ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Review on Rule Based Entity Identification for Data Cleaning

Ankita Saxena[1], Prof. Ranjana Dahake[2]

[1]ME Student, Dept. of Computer Engineering, MET BKC, University of Pune, Nashik, Maharashtra, India

[2]Professor, Dept. of Computer Engineering, MET BKC, University of Pune, Nashik, Maharashtra, India

ABSTRACT: In real-world scenario entity may appear in multiple data sources so that the entity may have quite different descriptions. Hence, it is necessary to identify the records referring to the same real-world entity, which is named as Entity Resolution (ER).This paper highlightsER as one of the most important problems in data cleaning and arises in many applications such as information integration and information retrieval. Traditional ER approaches are in sufficient to identify records based on pair wise similarity comparisons, which assumes that records referring to the same entity are more similar to each other than otherwise. However for certain circumstances this assumption does not always hold in practice and similarity comparisons do not work well when such assumption breaks. So to overcome traditional ER drawback a new set of rules which could describe the complex matching conditions between records and entities is proposed such as rule discovery algorithm and rule based ER algorithm.

KEYWORDS: Entity Resolution, Data Cleaning, Rule Learning**.**

I. INTRODUCTION

In various application areas, data from multiple sourcesoften needs to be matched and aggregated before it can be used for further analysis or data mining. Data quality is high priority in all information systems. As it is a key step in obtaining clean data, record linkage, entity identification or entity resolution (ER) to analyze the records referring to the same real-world entity. Entity resolution can also be referred as object matching, duplicate identification, record linkage, or reference reconciliation as essential task for data integration and data cleaning. For example, two firms may want to merge their customer records. In such situation the same customer may be represented by multiple records, so these matching records must be identified and combined (into what we will call as a cluster). This ER process is highly expensive due to very large data sets and complex logic that decides when records represent the duplicate entity. It is the objective of ER identifying entities referring to the same or duplicate real-world entity. The high importance and difficulty of the entity resolution problem has given rise to a huge amount of researchers to focus on different variations of the problem and numerous approaches have been proposed to resolve such problem.

A common scenario with rule-based matching can be taken as paper publish with respective paper author and coauthor, where the goal is to group and merge paper author records according to the real-life entities. Here pairwise matching is carried out based on name or coauthor equality, until we get an entity consisting all four records resolve to its respective entity.[1]Note, that e.g. the third and fourth records do not match directly, we can reason only indirectly that they belong to the same person. As shown in Table 1.Traditional ER approaches obtain a result based on similarity comparison among records, assuming that records referring to the same to each other. However, such property may not hold in some cases traditional ER approaches cannot identify records correctly. A Review on Rule Identification for data cleaning approach is based on based paper [1].

Name	Coauthor	Title
Wei Wang	Zang	Inferring…
Wei Wang	Lin,Pei	Threshold…
Wei Wang	Lin,Hua,Pie	Ranking…
Wei Wang	Shi,Zang	Picture Book…

Table 1: Matching Customer records

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Example 1. Table 2 shows seven authors with name "weiwang" identified by o_{ij}s. By viewing to the authors home pages containing their publications manly divide the seven authors into three clusters. The records with IDs o_{11}, o_{12}, and o_{13} refer to the person in UNC, express as e_1, the records with IDs o_{21} and o_{22} refer to the person in UNSW, express as e_2, and the records with IDs o_{31} and o_{32} refer to the person in Fudan University, denoted as e_3. The function of entity identification is to identify e_1, $e_2$and $e_3$using the information in Table 2.

	id	name	coauthors	title
e_1	o_{11}	wei wang	zhang	inferring...
	o_{12}	wei wang	duncan, kum, pei	social...
	o_{13}	wei wang	cheng, li, kum	measuring...
e_2	o_{21}	wei wang	lin, pei	threshold...
	o_{22}	wei wang	lin, hua, pei	ranking...
e_3	o_{31}	wei wang	shi, zhang	picturebook...
	o_{32}	wei wang	pei, shi, xu	utility...

Table 2 Paper-Author Records

Based on the observations, we can develop the following rules to identify records in Table 2.

- R1: ∀o_i, if oi[name] is "weiwang" and oi[coauthors] includes "kum", then o_i refers to entity e_1;
- R2: ∀ o_i, if oi[name] is "weiwang" and oi[coauthors] includes "lin", then o_i refers to entity e_2;
- R3: ∀oi, if o_i [name] is "weiwang" and oi[coauthors] includes "shi", then o_i refers to entity e_3;
- R4: ∀oi, if oi[name] is "weiwang" and o_i [coauthors] includes "zhang" and excludes "shi",then o_i refers to entity e_1.

Rule based method for Entity Resolution (ER) is being posed when a user want to retrieve data to identity the records referring to the same real world entity. Rule based method has defined its Entity Resolution rule such as it consist of two clauses (1) The If clause includes constraints on attributes of records and (2) the Then clause indicates the real world entity referred by the records that satisfy the first clause of the rule. Thus, we use A => B to express the rules "∀o, If Record o satisfies A Then o refers to B" for ER. Thus the left-hand side and the right-hand side of a rule r denoted as LHS(r) and RHS(r) respectively.

The rest of the paper is structured as follows Section II relevant literature work, Section III provides details about system flow, Section IV describes actual algorithmic strategy and Section V concludes the paper.

II. RELATED WORK

The work on entity resolution can be mainly divided into three categories.

A] Pairwise ER: Most works on ER focus on record matching which comprise of comparing record pairs and identifying whether they match to same real world entity. Most of the work limelight on record matching similarity functions. Acquisition string variations is proposed for transformation-based framework to match records based on both with and without using machine learning to find suitable parameterization and combination of similarity functions. Traditional ER in which records are compared with each other but in R-ER is orthogonal record matching is used. However, string similarity functions can be applied to fuzzy match operator (denoted by ≈) in ER-rules. For example, given a string *s*, we say as s ≈"wei wang" if the edit distance between *s* and "weiwang" is smaller than a given threshold. Decision trees are employed to get exact record matching rules as describe by S. Tejada, C. Knoblock, and S. Minton [11].As decision trees cannot be used to discover ER-rules because the domain of the right hand side of record matching rules depend on {yes, no} (two records are mapped or not mapped), while the domain of the right hand side of ER-rules result as an entity set.

B] Non-pairwise ER: Research on non-pairwise ER includes clustering strategies [10] and classifiers. Most strategies resolve ER based on the relationship graph among records, by modelling the records as nodes and the relationships as edges. Machine learning approaches [6] are also proposed by using global information to resolve ER effectively. However, these methods are not suitable for huge data because of efficiency issues.

C] Scaling: ER algorithm treated as black box and limelight on developing scalable framework for ER. Indexing techniques used for ER have been surveyed by Christen[2]. In [5] S. E. Whang and H. Garcia-Molina limelight on how

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

to update ER results efficiently when ER logic evolves. These techniques are orthogonal can be used to get accelerate rule-based ER algorithm.
R-ER focus on pair-wise ER rule-based approaches [7] are closer to the approaches define in [1] these rules differ as they focus on determining whether two records refer to the same entity while the paper focus ondetermining whether a record refers to an existing entity.

III. SYSTEM FLOW

It is based on the work of scientist Lingli Li, Jianzhong Li, and Hong Gao has introduce a new class of rules which could describe the complex matching conditions between records and entities. Based on this class of rules, the rule-based entity resolution problem describes an on-line approach for ER. In this framework, by applying rules to each record and to identify which entity the record refers to is major objective of rule based entity identification method. As below figure 1 explain the whole system flow process such as Input Data set comprise of dblp data is a selection from DBLP Bibliography3 and kdd data4 is the validation data set for Track 1 of KDD Cup 2013.Rule Discovery algorithm which comprise of few requirements based on syntax and semantics rules are define and for solving entity resolution problem an efficient rule-based algorithm is introduce while entity if information is changed a rule maintaining method refer as rule update is used. This all process results classified entity set which comprise of scan records one by one and determines the entity for each record.

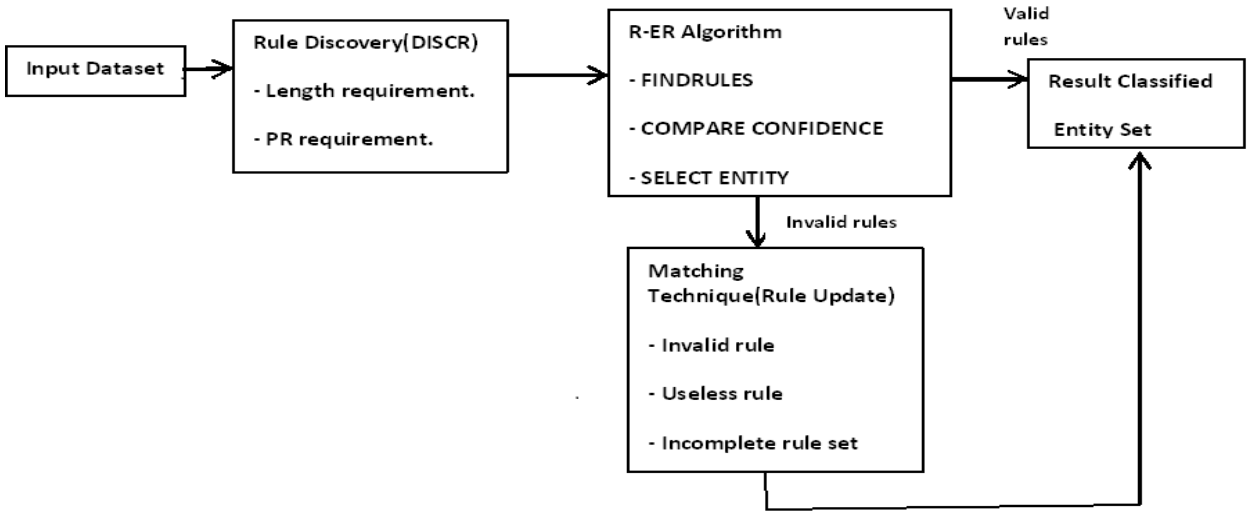

figure1: System Flow

IV. ALGORITHM STRATEGY

A] RULE DISCOVERY (DISCR)

For the convenience of forthcoming discussion some concepts are introduced first related to rule discovery in the figure 2.
ER-rules into two categories:

(i) PR is an ER-rule which only includes positive clauses.
Example

(ii) NR is an ER-rule which includes at least one negative clause.

Syntax are define as per the based paper[1]:- An ER-rule is syntactically defined as $T_1 \wedge ... \wedge T_m$) e, where $T_i (1 \leq i \leq m)$isaclause with the form of$(A_i o_{pi} v_i)$,$(v_i op_i A_i)$, ¬ $(A_i op_i v_i)$ or¬ $(v_i op_i A_i)$, where A_i is an attribute, vi is a constant in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

the domain of A_i and op_i can be any domain- dependent operator defined by users, such as exact match operator =, fuzzy match operator ≈ for string value ≤, for numeric value, or ∈ for set value. The clause with form ($A_i op_i v_i$) or ($v_i op_i A_i$) is called positive clause, and the clause with form ¬($A_i op_i v_i$) or¬ ($v_i op_i A_i$) is called negative clause.

Each ER-rule r can be assigned a weight *w*((r) in [0,1] as specify in the eq (1) is to reflect the level of confidence that r is correct. Intuitively, the more records are identified by an ER-rule r, the more possible r is correct. Therefore, given a data set S, we define the weight of each ER-rule r as:

$$w(r) = \frac{|S(r)|}{|S(RHS(r))|} \qquad \text{eq (1)}$$

where S(r) denotes the records in S that are identified by rand S(RHS(r)) denotes the records in S that refer to entity RHS(r)

Semantics are define as per the based paper [1]:- In the following definitions, we let o be a record, S be a data set, r be an ER-rule and R be an ER-rule set such as: Definition 1: o matches the LHS of r if o satisfies all the clauses in LHS(r).o matches the RHS(r) if o refers to entity RHS(r). Definition 2: o satisfies r, denoted by o ‘ r, if o does not match LHS(r) or matches RHS(r). Definition3: o is identified by r, if o matches both LHS(r) and RHS(r). Note that, if o is identified by r, o must satisfy r. If o satisfies r, o might not be identified by r.

Properties of ER-Rule Set comprise of :- Given an ER-rule set R and a data set S, to ensure R performs well on S, it require (1) there is no false matches between record and entity (validity); (2) there is no conflicting decisions by R (consistency); (3) each record in S can be mapped to an entity by R (completeness) and (4) there is no redundant rules in R (independence).Based on the syntax and semantics of the Rule Based Entity is used for an efficient Rule Based algorithm.

Example 2: The below rules are defining taken into consideration syntax and semantics as describe above for given Example 1 can be expressed as the following ER-rules respectively. For simplicity we write coa rather than coauthors.

r_1: (name =“weiwang”) ^ (“kum”2∈ coa) =>e_1,

r_2: (name = “weiwang”) ^ (“lin” ∈coa) => e_2,

r_3: (name=“weiwang”) ^ (“shi” ∈coa) =>e_3,

r_4: (name = “weiwang”) ^ (“zhang” ∈coa) ^ (“shi” ∈ coa)) => e_1,

For example, r_1, r_2 and r_3 in Example 2 are all PRs while r_4 is an NR.

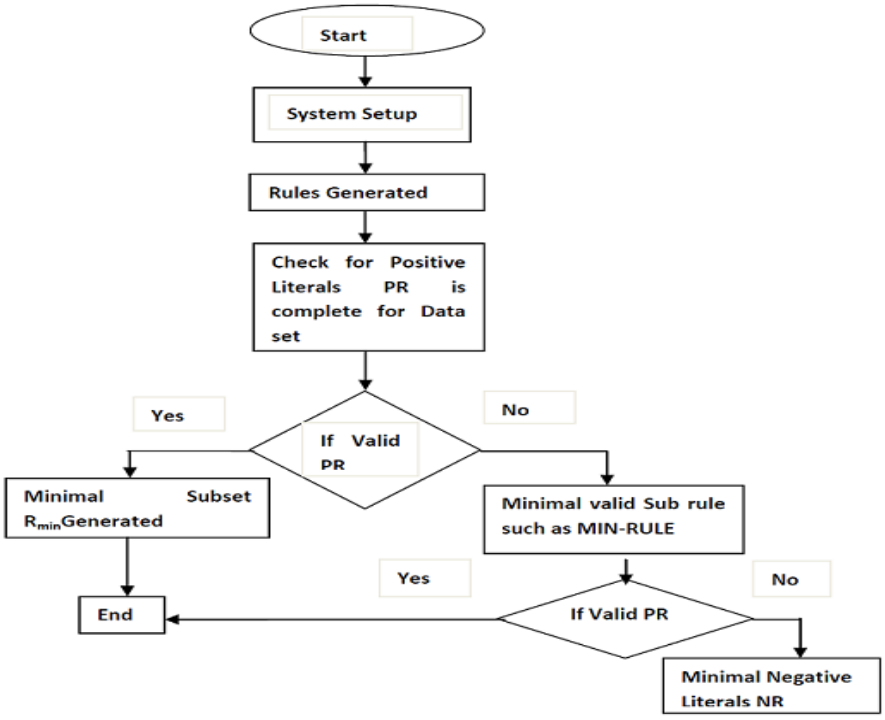

Figure 2: Flow Chart for DISCR Algorithm

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Coverage: Coverage of clause T on dataset S can be express as $Cov_S(T)$, is the subset of S such that $Cov_S(T)$ ={o|o∈ S, o satisfies T}.

Basic Requirements for Rule Discovery:

(i) Length Requirement: Given a threshold *l*, each rule r in R satisfies that $|r| \leq l$.

(ii) PR Requirement: each rule r in R is a PR.PR are describe as positive literal.

Rule Generated: Rules are generated using ER rules categories based on requirements of DISCR.

R_{min} Generated: Consist of Minimal subset rules which are generated using its basic requirements.

MIN-RULE: If rules generated do not satisfied the requirements of DISCR then they are termed as negative literals NR.

B] RULE BASED ENTITY RESOLUTION (R-ER)

Rule-based ER algorithm R-ER scans records one by one and determines the entity for each record. The process mainly divided into 3 main steps such as describe in system flow figure 1:

(i) FINDRULES: It is used to find all the rules satisfied by record.

(ii) COMPARE CONFIDENCE: It is to evaluate for each entity that which record might refer to the compute confidence specify as record o refers to entity e according to the rules of entity that are satisfied by record.

(iii) SELECT ENTITY: To select the entity with the largest confidence to which record might refer, and if this confidence is larger than a confidence threshold, it results that records o refers to entity e.

(iv) RULE UPDATE: The discover rules set might be invalid, incomplete, or contain useless rules if the training data is incomplete or out-of-date. Thus,to ensure the performance of the discover rule set on new records, below an evolution method of rules are introduce.

a) Invalid rule: A rule r is invalid if their exist records that match LHS(r) but do not refer to RHS(r).
b) Useless rule: An ER-rule r is called a useless rule if Cov(r) =Ø, since no records are identified by r.
c) Incomplete rule set: An ER-rule set R of entity set E is incomplete if there are records referring to entities in E that are not covered by R.

When rules are update or new rules are discover by exploiting users feedback at last it is simple to determine among the set of rules, which one should be deleted or inserted so as to update rule set accordingly to get the final result as classified entity set.

V. CONCLUSION AND FUTURE WORK

Rule based method is proposed to match complex matching conditioned between records and entities which comprise of new class of rules. Based on the syntax and semantics of the Rule Based Entity an efficient Rule Discovery (DISCR) algorithm is defined which includes few primary requirements. Rule based ER algorithm scans records one by one and determines the entity for each record. R-ER achieves good conduct both on efficiency and accuracy. Rule based method and traditional ER approaches can be considered as the complementary to each other and be applied together because rule-based method can identify records which cannot be solved by traditional ER methods and traditional ER methods can identify most of the records effectively which do not require availability of correct entity set.

REFERENCES

1. LingliLi, JianzhongLi, and Hong Gao," Rule-Based Method for Entity Resolution" IEEE Trans. Knowl. Data Eng.,vol. 27, no.1, pp. 250–263, Jan. 2015.
2. P. Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication",IEEE Trans. Knowledge Data Eng., vol. 24, no. 9, pp. 1537-1555, Sept. 2012.
3. M. Herschel, F. Naumann, S. Szott, and M. Taubert, "Scalable iterative graph duplicate detection",IEEE Trans. Knowl. Data Eng.,vol. 24, no. 11, pp. 2094–2108, Nov. 2011.
4. H. Kopcke, A. Thor, and E. Rahm, "Evaluation of entity resolution approaches on real-world match problems",Proc. VLDB Endowment,vol. 3, no. 1, pp. 484–493, 2010.
5. S. E. Whang and H. Garcia-Molina, "Entity resolution with evolving rules", Proc. VLDB Endowment, vol. 3, no. 1, pp. 1326–1337,2010.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

6. I.Bhattacharya and L.Getoor, "Collective entity resolution in relational data," Proc. VLDB Endowment, vol. 3, no. 1, p. 5, 2010.
7. F.Wenfei, J.Xibei, L.Jianzhong, and M. Shuai, "Reasoning about record matching rules" ,Proc. VLDB Endowment, vol. 2, no. 1,pp. 407–418, 2009.
8. A.Arasu, S.Chaudhuri, and R.Kaushik, "Transformation-based framework for record matching" ,in Proc. 24th Int. Conf. Data Eng.,2008, pp. 40–49.
9. R.Bekkerman and A. McCallum, "Disambiguating web appearances of people in a social network," in Proc. 14th Int. Conf. World Wide Web, 2005, pp. 463–470.
10. N.Bansal, A.Blum, and S.Chawla, "Correlation clustering",Mach. Learn., vol. 56, no. 1–3, pp. 89–113, 2004.
11. S.Tejada, C. Knoblock, and S. Minton, "Learning object identification rules for information integration", Inf. Syst., vol. 26, no. 8,pp. 607–633, 2001.