A Strategy for Selecting Relevant Attributes for Entity Resolution in Data Integration Systems

Gabrielle Karine Canalle, Bernadette Farias Loscio and Ana Carolina Salgado
Center of Informatics (CIn), Federal University of Pernambuco (UFPE), Recife, Pernambuco, Brazil

Keywords: Attribute Selection, Entity Resolution, Data Integration.

Abstract: Data integration is an essential task for achieving a unified view of data stored in heterogeneous and distributed data sources. A key step in this process is the Entity Resolution, which consists of identifying instances that refer to the same real-world entity. In general, similarity functions are used to discover equivalent instances. The quality of the Entity Resolution result is directly affected by the set of attributes selected to be compared. However, such attribute selection can be challenging. In this context, this work proposes a strategy for selection of relevant attributes to be considered in the process of Entity Resolution, more precisely in the instance matching phase. This strategy considers characteristics from attributes, such as quantity of duplicated and null values, in order to identify the most relevant ones for the instance matching process. In our experiments, the proposed strategy achieved good results for the Entity Resolution process. Thus, the attributes classified as relevant were the ones that contributed to find the greatest number of true matches with a few incorrect matches.

1 INTRODUCTION

The growing ease of data generation and sharing has contributed to the rapid rise of data volume available in digital environments. However, this growth has occurred in an uncontrolled manner, in such a way that a lot of data were published containing erroneous, missing or duplicate values, which may hinder its use. Nevertheless, there is a rising demand for data integration solutions over distributed and heterogeneous data sources. Examples are price comparison websites, as Booking.com[1] and Trivago.com[2].

Data integration solutions aim to combine data from different sources to provide a unified view of these data for users. An important step in this process is Entity Resolution (Christen, 2012), which aims to identify equivalent instances, i.e., which ones represent the same real-world concept (Dong and Srivastava, 2015). Entity Resolution consists of several stages, including a comparison step between instances pairs. During this step, the similarity between the attribute values that describe the compared instances is evaluated. In this context, one of the main challenges concerns the choice of attributes to use in the comparison step.

Generally, the choice of attributes is performed manually (Gruenheid et al., 2014). However, given the existence of a lot of attributes and the lack of prior knowledge of the sources domain, this choice may take into consideration attributes that could not contribute efficiently for the Entity Resolution process. Once the quality of Entity Resolution results is directly affected by the selected attributes, the use of strategies to help the selection of the best ones becomes essential.

In this paper, we propose a strategy that automates the Attribute Selection task in order to facilitate the identification of good attributes to be used in the comparison of instances. Our proposal considers specific criteria related to attribute values, like density and repetition to evaluate the attribute relevance. An attribute is considered relevant if it contributes positively for the identification of true correspondences, and irrelevant if it contributes in identifying incorrect matches (false positives and false negatives). By using the proposed strategy, it is expected that, at the end of the Entity Resolution process, we obtain the largest possible number of true matches with the lowest number of incorrect matches.

The rest of this paper is organized as follows. Section 2 presents a motivational example to illustrate the relevance of the proposed strategy. Section 3 presents

[1] http://www.booking.com
[2] http://www.trivago.com

Canalle, G., Lóscio, B. and Salgado, A.
A Strategy for Selecting Relevant Attributes for Entity Resolution in Data Integration Systems.
DOI: 10.5220/0006316100800088
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 1*, pages 80-88
ISBN: 978-989-758-247-9

Table 1: Result of a query about Data Integration in the sources CiteSeerX and DBLP.

ID Paper	ID	Source	Author	Title	Year	Venue	Pages
1	1	CiteSeerX	M. Lanzerini	Data Integration: A Theorical Perspective (2002)	2002	Symposium on Principles of Database Systems	NULL
	2	DBLP	Maurizio Lanzerini	Data Integration: A Theorical Perspective	2002	PODS 2002	233-246
2	3	DBLP	Guy Pierra	The PLIB ontology-based approach to data integration	2004	IFIP Congress Topical Sessions	13-18
3	4	CiteSeerX	Patrick Ziegler and Klaus R. Dittrich	Three decades of data integration - all problems solved	NULL	In 18th IFIP Computer Congress (WCC)	NULL
	5	DBLP	Patrick Ziegler and Klaus R. Dittrich	Three decades of data integration - All problems solved?	2004	IFIP Congress Topical Sessions	NULL

a summary of the main concepts related to Entity Resolution and Attribute Selection. In Section 4, it is specified the proposed Attribute Selection Strategy. Section 5 discusses the results obtained through experiments. In Section 6, some related works are presented, and finally, Section 7 presents the final considerations and future works.

2 MOTIVATIONAL EXAMPLE

To illustrate the need to select the best attributes to be used in the comparison step of the Entity Resolution process, and how these attributes impact on the quality of the process results, consider the following example.

An on-line digital library service, in the domain of Computer Science, integrates data from multiple sources, such as CiteSeerX [3] and DBLP[4], and makes feasible searches by title, author, or keyword. Assume that a user is interested in searching for papers about "Data Integration". The integration service submits the query to the data sources CiteSeerX and DBLP, and get a set of papers as the result for the query. A small fraction of this result can be seen in Table 1.

In this example, we present five instances, each one containing a proper identifier (ID). These instances are related to three different papers, identified by column ID Paper. In order to provide the integrated result to the user, the Entity Resolution process should be performed. For this, in the comparison step, the attributes that describe instances are compared. As mentioned before, the attribute selection is done manually or considering all attributes in the comparison step, we will suppose both situations.

Suppose an Entity Resolution process that is performed considering all attributes. Possibly, the instances 1 and 2 would be considered as not duplicated, given that for each of the five considered attributes, two (Venue and Pages) have a low value of similarity. The same would happen with the instances 4 and 5, whereby the attributes Year and Pages also have a low value of similarity. We can observe that attributes containing null values may affect negatively the Entity Resolution result. This happens because we consider that a null values led to similarity equals to 0 (zero). For this reason it does not make sense to use attributes that contain a large number of missing values, because this can make two equivalent instances as distinct ones. Therefore, the comparison considering all attributes would result in incorrect matches, known as false negative, given that instances of papers 1 and 3 are duplicate, and possibly the Entity Resolution algorithm would consider them as not correspondent.

Now, consider that a subset of attributes was selected randomly, without considering the attribute values. Suppose a subset composed by the attributes Year and Venue. Probably, the instances 3 and 5 would be considered as duplicate ones, given that the values of these attributes have a high similarity value. This happens because repeated values can contribute to increase the similarity value, which can make two instances are given as corresponding even being distinct. This would result in an incorrect correspondence, known as false positive, which refers to an instance pair not correspondent as correspondent. On the other hand, the instances 1 and 2, and 4 and 5 would be given as not correspondent, resulting in two false negatives.

We can notice that using all attributes, or a subset of attributes selected randomly, may led to Entity Resolution results with a low F-measure[5]. Because of this, a Strategy of Attribute Selection, in which only the relevant attributes for the Entity Resolution process are selected, becomes necessary.

[3] http://citeseerx.ist.psu.edu

[4] http://dblp.uni-trier.de

[5] Measure that indicates the quality of the Entity Resolution process, calculating the harmonic average between the precision and recall values. More details will be presented in Section 5.

3 ENTITY RESOLUTION AND ATTRIBUTE SELECTION

Data Integration is a complex process, which can be divided into three main steps: Schema Alignment, Entity Resolution, and Data Fusion (Dong and Srivastava, 2015). In this work, we are interested in the Entity Resolution step, whose objective is to identify different instances referring to the same real world entity. The Entity Resolution, in its turn, is subdivided into the following activities: Blocking, Pair Comparison and Clustering (Christen, 2012; Dong and Srivastava, 2015), as shown in the Figure 1.

Figure 1: Activities of the Entity Resolution process (Dong and Srivastava, 2015).

In the Blocking activity, the goal is to decrease the number of comparisons between instances. For this, blocking techniques are used. The standard blocking uses a blocking key to partition the instances where preferably all instances of the same entity are within the same block. From this, in the Pair Comparison step a detailed comparison between the instance pairs is performed. Finally, in the Clustering the instances are clustered so that each cluster is composed of instances that refer to the same entity (Christen, 2012). In this work, we focus on the Pair Comparison activity.

During the Pair Comparison activity, similarity values of attributes describing two compared instances are evaluated. The higher the similarity values of these attributes, higher the chances these two instances will be equivalent. One of the main challenges is to determine which attributes to consider for the comparison. Intuitively, one can think that the higher the number of attributes, the better is the result of the comparison, given that we consider the maximum amount of information. However, some attributes may not be relevant for the pair comparison. Therefore, it is necessary to select a subset of attributes to be used during the comparison step, in order to maximize the quality of Entity Resolution results.

In general, the attribute selection aims to find a subset of attributes from the original set, removing those that are irrelevant or redundant, which may significantly improve the efficiency of the task performed (Dash et al., 2002). For this purpose, the following issues should be treated: identification of attributes common to all instances participating in the Entity Resolution and selection of an optimal subset of common attributes. In other words, we look for a set of attributes able to support the generation of entity correspondences that are true (Gu et al., 2003). In addition to data integration, the attribute selection has been subject of research in other Computer Science areas, such as: Data Mining and Machine Learning (Jouve and Nicoloyannis, 2005; Li et al., 2006). In these studies, it is intended, beyond improving process efficiency, reducing the dimensionality of the samples, minimizing the computational costs and also the data storage space.

4 A STRATEGY FOR SELECTING RELEVANT ATTRIBUTES FOR THE ENTITY RESOLUTION PROCESS

In this section, we present an overview of our strategy for attribute selection in the Entity Resolution process and its main tasks.

4.1 Overview

As mentioned before, the Entity Resolution process can be seen as part of a Data Integration process, whose main goal is to provide an integrated view of data distributed in multiple data sources.

An overview of the proposed attribute selection strategy is presented in the Figure 2 and described below. In the following, consider a set of data sources $F = \{f_1, f_2, ..., f_n\}$ that provides data about one or more entities E of the real world. An entity E in a given data source f_i has a set of instances $f_i.E = \{e_1, e_2, ..., e_m\}$ such that each e_j represents an instance of the real world, and is described by a set of attributes $E.A_i = \{a_{i1}, a_{i2}, ..., a_{ip}\}$. An instance e_j of an entity E in the data source f_i, denoted by $f_i.E.e_j$, is defined by a set of pairs $\{(a_{i1}, v_{i1}), ..., (a_{ip}, v_{ip})\}$, where $a_{ik} \in E.A_i$, and v_{ik} is the value of a_{ik} for the entity E in the data source f_i.

The Entity Resolution process receives as input a set of instances $E' = \{e_1, e_2, ..., e_v\}$ referring to the same entity of the real world. Each instance e_j from E' can be from a different data source f_i and is described by a set of attributes, which depends from the data source f_i where the instance comes from. As shown in Figure 2, the attributes selection task is performed as an auxiliary task of the Entity Resolution process and receives as input the set E' and the set A_{int}, which denotes the set of attributes common to all instances from E' (Gu et al., 2003).

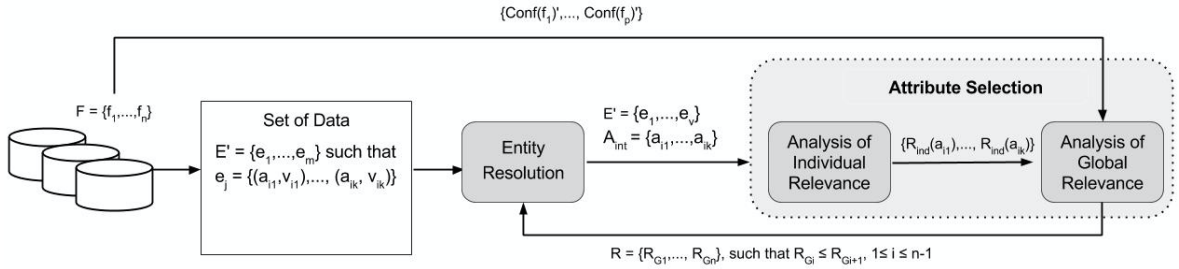


Figure 2: Overview of the Attributes Selection Strategy.

The attributes selection is divided in two main steps: (i) Analysis of Individual Relevance, which is based on characteristics related to the data, such as repetition and density values; (ii) Analysis of Global Relevance, which considers characteristics of the data sources, such as reliability. In order to support our strategy, we consider the existence of a metadata catalog (Oliveira et al., 2015) that stores information about the data sources being integrated, such as: name, url, schema (containing entities and attributes), and quality metadata, including the reliability of the data source.

During the Analysis of Individual Relevance, values for the repetition and density criteria are calculated, and so the Individual Relevance value is obtained for each $a_{ij} \in A_{int}$. The Analysis of Global Relevance of a given attribute a_{ij} consists of calculating a new value based on the value of the Individual Relevance of a_{ij} and the reliability value of the data source where the instances from E' come from.

As output of the attributes selection process, a set of global relevance values $R = \{RG_1, ..., RG_n\}$ is given. Each RG_k is a pair (a_{ij}, r_{ij}), where r_{ij} is the Global Relevance value of the attribute $a_{ij} \in A_{int}$.

4.2 Evaluation Criteria

In order to identify whether an attribute is relevant or not, first we should define the meaning of relevant attribute for the task in hand. Then, we can define the criteria to evaluate the relevance of a given attribute. In our case, the task in hand is the Entity Resolution process and the relevance of an attribute is proportional to its capability for discriminating instances that belong to different entities and in not discriminating instances that belong to the same entity. We also consider that the quality of the data sources can have a direct impact on the relevance of an attribute. For example, a data source that is likely to provide data that is not true. In this sense, weighting the relevance of an attribute according to the quality of the data sources can help to verify if the result of the selection of attributes is really reliable.

We identified six criteria to measure the relevance of attributes for the Entity Resolution process (Canalle, 2016). However, in this work only four criteria are used, two of them are related to data and are used to measure the individual relevance of attributes. In general, these characteristics include the level of errors and the number (and distribution) of attribute values, i.e., its informative content. For example, a field such as *sex* has only two possible values and, consequently, could not give enough information to identify similarity between instances. On the other hand, an attribute like *surname* contains more, however it is more prone to errors (Gu et al., 2003). The other two, related to the data sources, are used to measure the global relevance of a given attribute.

4.3 Analysis of Individual Relevance

In this section, we will detail each of the criteria chosen for the evaluation of the Individual Relevance of each attribute, specifying how it is calculated.

4.3.1 Repetition and Density

The repetition of an attribute a_{ij} is given by the number of times the same attribute value appears in the set of instances E'. The choice of this criterion for the attribute selection was motivated by the fact that using attributes with high repetition values can contribute to the generation of false positives, i.e., distinct instances classified as similar.

Given the set of instances E', for each attribute $a_{ij} \in A_{int}$ the value for the repetition criterion is calculated. In this paper, the value for the repetition criterion is calculated according to the Equation 1, where τ is the total number of distinct values of a_{ij}, and η is the total number of values of a_{ij}.

$$Rep(a_{ij}) = 1 - \left(\frac{\tau}{\eta}\right) \quad (1)$$

We use a similarity function to identify whether attributes have the same value. For this, there is a

number of functions proposed in the literature (Christen, 2012). In this work, the Levenshtein or Edit Distance similarity function was adopted. This function is frequently used to compare relatively small strings, which do not have necessarily the same size.

Given the set of instances E', for each attribute $a_{ij} \in A_{int}$, the density criterion is calculated. The density is given by the percentage of non null values in the set of values that describes it Naumann and Freytag (2000). During the instances comparison, missing values can make two corresponding instances be considered as distinct, once the comparison with missing values results in a similarity equal to 0 (zero). In this sense, empty values may contribute to the identification of false negative, in other words, similar instances classified as distinct. In this paper, the density value can be calculated according to the Equation 2, where α is the total number of nonnull values of a_{ij}, and β is the total number of values of a_{ij}.

$$Den(a_{ij}) = \frac{\alpha}{\beta} \quad (2)$$

4.3.2 Individual Relevance Calculation

The Individual Relevance R_{ind} of a given attribute a_{ij} is calculated based on repetition and density criteria, according to the Equation 3.

$$R_{ind}(a_{ij}) = Den(a_{ij}) * p_d + (1 - Rep(a_{ij})) * p_r \quad (3)$$

Where *Den* is the value of the density criterion for the attribute a_{ij}, *Rep* is the value of the repetition criterion for the attribute a_{ij},p_d is the weight for the density criteria, and p_r is the weight for the repetition criteria. Values of p_d and p_r are defined according to the importance degree of each criteria for the relevance of a given attribute, so that the sum of the weights must be equal to 1 (one).

4.4 Analysis of Global Relevance

The Global Relevance of an attribute $a_{ij} \in A_{int}$ is denoted by $R_{glob}(a_{ij})$. We evaluate the global relevance of an attribute a_{ij} in order to include data source quality information as part of the attributes selection process. This becomes necessary because data sources from F, which provide instances to E', may have low quality. Therefore, to calculate the $R_{glob}(a_{ij})$, in addition to the value of $R_{ind}(a_{ij})$, we also consider data sources quality information.

Specifically, in this work, the quality of a source is given based on two criteria: Reliability and Coverage. The Reliability of a source f_i concerns the degree to which the data provided by f_i is true and reliable (Wang and Strong, 1996). We assume that the sources have quality metadata associated to them (Mihaila et al., 2000), where the reliability value, denoted by $Rel(f_i)$ such that $0 \leq Rel(f_i) \leq 1$, can be obtained by means of these metadata.

Coverage of a data source f_i is defined by the percentage of instances that f_i provides to the set of instances E'. In this work, we use the evaluation metric proposal by Naumann and Freytag (2000) (Equation 4). The coverage of a data source f_i, denoted by $Cov(f_i)$, is calculated dividing the total number of instances that f_i provides to E', denoted by π, by the total of instances contained in E', denoted by $|E'|$.

$$Cov(f_i) = \frac{\pi}{|E'|} \quad (4)$$

After calculating the values of reliability and coverage for each data source $f_i \in F$, the quality of F, denoted by $Q(F)$, can be given by Equation 5.

$$Q(F) = \sum_{k=1}^{|F|} Rel(f_i) * Cov(f_i) \quad (5)$$

Finally, the $R_{glob}(a_{ij})$ can be calculated according to the Equation 6, where $R_{ind}(a_{ij})$ is the Individual Relevance value of the attribute a_{ij}.

$$R_{glob}(a_{ij}) = R_{ind}(a_{ij}) * Q(F) \quad (6)$$

5 EXPERIMENTAL EVALUATION

To evaluate the proposed strategy we used the Cora database[6]. This database contains 1.879 instances of different data sources related to literary productions. The instances are described by 15 attributes: *id, author, title, journal, volume, pages, year, publisher, address, note, venue, editor, type, institution and month*. Before performing our experiment, we performed a pre-processing step on the database to clean the data. For example, the attribute *year* has inconsistent characters with the format assigned to it (e.g.: "(1989)" instead of "1989").

Table 2: Scenarios with the duplicate data percentages.

Scenario	Scenarios of Duplicate Data
Scenario 1	5% - 10%
Scenario 2	15% - 30%
Scenario 3	35% - 50%
Scenario 4	55% - 70%
Scenario 5	>75%

Furthermore, the Cora database has a high percentage of duplicate data ($\pm$90%). Intuitively, we

[6]https://people.cs.umass.edu/ mccallum/code-data.html

believed that for a dataset with a high percentage of duplicate instances, any attribute selected for the instances comparison could provide a good result for the Entity Resolution process. For this reason, we expected that the huge number of duplication would lower the degree of effectiveness of our strategy. Then, we conducted our experiments in five different scenarios. Each scenario has a percentage of extracted duplicate data from Cora database, whose configuration is shown in Table 2. The goal of this experiment was to investigate how the proposed strategy would behave in each scenario and to confirm the following hypotheses:

H1 - The usage of all attributes in the comparison step of the Entity Resolution process results in a low F-measure, containing a high number of incorrect correspondences (false positives and false negatives), and a low number of correct correspondences (true positives and true negatives).

H2 - Considering just the most relevant attributes, identified by the proposed strategy, in the comparison step of the Entity Resolution process results in a high F-measure, i.e., a greater number of correct correspondences, with a smaller number of incorrect correspondences.

H3 - As less relevant attributes are added to the group of attributes considered in the comparison step of the Entity Resolution process, the number of incorrect correspondences increases, and the F-measure of the result decreases.

To evaluate our strategy, we considered the results obtained from an existing Entity Resolution tool, performed in different scenarios, using the attributes chosen by the proposed Attribute Selection. We used the DuDe Toolkit as Entity Resolution tool (Draisbach and Naumann, 2010), and the resolution algorithm was the Naive Duplicate Detection. For the similarity calculation, we adopt the Levenshtein function.

The Cora database has a Gold Standard containing the duplicate instance pairs, allowing to evaluate how good is the result of the Entity Resolution process. Specifically, we consider the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) obtained as the result of the Entity Resolution process. Based on these values, different quality measures can be considered (Christen, 2012). For the evaluation of the Entity Resolution, we adopted the F-measure, considering that this measure is recommended as the best to evaluate the quality of the Entity Resolution process (Christen, 2012). This measure uses the values obtained by calculating precision by Equation 7, and recall, by Equation 8, and it is calculated as shown in Equation 9.

$$precision = \frac{TP}{TP+FP} \tag{7}$$

$$recall = \frac{TP}{TP+FN} \tag{8}$$

$$F-measure = 2\left(\frac{precision * recall}{precision + recall}\right) \tag{9}$$

In these equations, TP is the number of corresponding instances correctly identified, FP is the number of not corresponding instances identified as corresponding, and FN is the number of corresponding instances identified as not corresponding.

5.1 Analysis of Results

Figure 3 presents the classification of each attribute according to the Attribute Individual Relevance analysis, as described in Section 4.3.2. In Figure 3, column *A* represents the attributes, *D* the density values, *R* are the repetition values, and *RI* the Individual Relevance values. It is important to emphasize that in the Entity Resolution process, instead of using a single attribute (or first attribute of the ranking), a set of attributes for comparing instance pairs should be used, because a single attribute may not be sufficient for the comparison phase.

In this sense, to validate our hypothesis, we have created four groups of attributes. These groups contain respectively the number of two, three and four most relevant attributes. In order to evaluate our approach, the Entity Resolution process was performed for each group. The obtained results are presented in Figure 4.

To evaluate the impact of Attribute Global Relevance, we consider the Group 1, which has a dataset with 40 instances about scientific publications. Taking into account the provenance of these instances and by assuming that 25 instances are from the data source 1 (DS1), that have a reliability of 90%. First, for each data source F_k, we calculate the coverage with regard to the dataset, according to Equation 2, obtaining as result $Cov(F1) = 0.625$, $Cov(F2) = 0.375$. Then, we calculate the quality of the data source set *DS*, composed by $DS1$ and $DS2$, following the Equation 3, obtaining $Q(F) = 0.76$. Thereby, for each $a_{ij} \in A_{int}$, the value of the Attribute Global Relevance is calculated considering quality information of the set of data sources $Q(F)$, based on the Attribute Individual Relevance value. In our experiment, for every attribute, the value of R_{glob} has been lower than R_{ind}, given that the source contributes with a lower quantity of instances and has a lower value of reliability.

A	D	R	RI
Title	1,00	0,05	0,98
Author	1,00	0,08	0,96
Pages	0,68	0,30	0,69
Year	0,85	0,55	0,65
Venue	0,63	0,48	0,58
Address	0,35	0,68	0,34
Publisher	0,45	0,78	0,34
Editor	0,30	0,75	0,28
Volume	0,23	0,75	0,24
Journal	0,15	0,83	0,16
Institution	0,15	0,85	0,15
Type	0,13	0,88	0,13
Note	0,10	0,88	0,11
Month	0,00	1,00	0,00

(a) Scenario 1

A	D	R	RI
Title	0,97	0,17	0,90
Author	0,77	0,20	0,79
Year	0,90	0,60	0,65
Pages	0,60	0,43	0,58
Venue	0,47	0,60	0,43
Volume	0,43	0,63	0,40
Address	0,3	0,67	0,32
Publisher	0,33	0,77	0,28
Journal	0,33	0,83	0,25
Editor	0,20	0,77	0,22
Institution	0,20	0,80	0,20
Note	0,13	0,83	0,15
Type	0,17	0,87	0,15
Month	0,00	1,00	0,00

(b) Scenario 2

A	D	R	RI
Title	1,00	0,34	0,83
Author	0,98	0,44	0,77
Year	0,90	0,62	0,64
Venue	0,68	0,60	0,54
Publisher	0,66	0,70	0,48
Pages	0,46	0,62	0,42
Address	0,46	0,74	0,36
Editor	0,38	0,76	0,31
Volume	0,34	0,78	0,28
Journal	0,20	0,80	0,20
Note	0,06	0,94	0,06
Institution	0,04	0,94	0,05
Type	0,04	0,94	0,05
Month	0	1	0

(c) Scenario 3

A	D	R	RI
Author	1,00	0,65	0,68
Title	0,90	0,63	0,64
Year	1,00	0,78	0,61
Pages	0,68	0,75	0,46
Volume	0,48	0,75	0,36
Journal	0,55	0,85	0,35
Venue	0,37	0,80	0,29
Address	0,35	0,80	0,28
Publisher	0,33	0,88	0,23
Editor	0,23	0,85	0,19
Note	0,10	0,88	0,11
Type	0,08	0,93	0,08
Institution	0,03	0,95	0,04
Month	0,00	1,00	0,00

(d) Scenario 4

A	D	R	RI
Author	1,00	0,74	0,63
Title	1,00	0,76	0,62
Year	0,84	0,80	0,52
Venue	0,85	0,82	0,52
Editor	0,54	0,78	0,38
Publisher	0,62	0,88	0,37
Pages	0,56	0,86	0,35
Address	0,44	0,82	0,31
Institution	0,10	0,88	0,11
Type	0,08	0,92	0,08
Volume	0,04	0,96	0,04
Journal	0,04	0,96	0,04
Note	0,02	0,96	0,03
Month	0,00	1,00	0,00

(e) Scenario 5

Figure 3: Individual Relevance values of the attributes.

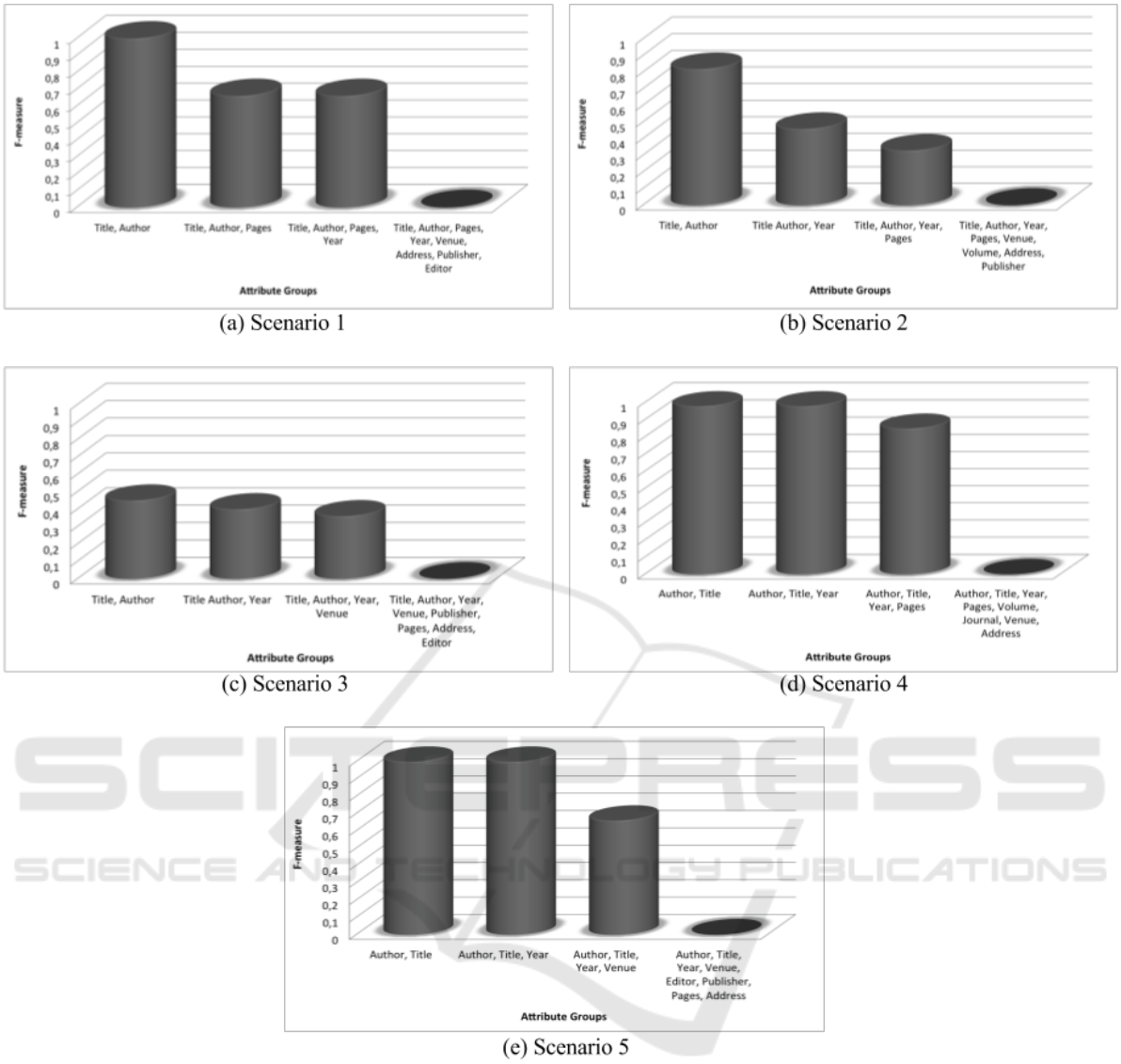
Analyzing the results, we can conclude that our strategy is efficient in all scenarios. In scenarios with a significant percentage of duplication (Scenario 4 and Scenario 5), we can observe that the relevance value of attributes is not so high when compared to other scenarios. This is due to the fact that, in scenarios with a significant percentage of duplication the value of the repetition criterion is always very high. Our strategy was able to correctly list the best attributes for such scenarios, proving its efficiency. Furthermore, we confirmed that using a large quantity of attributes in the Entity Resolution process is not acceptable, given the results obtained using the set of attributes of Group 4. It was also observed that the greatest F-measure of the Entity Resolution process was obtained using the group 1, with only the two most relevant attributes. We also realize that as 1. Thus, the results obtained through experiments validated our hypothesis.

With respect to the Global Relevance, we can verify that whether a source contributes to the majority of instances for a data set, but if this source has a low reliability, probably the attribute relevance should be contested. Thus, we concluded that the analysis of the Global Relevance can be useful, mainly to evaluate if the results of the Attribute Selection is really reliable.

6 RELATED WORKS

Entity Resolution is a research area that has attracted attention of researchers from different areas of Computer Science, such as Data Mining, Artificial Intelligence and Database. Because of this, several studies have been carried out proposing to solve the Entity Resolution problem in different ways, as using active learning (Sarawagi and Bhamidipaty, 2002), genetic programming (de Carvalho et al., 2010), and functional dependencies (Fan et al., 2009) (Caruccio et al., 2016). Several tools for the Entity Resolution process have also been proposed in the literature. In (Kopcke and Rahm, 2010), for example, a comparative evaluation of some tools that help in the Entity Resolution process is presented.

Considering the context of Entity Resolution, specifically in the data integration process, just few works discuss the attribute selection problem. Among them, we highlight Chen et al. (2012) and Su et al. (2010). Chen et al. (2012) proposes a method for Entity Res-



(a) Scenario 1 (b) Scenario 2

(c) Scenario 3 (d) Scenario 4

(e) Scenario 5

Figure 4: Results of the Entity Resolution process considering the attribute groups.

olution based on Machine Learning. This method searches, by means of a training set, the best group of attributes to be used in the comparison step of the Entity Resolution process. First, for each attribute the proper similarity measure is found, analyzing the F-measure value that each measure provided. Then, the groups of attributes are evaluated, where the groups with the highest values of F-measure are chosen as the best ones for the comparison step.

Su et al. (2010) proposes duplicate detection for results of queries in Web data sources. In this scenario, the work focuses on a Machine Learning algorithm that aims to adjust the weights of the attributes for the similarity evaluation. The algorithm is able to learn how to adjust the weights of the attributes using a data sample that contains instances without correspondence from different data sources. The authors argue that using this sample facilitates the Entity Resolution process, since in scenarios of queries for Web databases, the percentage of duplicate instances is far lower than the percentage of non-duplicate instances. To balance the attribute weights according to their importance, the authors use a non-duplicate data vector and a vector of duplicate data. For duplicate data vector higher weights are assigned to the attributes with high similarity between their values and low weights for attributes with low similarity in their values, and the opposite is done for the not duplicated vector.

Differently of (Chen et al., 2012) and (Su et al., 2010), this work proposes an Attribute Selection strat-

egy for the evaluation of relevant attributes using criteria related to the data and by means of metadata related to the data sources. Another differential of our work to the Chen et al. is that we do not need a training set. The definition of a training set can be a difficult task, specially in scenarios containing large volumes of data. Recently, some studies have been proposed in order to facilitate this task (Bianco et al., 2015).

7 CONCLUSIONS

In this work, we propose a strategy for selection of relevant attributes for the Entity Resolution process. This strategy consists of the following two steps: (i) Individual Relevance Analysis and (ii) Global Relevance Analysis. In the former, we analyze data features, such as repetition and density, to measure the individual relevance of an attribute. In the later, we refine results from earlier stages to weight the relevance of each attribute based on quality criteria of the data sources considered in the Entity Resolution.

For the purposes of evaluating the proposed strategy, we performed several experiments using the CORA dataset. These experiments have demonstrated that the groups of attributes selected by our strategy provide the best result for the Entity Resolution process, resulting in the validation of our hypothesis. In addition, we have made experiments with the Febrl dataset obtaining similar results.

As future work, we intend to include other criteria in the attribute selection process, such as the susceptibility of an attribute to contain errors (e.g. surname), the attribute dynamism, i.e., if the attribute contains values that may change over time (e.g. age). We believe that such characteristics can also be helpful for the selection of relevant attributes in the Entity Resolution process.

REFERENCES

Bianco, G. D., de Matos Galante, R., Gonalves, M. A., Canuto, S. D., and Heuser, C. A. (2015). A practical and effective sampling selection strategy for large scale deduplication. *IEEE Trans. Knowl. Data Eng.*, 27(9):2305–2319.

Canalle, G. K. (2016). Uma estratgia para seleo de atributos relevantes no processo de resoluo de entidades.

Caruccio, L., Deufemia, V., and Polese, G. (2016). Relaxed functional dependencies - a survey of approaches. *IEEE Trans. Knowl. Data Eng.*, 28(1):147–165.

Chen, J., Jin, C., Zhang, R., and Zhou, A. (2012). A learning method for entity matching. In *In Proceedings of 10th International Workshop on Quality in Databases*, East China Normal University, China.

Christen, P. (2012). *Data Matching*. Springer, Heidelberg.

Dash, M., Choi, K., Scheuermann, P., and Liu, H. (2002). Feature selection for clustering - a filter solution. In *ICDM*, pages 115–122. IEEE Computer Society.

de Carvalho, M. G., Laender, A. H. F., Goncalves, M. A., and da Silva, A. S. (2010). A genetic programming approach to record deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints).

Dong, X. L. and Srivastava, D. (2015). *Big Data Integration*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.

Draisbach, U. and Naumann, F. (2010). Dude: The duplicate detection toolkit. In *In Proceedings of the International Workshop on Quality in Databases (QDB)*.

Fan, W., Jia, X., Li, J., and Ma, S. (2009). Reasoning about record matching rules. *PVLDB*, 2(1):407–418.

Gruenheid, A., Dong, X. L., and Srivastava, D. (2014). Incremental record linkage. *PVLDB*, 7(9):697–708.

Gu, L., Baxter, R., Vickers, D., and Rainsford, C. (2003). Record linkage: Current practice and future directions. Technical report, CSIRO Mathematical and Information Sciences.

Jouve, P.-E. and Nicoloyannis, N. (2005). A filter feature selection method for clustering. In Hacid, M.-S., Murray, N. V., Ras, Z. W., and Tsumoto, S., editors, *ISMIS*, volume 3488 of *Lecture Notes in Computer Science*, pages 583–593. Springer.

Kopcke, H. and Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data Knowl. Eng.*, 69(2):197–210.

Li, Y., Lu, B.-L., and Wu, Z.-F. (2006). A hybrid method of unsupervised feature selection based on ranking. In *ICPR (2)*, pages 687–690. IEEE Computer Society.

Mihaila, G. A., Raschid, L., and Vidal, M.-E. (2000). Using quality of data metadata for source selection and ranking. In *WebDB (Informal Proceedings)*, pages 93–98.

Oliveira, M. I. d. S., Lscio, B., and Gama, K. (2015). Anlise de desempenho de catlogo de produtores de dados para internet das coisas baseado em sensorml e nosql. *XIV Workshop em Desempenho de Sistemas Computacionais e de Comunicao*.

Sarawagi, S. and Bhamidipaty, A. (2002). Interactive deduplication using active learning. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–278, New York, NY, USA. ACM.

Su, W., Wang, J., Lochovsky, F. H., and Society, I. C. (2010). Record Matching over Query Results from Multiple Web Databases. *IEEE Transactions on Knowledge and Data Engineering*, 22(4):578–589.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33.