2001, 2003, 2005, 2007, and 2011 editions appear in the list as

- *3rd International Workshop on Web Site Evolution (WSE 2001) - Access for All, 10 November 2001, Florence, Italy*
- *5th International Workshop on Web Site Evolution (WSE 2003) - Architecture, 22 September 2003, Amsterdam, The Netherlands*
- *Seventh IEEE International Workshop on Web Site Evolution (WSE 2005), 26 September 2005, Budapest, Hungary*
- *Proceedings of the 9th IEEE International Symposium on Web Systems Evolution, WSE 2009, 5-6 October 2007, Paris, France*
- *13th IEEE International Symposium on Web Systems Evolution, WSE 2011, Williamsburg, VA, USA, September 30, 2011*

Notice that in 2001 and 2003 the main topic of the workshop was also included as part of the title ("Access for all" and "Architecture", respectively). Other major differences include the presence of a sponsor (IEEE), the use of ordinal numbers written in letters (e.g., "Seventh" vs "9th"), the use of different event types ("Workshop" vs. "Symposium"), the acronym given between commas or in parentheses, the presence of "Proceedings of", the different positions and structure of conference dates and location, as well as the different keywords in the title ("Site" vs. "Systems"). These differences make it rather difficult to associate different editions to the same main event, looking only at the titles, by using string matching algorithms.

4 APPROACH

The linking problem addressed in this article can be summarized as follows: given a conference edition, which is the conference record associated to that event? Our approach is to compute a *score* for each pair (e, c), where e is an edition and c is a conference record in the authority file. The higher the score, the most likely is the match between edition e and conference c. Our algorithm picks as a match the conference record c that yields the highest score, according to the following formula:

$$best\_match(e) = \begin{cases} unmatched & \text{if } max_c\{score(c,e)\} = 0 \\ argmax_c\{score(c,e)\} & \text{otherwise} \end{cases}$$

Just Accepted

If edition e cannot be matched with any conference record c (i.e., for all records c, $score(c, e) = 0$), the function returns *unmatched*. If there is more than one conference record yielding the maximum score, one of them is arbitrarily chosen as the best match.

Our score function is presented in Section 4.3, after describing text normalization (Section 4.1) and conference descriptors (Section 4.2) that will be crucial in the score computation.

4.1 Text normalization

Before computing match scores, the strings corresponding to conference editions and conference titles are normalized by changing the text to uppercase letters, removing diacritical marks, replacing quotes, apostrophes, dashes, parentheses, punctuation marks, and slashes with white spaces. Next, separators (i.e., consecutive occurrences of white spaces) are collapsed, stopwords (such as articles, conjunctions, and prepositions) and escape sequences (e.g., &amp) are dropped. We also remove all the trailing lists of tokens starting with expressions such as *part of*, *held with*, *in conjunction with*, *colocated with* or *collocated with*.

4.2 Conference descriptors and authority file preprocessing

From each conference record c in the authority file, we automatically extract five *conference descriptors*, i.e., acronym, regions, qualifiers, sponsors, and core. Each descriptor represents a semantically different aspect of the conference record. The acronym is a shorthand for the conference and in most cases is explicitly represented in the authority file. The remaining descriptors are extracted from the conference title. Sponsors and regions obtained from the title are matched against domain-specific lists which include, e.g., ACM, IEEE, IFIP, SIAM, AAAI, USENIX, DIMACS for sponsors and Asian, Australian, Canadian, European, German, Italian for regions. The qualifier describes the different nuances of conference types, e.g., symposium, meeting, congress, workshop, seminar, summit, summer school, as well as characterizations such as annual or international. Whatever remains in the conference title after extracting regions, sponsors, and qualifiers is part of the conference name core.

Example 3: OOPSLA. Let us consider Example 1 of Section 3, related to a conference with the title "ACM SIGPLAN International Conference on Object-Oriented Programming Systems, Languages, and Applications". Besides the acronym OOPSLA, we extracted from the title "ACM SIGPLAN" as a sponsor, "INTERNATIONAL CONFERENCE" as a qualifier, "OBJECT ORIENTED PROGRAMMING SYSTEMS LANGUAGES APPLICATIONS" as the core. There is no region for this conference.

4.3 Score computation

In addition to the conference descriptors introduced in Section 4.2, our score computation algorithm uses two auxiliary string analysis functions, called `loose_match` and `strict_match`, respectively. Both functions take as input two strings X and Y, which are first normalized as described in Section 4.1 and then tokenized.

- Function `loose_match(X,Y)` counts the number of tokens of X that appear as tokens in Y, not necessarily in the same order or in consecutive positions. If X is empty, the function returns 0. This function is used when looser similarities between conference descriptors and tokens in a conference edition string are looked for.

  Example. If X = "MANAGED PROGRAMMING LANGUAGES RUNTIMES" and Y = "OBJECT ORIENTED PROGRAMMING SYSTEMS LANGUAGES APPLICATIONS", the output of function `loose_match(X,Y)` is 2, due to the tokens "PROGRAMMING" and "LANGUAGES".
- Function `strict_match(X,Y,D)` receives two strings X and Y and an integer D that is used as edit distance between the two input strings. The function checks whether all tokens of X appear as tokens in Y, in the same order but not necessarily in consecutive positions, and with an edit distance – computed across all tokens – smaller than or equal to D. If this is the case, the function returns the number of tokens in X. Otherwise, or if X is empty, the output is 0. Differences for the edit distance computation are at a character level and may be spread across different tokens. This function is used to check relatively accurate correspondences between conference descriptors and tokens in a conference edition.

  Example. If X = "PROGRAMMING SYSTEM APPLICATION" and Y = "OBJECT ORIENTED PROGRAMMING SYSTEMS LANGUAGES APPLICATIONS", function `strict_match(X,Y,2)` returns 3, because the three tokens of X appear in Y in the same order. Notice that tokens "SYSTEM" and "APPLICATION" have both edit distance 1 from the corresponding tokens in Y. Hence, the total edit distance across all tokens of X is 2, which is allowed

Just Accepted

```
   function score(conference record $c$, conference edition $e$)
1.    if strict_match($core(c), e, 1$) = 0
2.    then return 0
3.    else return $C \cdot$ strict_match($core(c), e, 1$)
4.              + $A \cdot$ strict_match($acronym(c), e, 0$)
5.              + $S \cdot$ loose_match($sponsor(c), e$)
6.              + $R \cdot$ loose_match($regions(c), e$)
7.              + $Q \cdot$ loose_match($qualifier(c), e$)
```

Fig. 1. Score function: the returned numeric value is obtained by analyzing the conference descriptors. If no token of $core(c)$ is included in e, even at edit distance 1, the function returns 0.

by the D value used in the function invocation. Instead, call `strict_match(X,Y,1)` would return 0 because the edit distance constraint is not satisfied.

The score function is shown in Figure 1. It computes a numeric value by analyzing the conference descriptors. The core of conference c must have at least one token at edit distance ≤ 1 from the tokens in edition e, otherwise 0 is returned. Common tokens must appear in the same order in $core(c)$ and e, hence function `strict_match` is used at line 3. The edit distance value is quite restrictive and is set to 1. A similar inclusion property must hold for the acronym, which however must be exactly equal to a sequence of consecutive tokens in e (i.e., the edit distance should be 0). Sponsors, regions, and qualifiers, if present, might not necessarily be complete (i.e., some tokens might be missing). Also, the tokens may appear in a different order in different conference editions. Hence, we check inclusion using function `loose_match`. These issues are illustrated in the following examples.

Example. Consider the DBLP authority file entry with title "ACS/IEEE International Conference on Computer Systems and Applications" and key *aiccsa*. The DBLP conference edition list contains 14 items with this key. Titles associated with different AICCSA editions differ in several respects. For instance:

- The 2005 edition appears in the list as

  2005 ACS / IEEE International Conference on Computer Systems and Applications (AICCSA 2005), January 3-6, 2005, Cairo, Egypt

- The 2009 edition appears in the list as

  The 7th IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2009, Rabat, Morocco, May 10-13, 2009

In these two examples, conference sponsors are separated differently (this is taken care of by our string normalization procedure) and, most notably, they appear in a different order. Transforming one sponsor sequence into the other would require 8 edit operations, making the use of `strict_match` not advisable in this case (used with such large edit distance, this function would result in many false positive matches). On the other hand, `loose_match` allows us to capture token inversions easily.

Example. Consider the DBLP authority file entry with title "International/Italian Conference on Algorithms and Complexity" and key *ciac*. In this case, the title tries to capture the fact that the conference changed denomination in 2010, switching from "Italian" to "International" conference. Our preprocessing extracts "Italian" as region and "International Conference" as qualifier.

The DBLP conference edition list contains 10 items with key *ciac*. Consider for instance the following entries:

- The 1994 edition appears in the list as:

  Algorithms and Complexity, Second Italian Conference, CIAC '94, Rome, Italy, February 23-25, 1994, Proceedings

- The 2013 edition appears in the list as:

  Algorithms and Complexity, 8th International Conference, CIAC 2013, Barcelona, Spain, May 22-24, 2013. Proceedings

Just Accepted

In the first case, our algorithm finds a match for the region and a partial match for the qualifier, while in the second case it finds a full match for the qualifier. Had we used `strict_match` rather than `loose_match` to check for qualifiers, the partial match would have been missed.

In summary, `strict_match` is used when a higher degree of similarity is needed in order to declare a match, while `loose_match` detects less tight resemblances.

A weighted sum of the scores obtained from the five conference descriptors is returned. In order to set the weights, a preliminary assumption was made to regard a match of the acronym as most significant: our preliminary tests indicated indeed the acronym to be the most reliable conference feature. This was followed by the conference core, and then by sponsors, regions, and qualifiers, which may not always be present. The final weights used in our implementation were determined empirically through a process of trial and error and are $C = 100$, $A = 200$, $S, R, Q = 30$. In particular, we tried different sets of parameters with the goal of maximizing the number of correct matches. We started from the initial set of parameters $C = 100$, $A = 150$, $S, R, Q = 5$ and we iterated by attempting to follow the gradient of the number of correct matches. We ended up in the configuration used in our implementation, which is a (possibly local) maximum.

4.4 Limitations and extensions

Our algorithm does not cover cases where abbreviated entries have to be matched. For instance, the entry "SIGIR Conference" is matched to "Annual International ACM SIGIR Conference on Research and Development in Information Retrieval" with a low score given that no token of the core "RESEARCH DEVELOPMENT INFORMATION RETRIEVAL" appears in the query and the matching only succeeds with the acronym (SIGIR) and the sponsor (ACM SIGIR).

Another case where abbreviations are difficult to be matched occurs with abbreviated tokens such as "Int. Conf. Soft. Eng.". In such cases, the matching algorithm may be relaxed to work with prefixes rather than with entire tokens. In the example above "Int." may be matched to "INTERNATIONAL", "conf." to "CONFERENCE", etc.

One more limitation is concerned with changes in the authority file that occur over time. New conferences may appear or existing conferences may change their names. For our approach to work correctly, we assume that the authority file is kept up to date, but we do not address how this can be done and we consider the authority file itself just as an input to our record linkage algorithm.

5 EVALUATION

In this section we present an evaluation of our record linkage algorithm on both the DBLP and the GII-GRIN-SCIE authority files. We also compare our results with those of general-purpose state-of-the-art record linkage algorithms.

We consider both positives (true and false), i.e., conference names matched to entries of the authority file and negatives (true and false), i.e., conference names that our algorithm declared as unmatched to the DBLP authority file. We recall that, by denoting with tp, fp, tn, fn the true positives, false positives, true negatives, and false negatives, respectively, we have that the *precision* is defined as $tp/(tp + fp)$, the *recall* as $tp/(tp + fn)$, and the *accuracy* as $(tp + tn)/(tp + fp + tn + fn)$.

5.1 Linkage of conference names to the DBLP authority file

As a first experiment, we describe our evaluation of record linkage using 42,395 conference names ("Proceedings" XML entities) excerpted from the DBLP repository. We applied the approach discussed in Section 4 by attempting to link each conference name to the DBLP authority file.

5.1.1 DBLP authority file cleaning. The initial run we performed indicated a precision of about 67.7% for our algorithm: namely, out of 43,218 conference names, we obtained 28,140 true positives, 14,255 false positives, 740 true negatives, and 83 false negatives. This leads to a recall and accuracy of 99.7%, and 68.1%, respectively.

A detailed analysis of the data revealed that most errors were due to issues in the authority file. For this reason, we focused on a random sample of 667 conference names (10%) along with their linked entries in the DBLP authority file and we manually validated them to check for true and false positives. We found out that there are a number of reasons for which the original authority file led to poor precision. Some of them are listed below.

- *Cumulative workshops co-located with a conference*: for instance, "IEEE International Enterprise Distributed Object Computing Conference Workshops". This is a common issue that we fixed by adding entries for individual workshops to the authority file.

Just Accepted

DBLP scenario	**Algorithm**	tp	fp	tn	fn	**Precision**	**Recall**	**Accuracy**
Before cleaning	This article	435	219	11	2	66.5%	99.5%	66.9%
Before cleaning	[7]	352	94	0	221	78.9%	61.3%	66.7%
After cleaning	This article	628	39	0	0	94.2%	100.0%	94.2%
After cleaning	[7]	425	142	0	100	74.8%	80.9%	63.6%

Table 1. Evaluation of the algorithms considered in this study on DBLP. In the table, tp, fp, tn, fn denote true positives, false positives, true negatives, and false negatives, respectively. The *precision* is defined as $tp/(tp + fp)$, the *recall* as $tp/(tp + fn)$, and the *accuracy* as $(tp + tn)/(tp + fp + tn + fn)$.

- *Tracks of a conference*: for instance, SPLASH has general tracks *OOPSLA*, *Onward!* and *Wavefront*. We fixed this issue by adding separate entries in the authority file for different tracks of the same conference.
- *Grant names*: some entries of the DBLP authority file contain umbrella names of grants, e.g., "EU Projects". We regard these types of issues as unsolvable with our proposed approach as they would be critical even for a human validator.
- *Multiple conference names in the same entry*: this is an issue of conference names rather than of the authority file itself. Solving the issue would require linking the name to multiple entries of the authority file, which is beyond the scope of our algorithm being a relatively infrequent case.
- *Acronyms instead of full conference names*: for instance, one of the names to be matched was "HCI International 2017 - Posters' Extended Abstracts - 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part II". We regard these types of issues as unsolvable with our proposed approach.
- *Same conference changed name*: this is a relatively frequent case that we fixed by adding extra entries for the same entity of the authority file. For instance, as anticipated in Section 3, "Symposium on Web Systems Evolution" was formerly "Workshop on Web Site Evolution".

Our manual validation led to a new version of the authority file where we added entries for workshops and co-located events, different names for the same conference, and other fixes described above. Since it was guided by a sample of conference names, our manual validation is by no means to be considered leading to a generally valid authority file. However, the data cleaning was instrumental in assessing the precision that one would get using our approach starting with a complete and accurate authority file of conferences. The cleaning was put in place to assess the garbage-in garbage-out phenomenon in our record linkage.

We chose to rely on manual data cleaning because we wanted our authority file to be as accurate as possible (within the limits of human error), so that the performance of our approach could be studied without any biases. However, nothing ties our techniques to a specific data cleaning method, and automatic cleaning procedures (see, e.g., Wang et al. [16]), could easily be integrated into our approach in order to meet the demands of specific applicative scenarios.

5.1.2 Results. Our results are reported in Table 1. After the data cleaning described in Section 5.1.1, the precision of the record linkage with our algorithm on the considered random sample of 667 conference names (10%) raised from 68% to 94%. False positives, which account for about 6%, were due to the unavoidable issues that we discussed in Section 5.1.1. We remark that our first run was based on the entire set of conference names, while our run after data cleaning was performed on the manually validated subset of conference names. On the full set of conference names and the original authority file, our algorithm featured a recall of about 97%. Since the second run after data cleaning was performed on a subset of matched conference names, the number of resulting negatives was zero.

5.2 Linkage of conference names to the GII-GRIN-SCIE authority file

In this section we discuss a second set of experiments where we attempt to link conference names to the GII-GRIN-SCIE authority file. The overall approach we adopted was to link the GII-GRIN-SCIE file to the DBLP file using our algorithm, hence transitively linking the conference names we experimented with in Section 5.1 to GII-GRIN-SCIE. The advantage of the GII-GRIN-SCIE authority file is that it assigns a ranking to a large number of conferences, which can be used when evaluating the quality of research. In particular, conferences in the GII-GRIN-SCIE file are

Just Accepted

GII-GRIN-SCIE Rank	Algorithm	tp	fp	tn	fn	Precision	Recall	Accuracy
Class 1	This article	70	9	0	0	88.6%	100.0%	88.6%
Class 2	This article	159	12	0	0	92.3%	100.0%	92.3%
Class 3	This article	320	25	2	0	92.7%	100.0%	92.8%

Table 2. Evaluation of our algorithm when linking DBLP and GII-GRIN-SCIE authority file entries. In the table, tp, fp, tn, fn denote true positives, false positives, true negatives, and false negatives, respectively. The *precision* is defined as $tp/(tp + fp)$, the *recall* as $tp/(tp + fn)$, and the *accuracy* as $(tp + tn)/(tp + fp + tn + fn)$.

partitioned into five sets: classes 1, 2, and 3, [WP] (work in progress) and [NR] (not ranked). As our aim was to assign ranks to conference editions (whenever possible), we focused on GII-GRIN-SCIE entries belonging to classes 1, 2 or 3.

Contrary to the experiment described in Section 5.1, in this case, we could not rely on a set of common keys to provide a ground truth against which to assess the quality of our results. We therefore resorted to conference acronyms as a proxy for uniquely identifying keys, complementing it with manual validation (exhaustive for class 1 entries, and sample-based for classes 2 and 3). Our experimental results, after filtering out a few sporadic cases that proved ambiguous even for human experts, are reported in Table 2. The results are good in terms of quality of the solutions but slightly lower than conference matches as discussed in Section 5.1.2. The main reason lies in a non-complete overlap between the two authority files.

5.3 Comparison with state-of-the-art algorithms

In this section, we report on the applicability of state-of-the-art record linkage (or deduplication) frameworks. We concentrate our investigation on two tools:

- Python Record Linkage Toolkit by de Bruin [2]: a well-known Python package for record linkage that implements a number of string matching algorithms;
- Dedupe by Gregg and Eder [7]: a state-of-the-art and commercial tool (https://dedupe.io/) frequently used in either record linkage and deduplication tasks. Dedupe uses cutting-edge machine learning techniques to identify matches.

We tested the Python Record Linkage Toolkit with 4 algorithms (jaro, jarowinkler, levenshtein, and damerau-levenshtein) and several thresholds. Even in the best combination (jarowinkler with a 0.75 threshold), the recall was never higher than 12%. For this reason, we considered another toolkit and started looking at Dedupe [7], a state-of-the-art machine learning tool for deduplication of entries.

Dedupe is an interactive tool that, first analyzes the two sets of entries, and then executes the training phase. As described in the "Active Learning" section of the documentation https://docs.dedupe.io/en/latest/Matching-records.html:

> *Basically, Dedupe keeps track of bunch unlabeled pairs and whether*
> - *the current learning blocking rules would cover the pairs;*
> - *the current learned classifier would predict that the pairs are duplicates or are distinct.*
>
> *We maintain a set of the pairs where there is disagreement: that is pairs which classifier believes are duplicates but which are not covered by the current blocking rules, and the pairs which the classifier believes are distinct but which are blocked together. Dedupe picks, at random from this disagreement set, a pair of records and asks the user to decide. Once it gets this label, it relearns the weights and blocking rules. We then recalculate the disagreement set.*

Dedupe claims to be able to compute the correct record linkage when it is presented with at least 10 positive and 10 negative examples. We trained the system with data of different sizes and we noticed that the performance tends to stabilize quickly without any significant improvement when the training dataset becomes larger. We thus report the case with 16 positive examples and 12 negative ones in the before cleaning case and 26 positive examples and 12 negative examples in the after cleaning case. The comparison with Dedupe is summarized in Table 1, which shows that our algorithm always achieves higher accuracy and recall. Notice that, using Dedupe, the precision is higher before the cleaning, but with a significantly lower recall value, while with our algorithm, described in Section 5.1.1, the reverse is true.

Just Accepted

We also experimented using Dedupe to solve the matching of the GII-GRIN-SCIE authority file to the DBLP one. The overall results were rather disappointing: Dedupe was only able to classify one fourth of the elements in the GII-GRIN-SCIE list in the classes 1, 2 and 3. In our opinion, this is due to the presence of instances in the two lists that are syntactically very similar but refer to different conferences. This similarity makes Dedupe very cautious about linking records in the two lists, thus leading to a low recall value.

6 CONCLUDING REMARKS

In this article we have considered the problem of linking conference names to authority files of conferences in the computer science domain. In our study, we have taken into account two authority files: 1) the prominent DBLP Computer Science Bibliography and 2) the GII-GRIN-SCIE list, created as a joint initiative sponsored by GII (Group of Italian Professors of Computer Engineering), GRIN (Group of Italian Professors of Computer Science), and SCIE (Spanish Computer Science Society). The main motivation of our work is to develop tools that can assist research quality assessment campaigns in areas where conferences, whose names are often difficult to identify correctly, play a major role such as in computer science. We have designed and implemented a record linkage algorithm especially tailored to conference names, testing it on both authority files. Our experimental results highlight that the quality of the outcome greatly depends on the quality of the adopted authority files.

REFERENCES

[1] Vinicius Da Silva Almendra, Denis Enachescu, and Cornelia Enachescu. 2015. Ranking computer science conferences using self-organizing maps with dynamic node splitting. *Scientometrics* 102, 1 (2015), 267–283.

[2] Jonathan de Bruin. 2016. Python Record Linkage Toolkit. https://recordlinkage.readthedocs.io/en/latest/index.html. Accessed April 2021.

[3] H.A. de Jesus and D.A. Pereira. 2017. Enriching an authority file of scientific conferences with information extracted from the web. *Journal of Computer Science* 13, 4 (2017), 68–77. https://doi.org/10.3844/jcssp.2017.68.77

[4] C. Demetrescu, I. Finocchi, A. Ribichini, and M. Schaerf. 2020. On bibliometrics in academic promotions: a case study in computer science and engineering in Italy. *Scientometrics* 124, 3 (2020), 2207–2228. https://doi.org/10.1007/s11192-020-03548-9

[5] Camil Demetrescu, Andrea Ribichini, and Marco Schaerf. 2018. Accuracy of author names in bibliographic data sources: an Italian case study. *Scientometrics* 117, 3 (2018), 1777–1791.

[6] Alfio Ferrara and Silvia Salini. 2012. Ten challenges in modeling bibliographic data for bibliometric analysis. *Scientometrics* 93, 3 (2012), 765–785.

[7] Forest Gregg and Derek Eder. 2019. Dedupe. https://docs.dedupe.io/en/latest/. Accessed April 2021.

[8] Danielle H. Lee and Peter Brusilovsky. 2019. The first impression of conference papers: Does it matter in predicting future citations? *Journal of the Association for Information Science and Technology* 70, 1 (2019), 83–95.

[9] Michael Ley. 2020. Personal communication.

[10] G. Papadakis, G. Papastefanatos, T. Palpanas, and M. Koubarakis. 2016. Boosting the Efficiency of Large-Scale Entity Resolution with Enhanced Meta-Blocking. *Big Data Research* 6 (2016), 43–63. https://doi.org/10.1016/j.bdr.2016.08.002

[11] George Papadakis, Dimitrios Skoutas, Emmanouil Thanos, and Themis Palpanas. 2020. Blocking and filtering techniques for entity resolution: A survey. *ACM Computing Surveys (CSUR)* 53, 2 (2020), 1–42.

[12] T.A. Paraizo and D.A. Pereira. 2020. PVAF: an environment for disambiguation of scientific publication venues. *International Journal on Digital Libraries* 21, 4 (2020), 407–421. https://doi.org/10.1007/s00799-020-00289-1

[13] D. A. Pereira, E. E. Braga da Silva, and A. A. A. Esmin. 2014. Disambiguating publication venue titles using association rules. In *IEEE/ACM Joint Conference on Digital Libraries*. 77–86. https://doi.org/10.1109/JCDL.2014.6970153

[14] Y. Tao and H. Kong. 2018. Entity matching with active monotone classification. In *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. 49–62. https://doi.org/10.1145/3196959.3196984

[15] George Vrettas and Mark Sanderson. 2015. Conferences versus journals in computer science. *Journal of the Association for Information Science and Technology* 66, 12 (2015), 2674–2684.

[16] Yan Wang, Hao Zhang, Yaxin Li, Deyun Wang, Yanlin Ma, Tong Zhou, and Jianguo Lu. 2016. A Data Cleaning Method for CiteSeer Dataset. In *Web Information Systems Engineering – WISE 2016*. Springer International Publishing, 35–49. https://doi.org/10.1007/978-3-319-48740-3_3

[17] J. Wu, A. Sefid, A.C. Ge, and C.L. Giles. 2017. A supervised learning approach to entity matching between scholarly big datasets. In *Proceedings of the Knowledge Capture Conference, K-CAP 2017*. https://doi.org/10.1145/3148011.3154470

Just Accepted