Data Mining based approach for authors disambiguation in large citation networks

Francisco Rodríguez Drumond
ISH CNRS University of Lyon, Lyon, France - frandres@gmail.com

Fabien Rico
ERIC University of Lyon, Lyon, France - fabien.rico@univ-lyon1.fr

Adrian Tanasescu
ISH CNRS, University of Lyon, Lyon, France - adrian.tanasescu@ish-lyon.cnrs.fr

Djamel Abdelkader Zighed
ISH CNRS, University of Lyon, Lyon, France - abdelkader.zighed@ish-lyon.cnrs.fr

Abstract

Entity resolution is a demanding research problem for which many approaches have been proposed. Disambiguating authors of scientific publications is an instance of this problem: when facing a corpus of publications we often wish to know which articles belong to which real life authors. This task is not always possible based only on the information present in an article citation, which usually includes authors, title, year and journal. Two types of problem arise: synonymy, which corresponds to the case where the name of an individual can be written in more than one way and homonymy which describes the case where two or more individuals have the same name. In this paper we introduce a technique for addressing this task in large datasets by clustering papers based on proposed similarity measures. One can expect that the obtained clusters can be mapped to real life authors under the assumption that authors tend to write similar papers throughout time. The similarity measures we propose include a mix of string-similarity measures between relevant available fields and a novel graph-based similarity metric. The similarities produced are used to cluster papers with hierarchical agglomerative clustering. We then use a corpus of manually disambiguated papers of diverse scientific disciplines as ground truth to evaluate the performance of our technique. Our contributions include: 1) the use of a name similarity mask to determine candidate papers to be clustered, 2) a novel similarity function using a combination of string similarities as well as graph similarity metrics, and 3) a useful technique for the visualization of these similarity metrics to determine the best features for the clustering task.

Keywords: Disambiguation; Entity Resolution; Data Mining; Machine Learning.

1 Introduction

There is an increasingly large corpus of scientific publications available online in digital libraries. When collecting and organizing these papers, there is a desire to know which articles correspond to what individuals. However, it is not always clear how to determine the individual who wrote the publication directly from the name written in an article citation. There are two types of problems to be addressed: synonymy, the case where there is more than one way to write the name of an individual and homonymy, the case where many individuals have the same name.

Note that henceforth we will use the term individual to refer to an actual person, and the term author to refer to an ambiguous name in our data set.

This problem, known as the Author Disambiguation, is an instance of a more general problem known as Entity Resolution or Record Linkage. There are two broad ways to deal with this issue: author assignment and author grouping methods [3]. The former consists of learning a model for a specific individual based on
the available information; the latter consists of grouping similar papers in the hope that the obtained groups can be mapped to individuals. In this paper we propose an author disambiguation technique which falls into the author grouping category.

The underlying assumption for author grouping methods is that researchers very often write similar papers over time. The author disambiguation task can then be solved by determining relevant distance measures between papers which can be fed to a clustering algorithm. By identifying repeated names within the resulting clusters we can identify individuals. That is, if within a group of very similar papers we can find a common name, then we can infer that these papers were written by one individual.

The literature presents a number of different methods to build distance measures between papers. Many of them use text-similarity approaches to compare titles, author names, institutions, subjects and other available fields. We cover some of them in the related work section. Our contribution in this article is a novel context based approach for extracting similarity measures between papers. Our approach allows us to determine whether two papers are similar by taking into account their context. We aim to find communities of authors that either publish together or work on similar topics to determine if two papers belong to the same individual. By finding similarity measures we can then produce distances between papers which in turn can be used as input for a clustering algorithm.

We use this approach together with the idea of a focus name. Rather than disambiguating the entire set at one time, we divide the set of author names into focus names. A focus name is a string that groups one or more authors who have a similar name. In section 4.1 we explain this idea in detail. By using focus names we avoid the comparison of authors that in the first place cannot correspond to the same individual. As a consequence, we minimize the number of mistakes and reduce the computational time necessary for disambiguating large datasets.

To evaluate our approach, we built a corpus of manually disambiguated papers from a subset of scientific publications from the Thomson Reuters Web of Science database. We show how our approach produces appealing results.

This paper is organized as follows. First, in section 2 we present a brief outline of the state of the art, with an emphasis on existing graph based approaches. To better explain our approach, we present some relevant graph definitions and notation in section 3. Next, we explain our proposal in detail in section 4. In section 5 we present the methodology and the results of the experiments on our manually disambiguated set of papers. Finally, in section 6 we present some concluding remarks and considerations for future work.

## 2 Related Work

There are many approaches for building similarity measures for comparing articles based on available citation information. These similarity measures, often in the $[0, 1]$ range are then used to construct distance measures; these distances are then fed to a clustering algorithm to obtain distinct clusters of articles that we expect to belong to distinct individuals. The similarity measures used depend on the citation data available for an article, which generally consists of a list of author names, the article title, and the publication venue in the minimal case, and sometimes contains other information like the author’s institutions or keywords. Common measures are based on string comparison techniques like Jaccard index, Euclidean distance or cosine similarity on a TF-IDF representation of the features. Other approaches are supervised techniques that try to learn a similarity measure from the data [1, 9].

Other authors have also explored the use of graphs for the author disambiguation task. For instance, in [5] the authors propose an author disambiguation technique based on a combination of text similarity between author names and the length of the shortest path between the two papers, and the quantity of these shortest paths, in a coauthorship graph. The graph they propose is an undirected graph in which vertices are papers and author names while edges indicate authorship. They use the length and quantity of the shortest paths between papers to compute the relationship strength. Given an integer $d$, the relationship strength between two authors $a1$ and $a2$ is the number of paths between them of length $d$ or less. Then they use different thresholds for the relation strength and the text similarity to determine if two papers belong to the same person. For example, a rule might state: if the string similarity is higher than $t1$ and the relationship strength is higher than $t2$, then the papers belong to the same individual.
Similarly, the authors of [6] present a graph based approach. The most distinguishable aspect of their work is that rather than using a coauthorship graph they build a citation graph in which the vertices are papers and the edges indicate a citation relationship. This graph exploits the fact that authors tend to cite their own previous work; it is reasonable to assume that if two papers share a similar author name and one of them cites the other, then they belong to the same person. They also use other features such as coauthorship and paper’s URL, achieving good results. However, they don’t specify whether a graph based approach was used for these last two features.

Last but not least, the authors of [2] propose a different graph based method in which vertices are author names and edges indicate coauthorship of some papers. Their work is similar to ours in that two publications belonging to the same individual are expected to share some coauthors either directly or by some other intermediate publications. Their work however focuses on determining what are valid paths in this graph, which can be used for calculating a connection strength, which is related to the length of the paths. As it can be seen, the existing literature on graph methods focuses on building graph models based on the available features and then determining similarity (i) if there is a direct link or (ii) using the length of the shortest path between papers (and/or the number of equivalent shortest paths). Our proposal is novel and different from the existing work in that we rather use the number of shared neighbors to compute similarity between papers. While the number of neighbors between two nodes can be seen as the number of paths of length two, the motivation and the way we construct the graph and the final similarity measure is different from the approaches studied in the literature.

3 Definitions and notation

Our context based approach uses directed graphs for building similarity measures between papers. Formally a directed graph is an ordered pair \( D = (V, E) \) in which:

- \( V \) is a set whose elements are called vertices or nodes; and
- \( E \) is a set of ordered pairs of vertices, called directed edges or arcs. An arc \( a = (x, y) \) is considered to be directed from \( x \) to \( y \); we call \( x \) the tail vertex, or node, of the arc and \( y \) the head vertex of the arc. Arcs can be labeled with a string that contains information about the relationship between the two nodes.

We will use the expression tail and head vertices to refer to nodes that are tail and head vertices of at least one arc. A node \( x \) is called a sink node if there is no edge such that \( x \) is the tail node of the arc. Two nodes are considered adjacent if there is at least one arc connecting them. The neighborhood set \( N \) of a node \( x \) is the set of nodes adjacent to it:

\[
N(x) = \{ y \in V | (x, y) \in E \lor (y, x) \in E \}
\]  

(1)

We shall call the elements of this set the neighbors of node \( x \). The degree of a node is the number of edges incident to it. The in-degree of a node is the number of edges ending at that node while the out-degree is the number of edges starting at that node. From this point on, in our figures we will use numbers to identify nodes and letters to label arcs.

4 The technique

Our author disambiguation technique consists of three steps. First, we use focus names to select subsets of papers that could belong to the same person in order to reduce the computational cost of the whole task and to avoid comparing papers that could very rarely belong to the same person. Using focus names allows us to divide the whole dataset into subsets for which we compute similarity measures using our context based approach. We use our approach to build graphs from two features: coauthors and references. Then similarity measures are computed and combined into a single similarity measure taking the maximum of the two. The
similarity is then used to obtain distance measures following the equation \( \text{distance}(x, y) = 1 - \text{similarity}(x, y) \). The distance measures are then used to run a hierarchical agglomerative algorithm to cluster the papers; it is to be expected that the resulting groups of papers belong to distinct individuals.

In this section we explain in detail each of the three steps, with a particular emphasis on the method we use to compute the similarities since that is the core of our proposal in this paper.

### 4.1 The use of a focus name

Attempting to disambiguate a large set of papers can be a difficult task for two reasons. First, it can be computationally expensive to find similarity measures between all pairs of papers, particularly when their number is very large. For instance, the Microsoft Academic Search database contained more than 50 million publications and over 19 million authors at the time this paper was written. Second, two papers with very different author names could almost never belong to the same person, regardless of how similar they are in terms of other available features. For instance, there are two articles with very similar titles and subjects, written in the same research institution and published in the same venue, they cannot belong to the same person if the names of the authors are “John Smith” and “Peter Johnson”. Because of these reasons, identifying blocking features is a useful stage for entity resolution problems, as shown in [4]. Blocking features are those that allow us to discriminate papers which could never be associated with the author and avoid the comparison of objects when it does not make sense.

We follow this approach: instead of disambiguating the entire dataset at one time, we divide the set of author names to be disambiguated into smaller subsets according to focus names. A focus name is a string that groups one or more authors who have a similar name. There are different ways to specify a focus name. For us, a focus name groups names that share (i) a surname and (ii) at least one common first name and/or middle name initial. The use of initials for first and middle names rather than complete names has proven useful before, as reported in [7]. It essentially allows to circumvent problems related to spelling variants of the first names. We work with the assumption that there is little variation in the way the surname of the author is written.

For instance, from the set of author names \{John Smith, Peter Johnson, John Michael Smith, Jane Smith\} we would compare papers which contain \{John Smith, John Michael Smith, Jane Smith\} on one side and \{Peter Johnson\} on another side. That is, we compare them in different batches.

The main disadvantage of using focus names is that authors with surnames written in different ways can be mistakenly split. This happens whenever there is a spelling mistake in the surname or when the author publishes using different surnames. For instance, women who change their civil status (e.g. get married or divorced) may choose to change surname. Other issues are contraction of surnames, presence or absence of accents - a common problem with names from romance languages - and other variants. Finally, there may be more than one way to correctly write a surname.

In the overall, using a focus name is vastly more helpful than harmful and there are ways to circumvent the mentioned difficulties. We believe that choosing a focus name that groups a greater number of authors can solve some of the problems. For instance, instead of choosing the last name, a set of last names with a given string-based similarity can be used (e.g Levenshtein distance or Jaro-Winkler distance). This would allow to better disambiguate last name variants like Rodríguez, Rodrígues, Rodriguez, etc. Though they are different last names, they are often mistakenly interchanged. We leave this idea for future work as the core of our proposal is the context based similarities, which we introduce next.

### 4.2 Context based similarities

A number of author disambiguation approaches consist of finding similarity measures between papers; many of these approaches use text-similarity techniques to compare titles, keywords, institutions, etc. For instance, when trying to extract a similarity measure from the set of authors of two papers we could use the Jaccard index. The Jaccard index is a set similarity measure which consists of computing the size of the intersection divided by the size of the union of two set. The higher the value for the two sets of authors the more similar two papers would be.
However, consider the following example. Suppose that we are disambiguating the hypothetical set of papers related to the focus name Smith, J., consisting of the following two papers, both belonging to the same person:

- (P1) Smith, J. and Roberts P. *Graph based approaches for entity resolution tasks.*
- (P2) Smith, J. and Williams C. *A novel author disambiguation technique.*

Setting aside author Smith, J. (common to both papers since it is the focus name we are trying to disambiguate), we could compute the Jaccard index to obtain a similarity measure. It’s value would be zero, as the intersection of the sets \{Roberts P.\} and \{Williams C.\} is the empty set.

Suppose that authors Roberts P. and William C. belong to the same research community; that is, they work and publish together. This is often the case: researchers usually work with other people who in turn publish together. In this hypothetical case is thus possible to find a third publication written by both Roberts P. and Williams C., the coauthors of papers P1 and P2, as follows:

- (P3) Roberts P. and Williams C. *Using graphs for automatic author detection.*

Although paper P3 would not be compared with papers P1 and P2 when disambiguating the focus name Smith, J., P3 can be useful to determine that P1 and P2 belong to the same person. The reasoning is that if papers P1 and P2 are written by authors that in turn publish together, then it is very likely that these three papers were written inside a common scientific community and, consequently, very probable that papers P1 and P2 were written by the same person called Smith, J.

Even in the case where two papers share coauthors, meaning that the intersection of their authors sets is not empty and the Jaccard index is greater than zero, it is reasonable to take into account the context of the two publications. This allows us to analyse their similarity using not only the information present in the citation, but also that contained in the whole set of papers. Therefore, we want to find (i) the research community in which authors publish by means of the coauthors and (ii) sets of papers that share the same topic by means of the references. We do this by constructing a graph of related papers according to the previously mentioned features: coauthors and references. It allows us to find other papers that are similar to the initial two and that act as bridges disclosing hidden relationships. The idea is that by finding similar papers to the ones we are trying to disambiguate we can discover links not present in the citations and exploit them to obtain similarity measures.

We use context graphs to implement this idea. We make a graph for finding papers that are similar according to the coauthors and another graph for papers that are similar according to the references. We shall call these coauthor community and reference graphs. They are both instances of a context graph.

We define a context graph as follows. The set of nodes of a context graph is the set of papers associated with the focus name we aim to disambiguate, together with additional papers from the dataset that are similar to the papers to be disambiguated. For the coauthor community graph, the additional papers are those that share at least one author - different from the focus name - with the papers associated with the focus name. For the reference graph the additional papers are those that share at least one reference with the papers related to the focus name.

The arcs of a context graph are drawn from papers associated with the focus name to papers that are similar. We restrict tail nodes to be only papers related to the focus name; we do this because in our model arcs between papers not associated by the focus name are not used for constructing our similarity measure. In the case of the coauthor community graph two papers are similar if they share a coauthor different from the focus name. In the case of the reference graph, two papers are similar if they share one reference.

For the reader interested in a formal definition of our graphs, our coauthor community graph is defined as follows.

- Let \( \Pi \) be the set of all papers in our data set
- Let \( f \) be the focus name to be disambiguated.
- Let \( A(i) \) be the set of the authors of paper \( i \).
Let $C(i, a)$ be the coauthors of author $a$ in paper $i$. We can write it in terms of $A(i)$ and $a$ as $C(i, a) = A(i) - a$.

Let $P(a)$ denote the set of papers written by author $a$.

Then our coauthor community graph $G_{CC} = (V_{CC}, E_{CC})$ is given by the following set of vertices:

$$V_{CC} = P(f) \cup \{ p \in \Pi | (\exists x| x \in P(f) \land C(x, f) \cap A(p) \neq \emptyset) \}$$

(2)

And the following set of arcs:

$$E_{CC} = \{(x, y) \in V_{CC}^2 | x \in P(f) \land C(x, f) \cap A(y) \neq \emptyset \}$$

(3)

Given the focus name $f$, a directed edge is drawn from a vertex $x$ to a vertex $y$ iff $x$ and $y$ share at least one author different from $f$ and $f$ is one of the authors of paper $x$. We label the arcs according to the name of the coauthors shared by the papers.

Analogously, our reference graph is defined as follows.

Let $\Pi$ be the set of all papers in our data set

Let $f$ be the focus name to be disambiguated.

Let $R(i)$ be the set of the references of paper $i$.

Let $P(a)$ denote the set of papers written by author $a$.

Then our reference graph $G_R = (V_R, E_R)$ is given by the following set of vertices:

$$V_R = P(f) \cup \{ p \in \Pi | (\exists x| x \in P(f) \land R(x) \cap R(p) \neq \emptyset) \}$$

(4)

And the following set of arcs:

$$E_R = \{(x, y) \in V_R^2 | x \in P(f) \land R(x) \cap R(y) \neq \emptyset \}$$

(5)

Given the focus name $f$, we draw edges from a vertex $i$ to a vertex $j$ iff $i$ and $j$ share at least one reference and $f$ is one of the authors of paper $i$. We also label the arcs according to the shared references.

Note that there are no restrictions impeding head vertices to be also tail vertices: if two papers to be disambiguated have coauthors - different from the focus name - or references they will also have arcs connecting them. Also note that the are no restrictions impeding a paper associated with a focus name to have an arc pointing to itself.

Notice as well that for coauthor community graphs we draw edges between two papers if they share one author name; that name can in turn be ambiguous. However, we work under the assumption that given a focus name, if two papers to be disambiguated have coauthors with the same name publishing together, then it is very likely that these coauthors actually know each other and belong to the same community.

Figure 1 shows the coauthor community graph for the example we presented at the beginning of this section. For the sake of simplicity we use letters instead of full names: the focus name to be disambiguated is $f$ and we use $a, b$, and $c$ to denote coauthors. The set of vertices is composed by papers $P1, P2, P3$. We show the authors of each of these papers in curly brackets (i.e. $P1$ is written by $f$ and $a$; $P2$ is written by $f$ and $b$ and $P3$ is written by $a$ and $b$). $P1$ and $P2$ are papers with the focus name $f$ and are consequently tail vertices. $P3$ is not in the list of papers to be disambiguated, but is in the graph as a head vertex since it shares coauthors with the other papers. Nodes $P1$ and $P2$ have arcs pointing to themselves ($P1$ and $P2$ respectively) since they have more than one author, and consequently share at least one coauthor.

As we will see, the graph we construct allows us to identify similar papers by looking at the number of neighbors they share. Recall that neighbors are the elements of the set previously defined in equation 1. However, we need a similarity measure which can be converted to a distance measure for running clustering algorithms. This measure should be proportional to the number of shared neighbors, but should also take into account the number of neighbors the two papers to be compared have. To obtain a normalized measure we divide the number of shared neighbors by the size of the smallest set of neighbors. The formula is:
Figure 1: The coauthor community graph for the example. P1 and P2 are similar if they are found to be similar to a third paper P3.

$$\text{Similarity}(a, b) = \frac{|\text{Neighbors}(a) \cap \text{Neighbors}(b)|}{\min(|\text{Neighbors}(a)|, |\text{Neighbors}(b)|)}$$ (6)

Before computing the similarity measure between two papers we first take two refinement steps, which we now explain in depth. To do so we illustrate the graph of a small subset of publications of our dataset, characterized by belonging to the focus name Lefebvre A. Figure 2 shows the coauthor community graph for this focus name. The big black vertices are tail vertices, papers we aim to disambiguate. The numerous small gray vertices are sink vertices, papers that share at least one author besides the focus names, with the papers to be disambiguated.

This subset is characterized by having 30 papers; this number is sufficient enough to be representative of the general case while having an adequately number of vertices to allow for an understanding of the graph.

For the sake of simplicity, in this section we will refer to coauthor community graphs. However, the refinement steps apply to reference graphs as well.

4.2.1 Removing sink nodes with in-degree equal to one

Figure 2 shows the coauthor community graph for the focus name Lefebvre A. The first thing one notices is that several tail nodes share a great number of neighbors and form clusters of papers that may belong to distinct individuals. The second thing we can see is that there is a large number of sink vertices that are only connected to one tail vertex. For this particular subset of papers, they account for over 80% of the vertices. We are only interested in keeping sink vertices with in-degree greater or equal to two. This only affects the denominator of our similarity measure. The numerator is defined as the size of the intersection of the neighbors of two papers, and nodes with in-degree one would never be in that intersection. Sink nodes that act as bridges can be used to establish that two papers are similar, either because a subset of their coauthors belong to one community or because they talk about similar topics as determined by a subset of their references.

Keeping sink nodes with in-degree equal to one could also make our similarity measures too low since we normalize dividing by the size of the smallest set of neighbors of the two nodes. This is harmful because publications that have many coauthors or references could have lower similarity measures between themselves than publications with less coauthors or references.

Additionally, the large number of sink nodes with in-degree equal to one would make the computation significantly more expensive.

Figure 3 shows an example of a case in which we need to filter out papers with in-degree equal to one. It is a small variation of the example presented in Figure 1. In this case P1 is written by authors f, a, and c. It is related to paper P4 because they both share author c. However, the fact that paper P4 has in-degree...
Figure 2: Unrefined coauthor community graph for the focus name Lefebvre A.

one is evidence that author c does not belong to the community of authors a and b. In that case, the similarity measure can change from 0.33 to 0.5 by pruning P4. Before pruning the set of neighbors of P1 is \{P1, P3, P4\}. The set of neighbors of P2 is \{P2, P3, P5\}. By pruning P4 the set of neighbors of P1 goes to \{P1, P3\} and so the minimum size of their sets of neighbors is two.

Figure 4 shows the resulting coauthor community graph for the focus name Lefebvre A after filtering out the unwanted vertices. The spatial distribution is the same, although tail vertices not connected to any other vertices (in black) have been grouped to the right. Papers that belong to the same individual and that are close one another in the graph are colored in light blue and red. We alternate between the use of red and blue to distinguish papers that despite being spatially close belong to different individuals. The small gray nodes are sink vertices.

Notice how papers that belong to the same individuals tend to share several neighbors. The only exception is the red isolated node in the lower-left corner. This paper belongs to the same individual as the other two vertices right next to it, but does not share any of their neighbors. This is because in this particular case the author wrote that paper by himself.

4.2.2 Removing equivalent sink nodes

We now present a second refinement of the graph. It is very likely that a paper shares the same coauthor with more than one paper, particularly when it is a very common author name. In other words, it is not hard to find that a tail vertex has edges with more than one different head vertex for sharing the same coauthor. This does not necessarily affect in a negative way our approach. However, we are concerned by the case in which there are multiple tail vertices that are connected to the same sink vertices by sharing the same coauthors. Figure 5 illustrates this situation by slightly modifying the example shown in Figure 1.

Notice that the edges from P1 to papers P4, P5, P6 are drawn because the latter three share coauthor a with P1. Similarly, the edges from P2 to papers P4, P5, P6 are drawn because of coauthor b. Since vertices P3, P4, P5 have the same tail nodes, and because the edges connecting to the tail nodes have the same labels, we can treat these three nodes as equivalent.

Formally, we define two sink vertices i and j to be equivalent iff they (i) share the same set of tail vertices T and (ii) for every node t in T the edge from t to i has the same label as the edge from node t to j.

We are concerned about equivalent sink nodes because they introduce unwanted redundancy and can negatively affect the quality of our similarity measure. As previously said, to obtain normalized similarity
measures it is necessary to take into account the total number of neighbors each vertex has. Notice that in Figure 5, $P1$ and $P3$ share two neighbors - themselves -. $P1$ has other three neighbors, but they are equivalent. Paper $P3$ has also other two neighbors, $P7$ and $P8$ which are not equivalent. The fact that there are more papers written by authors $a$ and $b$ than papers written by $c$ should not imply that papers $P1$ and $P2$ are more similar than papers $P1$ and $P3$.

It is necessary to take into account these equivalence sets because generally there are names which are more common than others. Also, frequently a group of authors publish a great number of papers together. This can significantly hinder our similarity measures. Our proposal is that for each equivalence set, we keep only one vertex and discard the rest. Figure 6 shows the result of applying this transformation to the graph of Figure 5.

Notice that now, $P1$ and $P2$ share one neighbor while $P1$ and $P3$ share two neighbors as before. Sink vertices, rather than representing papers now represent a match on a criterion. The more matches on different criteria between papers - the more non-equivalent neighbors - the better.

Figure 7 shows the result of applying this transformation to the coauthor community graph for the focus name Lefebvre A. Again, tail nodes have been colored in red and light blue in order to show papers that belong to the same individual. We alternate between red and blue to distinguish papers belonging to different individuals who are close to each other in the graph.

Notice how there are significantly less sink vertices in this new graph. This gives an idea of the amount of redundancy carried by the previously defined equivalence sets. Also, it is clear that nodes belonging to the same individual share many of their neighbors.

We make one final remark to conclude this section. In this paper we use our graph technique for finding similarities based on coauthors and references. However, similar graphs can be constructed from other features. The idea is that an edge is drawn between two vertices if they are similar according to any criteria. For instance, we can build a graph drawing edges between two nodes if the corresponding papers share a similar title or abstract. Then we can follow the same approach to refine it and finally extract a similarity measure.

In this work we decided to use coauthors and references because they provide: (i) information about scientific communities and (ii) information about papers treating similar topics. Using these fields do not require computing text similarities as it would be necessary for titles, as an example.
4.3 Running a clustering algorithm

The two similarity measures obtained from our two graphs are combined into a single measure by taking the maximum value. That is, for every pair or papers to disambiguate we take the maximum between the coauthorship and the reference similarity.

We decided to do this after discarding other options. Empirically, constructing a single graph with both features did not produce good results because the number of nodes in the coauthor community graphs is significantly bigger than the number of nodes in the references graphs. Taking a weighted sum does not work since two papers belonging to the same individual may have a very low similarity measure for one of the features. For instance when the author works in several different research communities and studies very different topics. By taking the maximum we are being optimistic in that, if two papers are similar according to just one of the features, then there is a high chance that they belong to the same person. As we will see, empirical results show that this is a good idea.

From the similarity measure we can build the distance measures by using the following equation: \( \text{distance}(a, b) = 1 - \text{similarity}(a, b) \). The obtained distance measures are used to run Hierarchical Agglomerative Clustering with single linkage. The stopping criteria for this algorithm is chosen automatically. Given a distance matrix, we choose the threshold that maximizes a consensus function defined as the weighted difference between intra and inter cluster distances. The intra and inter cluster distances are the within-cluster and between-cluster sum of squared distances, normalized by the size of the clusters. In particular, we maximize:

\[
C(\text{clusters}) = \text{IntraD}(\text{clusters}) - 0.75\text{InterD}(\text{cluster})
\]

where the values of the parameters were chosen using a cross-validation set. More information about the consensus function can be found in [8].

Finally, the obtained clusters represent groups of papers which very likely belong to distinct individuals. These can be compared with a ground truth set in order to verify whether the clustering algorithm yields good results. We explain how this is done in the next section.
5 Experiments

In this section we explain the experiments we performed to compare our proposal with other existing techniques and provide the obtained results. In section 5.1 we explain the methodology and in 5.2 we provide the experimental results along with an analysis.

5.1 Methodology

To assess the quality of our approach, we used a set of manually disambiguated papers - meaning that we know what the individuals are - to evaluate the performance of our measure for author disambiguation and compare it with other techniques. We worked with 901 manually disambiguated publications from a subset of the database from the Thomson Reuters Web of Science database. This subset contains all the publications from 2007 to 2010 with at least one French author. Despite the limited time period and the presence of at least one French author, we believe that this data set is representative of the general author disambiguation task; future work on a different dataset might help confirm this.

The manual disambiguation consists in finding the individuals of the publications on the Internet; this
includes looking for author’s email addresses, personal web spaces, visiting web pages of research institutions and other heuristics. Manually disambiguating publications for the evaluation of a proposal is a common step in several author disambiguation papers.

Our experiments consist in extracting similarity measures using our approach and comparing their quality with similarity measures extracted using text-based approaches. We compare them by (i) running a Hierarchical Agglomerative Clustering algorithm as explained in section 4.3 and evaluating the results based on the ground truth (ii) producing heat maps based on the similarity measures and visually comparing them. The two techniques chosen for the comparison are the Jaccard index and cosine similarity on a TF-IDF vector representation, both commonly used for the author disambiguation task.

We compute the three similarity measures as follows. For the Jaccard index, we consider the set of coauthors and references of every paper and compute the similarity between them as the size of the intersection of these sets divided by the size of their union. The TF-IDF approach computes a Term Frequency-Inverse Document Frequency vector model of the documents; the similarity between papers is the cosine of their corresponding vectors. The documents used for the TF-IDF technique are (i) the concatenation of all the author names of the articles and (ii) the concatenation of the venue and the first name of the author of each referenced article (the title of the referenced article was not available in our dataset). Finally, the similarity measure for our approach is computed as explained in section 4.2.

To evaluate the quality of our clustering we use pairwise F1-score, precision and recall evaluated against the ground truth set. This is the standard procedure in the literature on author disambiguation. To do this, all possible pairwise combinations of papers that belong to one cluster are considered; the ground truth pairs are the correct pairs while the found pairs are the ones generated according to the results of the clustering algorithm. Precision, recall and the F1-score are then computed according to their well known formulas:

\[
\text{precision} = \frac{|\text{correct_pairs} \cap \text{found_pairs}|}{|\text{found_pairs}|} \quad (8)
\]

\[
\text{recall} = \frac{|\text{correct_pairs} \cap \text{found_pairs}|}{|\text{correct_pairs}|} \quad (9)
\]

\[
F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (10)
\]

Precision thus gives information about the purity of the clusters: high precision means that papers grouped together indeed belong to the same individual. Recall measures the ability to actually identify individuals: high recall means that a high number of papers that belong to the same author were indeed put together. The F1-score uses both criteria to give an overall idea of the quality of the results.
Table 1: F1-scores for coauthors sim. measures

<table>
<thead>
<tr>
<th>Focus name</th>
<th>J. Index</th>
<th>TF-IDF</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arlot</td>
<td>0.71</td>
<td>0.71</td>
<td>0.85</td>
</tr>
<tr>
<td>Barba</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Bassetti</td>
<td>0.84</td>
<td>0.90</td>
<td>0.81</td>
</tr>
<tr>
<td>Beer</td>
<td>0.82</td>
<td>0.82</td>
<td>0.84</td>
</tr>
<tr>
<td>Bruce</td>
<td>0.87</td>
<td>0.89</td>
<td>0.94</td>
</tr>
<tr>
<td>Castella</td>
<td>0.37</td>
<td>0.66</td>
<td>0.95</td>
</tr>
<tr>
<td>Eklund</td>
<td>0.88</td>
<td>0.91</td>
<td>0.99</td>
</tr>
<tr>
<td>Gaillot</td>
<td>0.80</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td>Jouet</td>
<td>0.75</td>
<td>0.64</td>
<td>0.87</td>
</tr>
<tr>
<td>Karakiewicz</td>
<td>0.88</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>Kraft</td>
<td>0.56</td>
<td>0.45</td>
<td>0.91</td>
</tr>
<tr>
<td>Nikolic</td>
<td>0.68</td>
<td>0.68</td>
<td>0.96</td>
</tr>
<tr>
<td>Pita</td>
<td>0.94</td>
<td>0.92</td>
<td>1</td>
</tr>
<tr>
<td>Pujolle</td>
<td>0.14</td>
<td>0.20</td>
<td>0.15</td>
</tr>
<tr>
<td>Rohrmann</td>
<td>0.84</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>Stokes</td>
<td>0.73</td>
<td>0.64</td>
<td>0.91</td>
</tr>
<tr>
<td>Zighed</td>
<td>0.65</td>
<td>0.65</td>
<td>0.61</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.72</strong></td>
<td><strong>0.73</strong></td>
<td><strong>0.85</strong></td>
</tr>
</tbody>
</table>

Last but not least, we can also visualize the quality of the three techniques by generating a heap map from ordered distance matrices. The idea is that based on a ground truth set we can assign an id to each individual author and to each of their papers. We can then arrange the rows and columns of the distance matrix so that they are ordered by the cluster id. We can compare the resulting heat maps by also generating an *ideal heat map* based on the ground truth set: we assign a distance of 0 between members of the same cluster and 1 between members of different clusters. In this ideal heat map we can expect squares along the diagonals which identify clusters of papers written by the same individual. Comparing heat maps obtained according to different techniques and features against the ideal heat map gives an idea of the quality of the obtained measures.

After comparing our technique with the other two chosen methods by means of the F1-score and heat maps, we present the detailed results of using our context based technique on coauthors and references and combining them using the maximum of the two similarities as explained in section 4.3.

### 5.2 Results

Table 1 shows the result of running the clustering algorithm using similarity measures derived from the coauthors of a publication. We present the results for each of the focus name in our test-set.

For most of the focus names, our context based similarity measure is the one that yields the best result. Most of its F1-scores tend to be above 0.80 except for a few outliers (Pujolle and Zighed). Averaging the F1-score according to each technique confirms this idea: our similarity measure produces an average F1-score of 0.85 versus the 0.73 and 0.72 obtained when using the other measures.

In table 2 we present the results for the similarity measures extracted from the references. These results are slightly different. In general, the results obtained by using the Jaccard index are significantly worse than the results obtained by using TF-IDF with cosine distance or our context based measure. When comparing the two latter, TF-IDF performs slightly better than our measure. Most of the times TF-IDF is the best technique for references, though in a few cases (Arlot,Eklund,Kraft,Pita,Stokes) our technique fares better. This is confirmed by comparing the averages over the three techniques: the Jaccard index produces an F1-score of just 0.65 compared to 0.77 (our measure) and 0.84 (TF-IDF with cosine similarity).

These results show that our context based technique is very good for extracting similarity measures from the coauthors. By looking at other papers written by the same coauthors, we can expect to detect communities
Table 2: F1-scores for references sim. measures

<table>
<thead>
<tr>
<th>Focus name</th>
<th>J. Index</th>
<th>TF-IDF</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arlot</td>
<td>0.60</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>Barba</td>
<td>0.76</td>
<td>0.95</td>
<td>0.82</td>
</tr>
<tr>
<td>Bassetti</td>
<td>0.78</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>Beer</td>
<td>0.64</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>Bruce</td>
<td>0.89</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>Castellia</td>
<td>0.21</td>
<td>0.61</td>
<td>0.53</td>
</tr>
<tr>
<td>Eklund</td>
<td>0.79</td>
<td>0.85</td>
<td>0.96</td>
</tr>
<tr>
<td>Gaillot</td>
<td>0.53</td>
<td>0.96</td>
<td>0.87</td>
</tr>
<tr>
<td>Jouet</td>
<td>0.24</td>
<td>0.44</td>
<td>0.33</td>
</tr>
<tr>
<td>Karakiewicz</td>
<td>0.60</td>
<td>0.76</td>
<td>0.66</td>
</tr>
<tr>
<td>Kraft</td>
<td>0.79</td>
<td>0.88</td>
<td>0.94</td>
</tr>
<tr>
<td>Nikolic</td>
<td>0.93</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>Pita</td>
<td>0.94</td>
<td>0.96</td>
<td>1</td>
</tr>
<tr>
<td>Pujolle</td>
<td>0.09</td>
<td>0.97</td>
<td>0.25</td>
</tr>
<tr>
<td>Rohrmann</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>Stokes</td>
<td>0.85</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>Zighed</td>
<td>0.43</td>
<td>0.67</td>
<td>0.48</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.65</strong></td>
<td><strong>0.84</strong></td>
<td><strong>0.77</strong></td>
</tr>
</tbody>
</table>

of researchers working together. This allows to exploit information which is not present in the citation but still in the dataset.

In the case of references, the fact that our technique outperforms the Jaccard index indicates that by using the context we can obtain better results than by just looking at direct matches between the list of references. Papers with shared references allows us to identify communities of papers written about the same topics. However, in this case TF-IDF outperforms the other techniques because it can identify that two papers are similar even when they have different references, but similar journal or conference names. For instance, authors that work on Data Mining will usually publish in Data Mining conferences which will have similar names. Consequently, TF-IDF can detect that two papers are about the same topic even if they don’t share references or even if there are no other papers that link them.

These ideas are confirmed by looking at the heat maps obtained from each of these features and techniques. For the sake of brevity we only present the heat map for Gaillot, in figure 8. We chose this focus name because it is representative of the rest of the focus names in our dataset.

Notice that for coauthors, the context approach heat map better approximates the optimal heat map. This is particularly so when observing the square in the upper left corner. While the Jaccard index and TF-IDF with cosine distance allow to identify some of the members of that cluster, our context based measure identifies all but two of the members. Also note that since this is a heat map, the brighter the colors the stronger the similarity between two papers. The context based approach thus produces higher similarity values between papers that belong to the same person than its counterparts.

When it comes to references, our approach produces once more a better heat map than the one generated by using the Jaccard index. The latter allows to identify only links between some papers in the upper left region and some links in the central region. The same regions are better identified by our technique. However, TF-IDF is the only technique that allows to identify clusters in the lower right corner. Despite low similarity values, the clustering algorithm can still group these papers. Notice that because TF-IDF does not look for strict matches in the list of references, but rather for similar strings, there is some noise introduced in the distance matrix. We refer to the small light patches located far from the diagonal; these identify papers that belong to different authors but that are found to be slightly similar. In this case, because their similarity measure is too low, it does not hinder the results.

Note that as explained at the end of section 4.2, our approach does not exclude benefiting from string similarity measures. In that sense, we could change the criteria for drawing edges in a reference graph, so that arcs are drawn if there is at least one reference with a similar title or venue name, rather than if they
Table 3: Results of combining the two measures

<table>
<thead>
<tr>
<th>Focus name</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arlot</td>
<td>0.94</td>
<td>1</td>
<td>0.97</td>
</tr>
<tr>
<td>Barba</td>
<td>0.99</td>
<td>0.91</td>
<td>0.95</td>
</tr>
<tr>
<td>Bassetti</td>
<td>1</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>Beer</td>
<td>0.78</td>
<td>0.96</td>
<td>0.86</td>
</tr>
<tr>
<td>Bruce</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Casteilla</td>
<td>1</td>
<td>0.88</td>
<td>0.93</td>
</tr>
<tr>
<td>Eklund</td>
<td>0.97</td>
<td>1</td>
<td>0.99</td>
</tr>
<tr>
<td>Gaillot</td>
<td>0.97</td>
<td>0.80</td>
<td>0.88</td>
</tr>
<tr>
<td>Jouet</td>
<td>1</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>Karakiewicz</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Kraft</td>
<td>0.95</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>Nikolic</td>
<td>0.99</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td>Pita</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pujolle</td>
<td>1</td>
<td>0.21</td>
<td>0.35</td>
</tr>
<tr>
<td>Rohrmann</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Stokes</td>
<td>0.92</td>
<td>1</td>
<td>0.96</td>
</tr>
<tr>
<td>Zighed</td>
<td>1</td>
<td>0.48</td>
<td>0.65</td>
</tr>
<tr>
<td>Average</td>
<td>0.97</td>
<td>0.88</td>
<td>0.90</td>
</tr>
</tbody>
</table>

share the exact same reference. This is however left for future work.

Finally in table 3 we present the results of running a clustering algorithm on a similarity measure obtained by taking the maximum of our two context measures. The results obtained are very appealing: the average F1-score is 0.90 and except for two outliers (Pujolle and Zighed) the F1-scores are all above 0.85. Note that precision tends to be higher than recall; this means that our algorithm is better at correctly identifying papers written by the same real life individual than at identifying all the papers written by the same person. In other words, our algorithm does a better job at accurately assuring that two papers belong to the same individual than at finding all the papers of an individual.

6 Conclusions

In this paper we introduce a novel context based method for finding similarity measures between scientific publications in order to solve the author disambiguation problem. Most of the existing methods use text based similarity measures; a few approaches in the literature use graphs. Our approach is novel in that instead of using the length of paths between papers we build a context graph and compute a similarity measure proportional to the number of shared neighbors. By doing this, we are able to detect groups of papers that are written inside the same community or that are related to the same topic. Our context graph exploits the fact that two papers are similar if they are also similar to many other papers. We present in detail the methodology to construct the graph and two refinement steps consisting of removing vertices which do not add information and equivalent vertices which add redundancy to the model.

We combine this approach with the use of a focus names, which consists of disambiguating papers related to one name at a time. This allows us to avoid comparing papers which could rarely belong to the same person, reducing the time necessary to work with large databases and allowing us to obtain better results.

We extract similarity measures from coauthors and references using our graph technique, Jaccard index and TF-IDF and we compare their quality running a Hierarchical Agglomerative Clustering algorithm. We also produce heat maps from the similarity matrices for a visual quality assessment. Both the quantitative results and visualizations show that our approach is significantly better than the other two techniques for coauthors and better than the Jaccard index for references. With references, TF-IDF slightly outperforms our technique; this may be due to the fact that two different references can be still be similar with respect to conference and journal names, something that our graph approach does not take into account.
There are two lines for future work. First, we can experiment with using text similarity measures to construct a context graph. This could lead to better results for references and enable us to use other features such as title.

Second, it would be interesting to run a community detection algorithm to directly detect clusters from the graphs. In a way our context graphs resemble the structure of a community: edges are drawn between papers if they share coauthors or references. Thus, we could easily detect research communities (from the coauthors) or groups of researchers working in the same topic (from the references).

References


Figure 8: Heat maps from the similarity matrix obtained by using different techniques for the focus name Gaillot.