IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Using Rule Based and Blocking Approaches to accomplish Entity Identification for Data Cleaning

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

PG Student, Department of Computer Engineering, MET BKC, University of Pune, Nasik, Maharashtra, India[1]

Asst. Prof of Department of Computer Engineering, MET BKC, University of Pune, Nasik, Maharashtra, India[2]

Abstract: In today's scenario entity appear in multiple data sources so it is necessary to identify the records referring to the same real-world entity, which is named as Entity Resolution (ER).ER is one of the most substantial problems in data cleaning and ascends in many applications such as information integration and information retrieval. Familiar ER approaches are in sufficient to identify records based on pair wise likeness 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 likeness comparisons do not work well when such assumption breaks. So to overcome outdated ER drawback a new set of rules which could describe the complex matching conditions between records and entities is proposed such as rule discovery algorithm, rule based ER algorithm along with blocking scheme methods to get more resolved classified entity set.

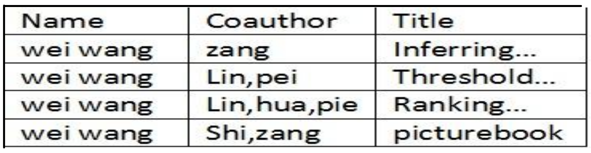
Keywords: Entity Resolution, Data Cleaning, Rule Learning and Meta blocking.

I. INTRODUCTION

In several application data from multiple sourcesoften needs to be matched and gathered before it can be used for further analysis or data mining. Also 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 stating to the same real-world entity. Entity resolution can also be stated as object matching, duplicate identification, record linkage, or reference reconciliation as essential task for data integration and data cleaning. ER can be performed in two ways using rules and blocking methods. Data blocking is the most populartechniques, groups like entity profiles into blocks and absolutely to perform the comparisons within each block. For example, two organizations may want to merge their customer records. In such situation customer may be represented as alikeby multiple records, so these matching records must be well-known and combined (into cluster). This ER process is highly costly due to very large data sets and complex logic that decides when records represent the duplicate entity.ER problem has given rise to a substantial amount of researchers to emphasis on different variations of the problem and numerous approaches.

A usual scenario with rule-based matching can be taken as paper publish with respective paper author and co-author, where the objective is to group and merge paper author records according to the real-life entities. Here pairwise matching is carried out based on name or co-author likeness, until we get an entity consisting all four records resolve to its respective entity.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. Usual ER approaches obtain a result based on similitude comparison among records, assuming that records referring to the same to each other. However, such property may not hold in some cases outdated ER approaches cannot identify records correctly [1].

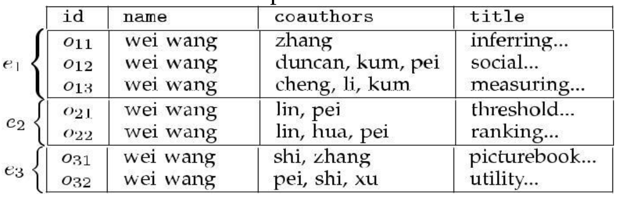
Table 1: Matching Customer records



Name	Coauthor	Title
wei wang	zang	Inferring...
wei wang	Lin,pei	Threshold...
wei wang	Lin,hua,pie	Ranking...
wei wang	Shi,zang	picturebook

Example:1. Table 2 shows seven authors with name "weiwang" acknowledged 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} state to the person in UNSW, precise as e_2, and the records with IDs o_{31} and o_{32}signify to the person in Fudan University, denoted as e_3. The function of entity affinity is to be identify ase_1, $e_2$and $e_3$using the information in Table 2.

Table 2 Paper-Author Records



	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...

Based on the observations, we can develop the following rules to identify records in Table 2 which is based on proposed system rule in section III

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

The rest of the paper is structured as follows. Section II reviews relevant literature survey, section III consists of proposed system, sections IV consists of experimental result and Section V concludes the paper.

II. LITERATURE SURVEY

Attempts are mainly taken to explore entity resolution into four categories.

A] Pairwise ER: ER emphasis on record matching which comprise of associating record pairs and recognizing whether they match to same real world entity. Most of the work fame 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 likeness functions. Outdated ER in which records are compared with each other but in R-ER is orthogonal record matching is used. However, string resemblance functions can be applied to fuzzy match operator (denoted by ≈) in ER-rules. For example, given a string s, we say s ≈ “wei wang” if the edit distance between s and “wei wang” is smaller than a given threshold. Decision trees are employed to get precise record matching rules as describe by S. Tejada, C. Knoblock, and S. Minton [14].As decision trees cannot be used to determine ER-rules because the area 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 embraces clustering approaches [13] and classifiers. Most methodologies resolve ER based on the relationship graph among records, by signifying the records as nodes and the relationships as edges. Machine learning methods [9] are also proposed by using global information to resolve ER resourcefully. However, these methods are not suitable for massive data because of efficiency issues.

C] Scaling: ER algorithm treated as black box and eminence on emerging scalable framework for ER. Indexing techniques used for ER have been surveyed by Christen[5]. [8] S. E. Whang and H. Garcia-Molina eminence on how to update ER results appropriately when ER logic evolves. These methods are orthogonal can be used to get stimulate rule-based ER algorithm. S. Whang, D. Marmaros has examines how to enlarge the progress of ER with a restricted amount of work using “hints,” which give information on records that are eventual refer to the same real-world entity. A hint can be represented in various formats (e.g., a grouping of records based on their likelihood of matching) and ER can use this information as a guideline for which records to compare first[3]. R-ER focus on pair-wise ER rule-based methods [10] are closer to the methods define in [1] these rules differ as they emphasis on determining whether two records refer to the same entity while the paper emphasis on determining whether a record refers to an existing entity.

D] Blocking Methods: Meta-blocking aims at extracting the most like pairs of entities by leveraging the information that is summarized in the block-to-entity relationships [2].The semantic-aware locality-sensitive hashing [LSH] blocking outline takes into consideration both textual and semantic likenesses in the ER blocking process. Semantic Information can be leveraged to progress the blocking quality and the integration of textual likeness and semantic likeness with the LSH technique provide resourceful and scalable blocking technique for ER with improved quality [4].

III.PROPOSED SYSTEM

System comprise of basically two methods such as Rule based and blocking approaches. Rule based method for Entity Resolution (ER) is being tendered when a user want to retrieve data to identity the records referring to the same real world entity. Blocking method which comprise of the groups of similar entity profiles of author co-author as a blocks to perform building blocks of entities, weighting scheme and pruning to get classified entity set which is used to examine records one by one and conclude the entity for each record.

A. System Flow

Input to the system is paper author’s data which is divided into groups corresponding to the authors identities. Input data set is pre-processed into clusters according to the user based then eventually rules based and blocking methods are used to perform objecting matching. Rule Discovery algorithm which comprise of few requirements which define syntax and semantics rules for generating ER set of Rules. Effectual rule-based algorithm is used to find rules, compare confidence and select entity with large confidence value as a set of entity profiles.In case if entity information is changed or incomplete or invalid a rule maintaining technique refers as rule update to produce set of resolved entity profiles. Set of resolved entity blocks caused by R-ER algorithm act as entity profiles which is been used as input profile for building blocks algorithm each entity blocks are created and using weighting scheme each entity is assign weight respective. Lastly pruning algorithm is functional using Weight Edge pruning(WEP) and Cardinality EdgePruning (CEP) to get more classified entity set.Fig.1 shows the System flow as per the proposed system.

DOI 10.17148/IJARCCE.2016.5776

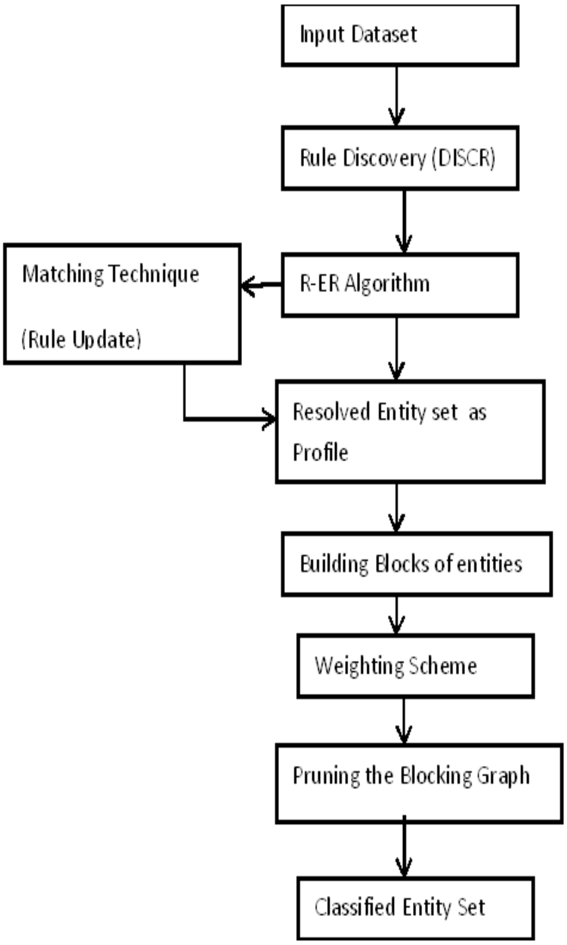


Fig.1.System Flow for Rule Based and Blocking Approaches

A. Algorithmic Strategy

Algorithmic strategy comprise of four vital algorithms such as:

i] Rule Discovery(DISCR).
ii] Rule Based Entity Resolution(R-ER).
iii] Building Blocks of Entities.
iv] Pruning the Blocking Graph.

i] Rule Discovery (DISCR)

Rule based method has defined its Entity Resolution rule such as it consist of two clauses (1)the If clause contains constraints on attributes of records and (2) the Then clause designates the real world entity referred by the records that satisfy the first clause of the rule. Thus, we use A => B to precise the rules "∀o, If Record o pleases A Then o refers to B" for ER. Thus the left-hand side and the right-hand sideof a rule r denoted as LHS(r) and RHS(r) correspondingly.

For the amenity some concepts are introduced first related to rule discovery in the Algorithm 1.

ER-rules into two categories:

a. PR is an ER-rule which should only embraces of positive clauses.
b. NR is an ER-rule which should embraces of at least one negative clause.

Syntax are define as per the based paper[1]:- An ER-rule is syntactically outline as (T_1 ^...^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 perpetual in the domain of $A_i$and op_i can be any domain-dependent operator prescribe by users, such as definite 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 theclause with form ¬(A_i op_i v_i) or¬ (v_i op_i A_i) is called negative clauses.

Semantics are define as per the based paper [1]:- In the following description, 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 equivalent to the LHS of r if o content all the clauses in LHS(r).o equivalent to the RHS(r) if o states to the entity RHS(r). Definition 2: o satisfies r, denoted by o ‘ r, if o does not equivalent to LHS(r) or RHS(r). Definition3: o is recognized by r, if o is corresponding to both LHS(r) and RHS(r). Note that, if o is recognized by r, o must satisfy r. If o satisfies r, o might not be recognized 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 untrue matches between record and entity (validity); (2) there is no conflicting decisions by R (consistency); (3) each record in S can be charted to an entity by R (completeness) and (4) there is no superfluous 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 co-authors.

r_1: (name ="wei wang") ^ ("kum"2∈ coa) =>e_1,
r_2: (name = "wei wang") ^ ("lin" ∈ coa) => e_2,
r_3: (name="wei wang") ^ ("shi" ∈ coa) =>e_3,
r_4: (name = "wei wang") ^ ("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.

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: Assumed a threshold l, each rule r in R contents |r| ≤l.

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

Algorithm 1: Rule Discovery (DISCR)

INPUT: Let length threshold l=2 and training data S = {S_1,….,S_m}.

```
OUTPUT: ER-rule set R
1.      Begin
Union of$S_1$,….,$S_m$ to get Training data set S.
GEN-PR are rules form using  (l, S) to get ER-rule set R
2.      IfCov(R) doesn't covered in  training data set S then S' is all set of rules not covered by R using S\ Cov(R);
3.      for each record o are in S' do
  if $r_0$ is valid then insert MIN-RULE($r_0$) in the ER-rule set R else insert in GEN-SINGLENR(o) set
     end if
   end for
end if
$R_{min} \leftarrow$ GREEDY-SETCOVER(R.S);
```

Rmin Generated: Comprise of Minimal subset rules which are produced using its basic requirements.

MIN-RULE: If rules produced do not fulfill the requirements of DISCR then they are termed as negative literals GEN-SINGLENR.

ii] Rule Based Entity Resolution (R-ER)

Rule-based ER algorithm R-ER scans all records one by one and determines the entity for each record. The process mainly divided into 3 main steps such as describe in Algorithm 2:

(a) FINDRULES: To find all the rules are fulfilled by record o.
(b) COMPARE CONFIDENCE: For each entity e to which record o might represent, then estimate the confidence that o specify to e according to the rules of e that are content by o.
(c) SELECT ENTITY: It is to select the entity e with the largest confidence to which o might represent, and if this confidence is more than a confidence threshold, it is determined that o refers to e.

Algorithm 2: Rule Based Entity Resolution (R-ER)

```
INPUT: Uis Data set,$R_E$ is an ER-rule set of entity set Eand $\theta c$ is largest confidence.
OUTPUT: U is set of resolved entity profiles

1.      Begin Initialize for each entity e in E do U belong to ∅.
2.      FINDRULES (o) for each record o using proposed system association rules then return R(o).
3.      COMPCONF(R) confidence is measure using equation 1 then return C.
4.      SELENTITY (o, $\boldsymbol{\theta}$c) Select entity e which as largest confidence among all the entities.
If C ≥ $\theta c$then add record o to U
end if
```

(d) RULE UPDATE: The discover rules set might be invalid, incomplete, or contain useless rules if the training data is incomplete or out-of-date. Then to confirm the performance of the discover rule set on new records evolution method of rules [8] are used to delete, insert or update rule set accordingly to get the final result as resolved entity set.

Each ER-rule r can be allocated a weight w(r) in [0,1] to reflect the level of confidence that r is correct is specify in the equation1. Apparently, the more records are recognized 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:

$$w(r) = \frac{|S(r)|}{|\,S(RHS(r))|} \quad ......(1)$$

Where,
S(r) denotes the records in S that are identified by r and S(RHS(r)) denotes the records in S that refer to entity RHS(r)

iii] Building Blocks of Entities

Resolved entity set obtained by R-ER algorithm act as entity profiles (such as p_i and p_j) and input for the building blocks of entities. It is the procedure of mining the blocking graph from a bilateral block collection B .Graph materialization block comprisesof a conceptual model that aims at simplifying the clarification and the development of blocking techniques. In the context of huge entity collections of entities (nodes) and comparisons of edges, its materialization actually poses noteworthy technical challenges. For this reason, it can be indirectly executed in two ways:(i) through inverted indices, which related each entity with the list of the blocks containing it, and (ii) with the help of bit arrays, which signify each entity as a vector with a zero value in all places, but those resultant to the blocks containing it.

Algorithm 3: Building Blocks of Entities

```
INPUT: B a block collection and WS a weighting scheme.
OUTPUT: $G_B$ corresponding blocking graph.

1.      BeginInitialize $V_B$ as node and $E_B$as edge an empty graph.
2.      {first iteration}
foreach $b_i \in$B  \\ to check all blocks
  {second iteration}
 foreach  $p_i \in b_i^1$do \\ tocheck all comparisons
$V_B \leftarrow V_B \cup \{v_i\}$; \\$v_i$ is degree of nodes.
{third iteration}
foreach$p_j \in b_i^2$do
$V_B \leftarrow V_B \cup \{v_j\}$; \\ to add node for $p_j$
$E_B \leftarrow E_B \cup \{e_{i,j}\}$; \\ to add edge ($p_i$,$p_j$)

3.      setEdgeWeights(WS,B,$V_B$,$E_B$) using equation 2 and 3.
4.      normalizedEdgeWeights ($E_B$).
```

Aggregate reciprocal comparisons scheme (ARCS): This outline is based on the postulate that the more entities a block contains, the less likely they are tobe pairs. The weight of an edge $e_{i;\,j}$ is denoted as follows in equation 2:

DOI 10.17148/IJARCCE.2016.5776

$$e_{i;j}.\text{weight}=\sum_{b_k \in B_{i;j}} 1/\|b_k\| \quad \text{……..(2)}$$

Where,
$B_i \subseteq B$ signifies the set of blocks comprising the entity p_i,
$B_{i;j} \subseteq B$ is the set of blocks shared by the entities p_i and p_j (i.e., $B_{i;j}= B_i \cap B_j$)
Common blocks scheme (CBS): A strong sign of the likeness of two entitiesis specified by the number of blocks they have in common; the more blocks they share, the more likely they are to be paired. Therefore, the weight of an edge connecting entities $p_i$and $p_j$is set given by equation 3:

$$e_{i;j}.\text{weight}= |B_{i,j} \quad \text{…….(3)}$$

In this way, weighting scheme of the entities is calculated for Building Blocks of Entities algorithm.

iv] Pruning the Blocking Graph
Introduce a series of pruning schemes that rely on uncertainpruning algorithms that can be applicable to any blocking graph. Edge-centric algorithms exclusive the globally best likenesses by repeating over the edges of a blocking graph in adequate to filter out those that do not fulfill the pruning standard. Thus pruning is done using two ways such as Weight Edge Pruning and Cardinality Edge Pruning.

a) Weight Edge Pruning (WEP): This method involves of the edge-centric algorithm fixed with a global weight threshold with the minimum edge weight as specify in the algorithm 4.

b) Cardinality Edge Pruning (CEP): This method combines with a global cardinality threshold K that states the total number of edges engaged in the pruned graph. The aim is to maintain the K edges with the maximum weight in the algorithm 5.

Algorithm 4:Weight Edge Pruning (WEP)

INPUT: G_B^{in} the blocking graph and w_{min} the global weight pruning criterion.
OUTPUT: G_B^{out} the undirected pruned blocking graph

1. Iterates over all edges using foreach $e_{i,j} \in E_B$
2. To discard every edge with weight lower than w_{min} if $e_{i,j}.\text{weight} < w_{min}$
then $E_B \leftarrow E_B – \{e_{i,j}\}$

Algorithm 5:Cardinality Edge Pruning (CEP)

INPUT: G_B^{in} the blocking graph and K the global cardinality pruning criterion i.e specifies total number of edges retained in pruned graph.
OUTPUT: G_B^{out} the undirected pruned blocking graph.

1. Sorts edges in descending weight such as SortedStack ← {};
2. Add every edge in sorted stack as such SortedStack.push($e_{i,j}$);
3. Remove the edge with $(K+1)^{th}$ top weight using SortedStack.pop();
4. Discard all edges that are not among the top-K weighted ones

if $e_{i,j} \notin$ SortedStack then $E_B \leftarrow E_B – \{e_{i,j}\}$

IV.EXPERIMENTAL RESULTS

Dataset: The standard datasets like DBLP Bibliography containing 1,812 paper author's record has been used which is linked to data mining domain.

Experimental Setup: JDK environment is used for implementation. The experiment is done on Windows with Intel core i5 processor, speed 2.30 GHz and RAM 4 GB.
Existing system did not work on the users input i.e. it does not have facility to work on user precise input, so proposed system has facility to work on the user preciseinput. Thus proposed system is user approachable system.

The fig. 2.F-measure used by R-ER is measured using precision and recall to calculate accuracy on the data set. Sampling of the records is described in the table 3 to measure F-measure. Thus graph presents effective and resourceful measure for retrieval of information.

Table 3 Sampling of the records is used to measure F-measure.

F-measure	R-ER
"Xavier AlamÃ¡n"	0.59
"Andrew Thangaraj"	0.65
"Feng Wang"	0.71
"Tie Li"	0.75
Rui Wang"	0.80
"Jun Zhang"	0.85

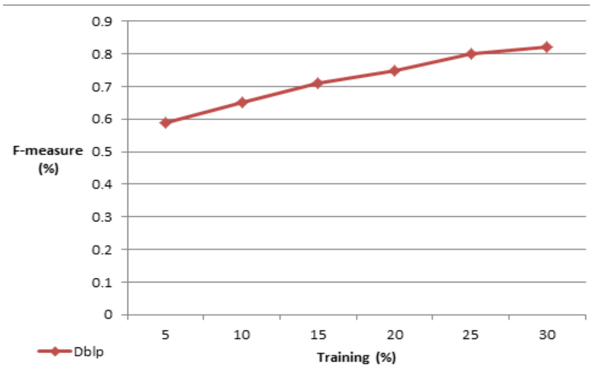


Fig.. 2 Outcome of data size on accuracy

The fig.3 It signifies the effect of unpredictable training data size on the number of generated rules. It also display number of rules is larger than number of training records on data set. Thus, it can be concluded that size of rules would not be large since it grows with training data size.

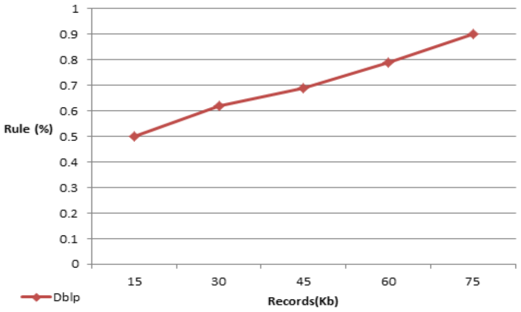


Fig.. 3 Outcome of Data size on rule

The fig.4 It specifies the performance of DISCR algorithm on dblp.Runtime for DISCR algorithm is quadratic the number of records.

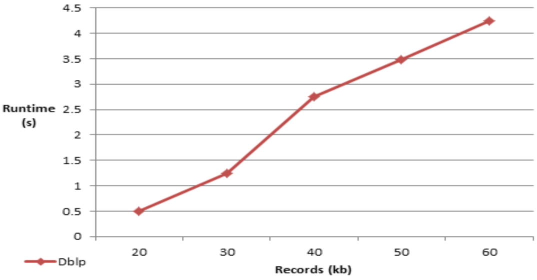


Fig. 4 Scalability of DISCR Algorithm

The fig.5 It specifies the performance of R-ER algorithm on dblp.Runtime for R-ER algorithm is approximately linear to the number of records.

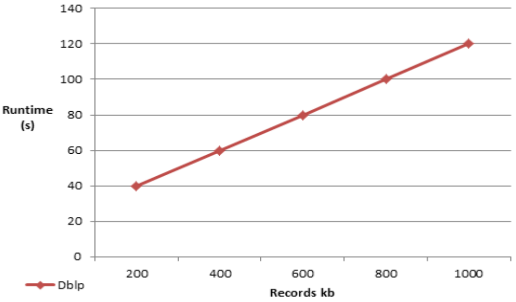


Fig. 5 Scalability of R-ER Algorithm

V. CONCLUSION

Rule Discovery and R-ER algorithm aremeant to match complex matching conditions between records and entities outcomes resolved entity profiles. This entity profiles are used byblocking method to provide exact matching with entities by performing building blocks of entities, weighting scheme and pruning blocks to obtain well classified entity set. Thus, the algorithms achieve good conduct in the sense of efficiency and accuracy for the purpose of recognizing records stating to match the real world entity. The experimental results shows that including blocking method has improved the accuracy on data size which can be seen in F-measure graph and also the time complexity is been reduced for both the algorithm rule discovery and R-ER .Also, blocking approach has increase the efficiency of record likeliness with respect to entities.

ACKNOWLEDGMENT

We are very thankful to all the authors of the papers, those we have referred in our work.

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] George Papadakis, Georgia Koutrika, Themis Palpanas, andWolfgang Nejdl,"Meta-Blocking: Taking Entity Resolution to the Next Level, "IEEE Trans. Knowledge Data Eng.,vol. 26, no. 8,pp. 1946-1960,Aug 2014.

[3] S. Whang, D. Marmaros, and H. Garcia-Molina, "Pay-as-You-Go Entity Resolution," IEEE Trans. Knowledge Data Eng., vol. 25, no. 5,pp. 1111-1124, May 2013.

[4] Qing Wang, Mingyuan Cui and Huizhi Liang, "Semantic-Aware Blocking for Entity Resolution," IEEE Trans. Knowledge Data Eng, 2015.

[5] 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.

[6] 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.

[7] 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.

[8] S. E. Whang and H. Garcia-Molina, "Entity resolution with evolving rules," Proc. VLDB Endowment, vol. 3, no. 1, pp. 1326–1337,2010.

[9] I.Bhattacharya and L.Getoor, "Collective entity resolution in relational data," Proc. VLDB Endowment, vol. 3, no. 1, p. 5, 2010.

[10] 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.

[11] A.Arasu, S.Chaudhuri, and R.Kaushik, "Transformation-based framework for record matching,"in Proc. 24th Int. Conf. Data Eng.,pp. 40–49,2008.

[12] 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.

[13] N.Bansal, A.Blum, and S.Chawla, "Correlation clustering,"Mach. Learn., vol. 56, no. 1–3, pp. 89–113, 2004.

[14] S.Tejada, C. Knoblock, and S. Minton,"Learning object identification rules for information integration," Inf. Syst., vol. 26, no. 8,pp. 607–633, 2001.

BIOGRAPHIES

Ankita Saxena completed her graduation from MET Bhujbal Knowledge City,IOE, Nasik,Maharashtra,India.Presently,she is Post-Graduate student at MET Bhujbal Knowledge City,Institute of Engineering,Nasik,Maharashtra,India.Her research of interest include Data Mining and Big Data.

Prof. Rajana Dahake presently she is working atMET Bhujbal Knowledge City Institute of Engineering, Nasik, Maharashtra, India as a Assistant Professor. She has presented/Publish papers on various thesis of the computer engineering in national/international conferences and journals. Her research of interest include image processing, cloud computing and data mining.

DOI 10.17148/IJARCCE.2016.5776