IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 33-36
www.iosrjournals.org

A Review on Rule Based Method for Entity Resolution

Sreejam M[1], Praveen K Wilson[2]
[1](*Department of Computer Science, College of Engineering Perumon, Kollam, India)*
[2](*Department of Information Technology, College of Engineering Perumon, Kollam, 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 highlights ER as one of the most important problems in data cleaning and arises in many applications like information integration and information retrieval. Traditional ER approaches are ineffective to find records based on pair wise similarity comparisons, which assumes that records referring to the same entity are very similar to each other than otherwise. However for certain circumstances this assumption does not always hold and similarity comparisons do not work very well when such assumption breaks. So to overcome traditional ER drawback a set of rules which could explain the complex matching conditions between records and entities is proposed such as rule discovery algorithm and rule based ER algorithm.*

Keywords: *Data Cleaning ,Entity Resolution, Rule Learning*

I. Introduction

In various application areas, data from multiple sources often 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 getting clean data, document linkage, entity identification or entity resolution (ER) to analyse 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 complexity 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 co-author , 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 between records, assuming that records referring to the same to each other. However, such property may not hold in some situations traditional ER approaches cannot identify records correctly.

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

Example 1. The table2 shows seven authors with name "weiwang" identified by oijs. By viewing to the authors home pages including their publications manly divide the seven authors into three clusters. The records with IDs o11, o12, and o13 refer to the person in UNC, express as e1, the records with IDs o21 and o22 refer to the person in UNSW, express as e2, and the records with IDs o31 and o32 refer to the person in Fudan University, denoted as e3. The function of entity identification is to identify e1, e2and e3using 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: ∀oi, if oi[name] is "weiwang" and oi[coauthors] includes "kum", then oi refers to entity e1;
- R2: ∀ oi, if oi[name] is "weiwang" and oi[coauthors] includes "lin", then oi refers to entity e2;
- R3: ∀oi, if oi [name] is "weiwang" and oi[coauthors] includes "shi", then oi refers to entity e3;
- R4: ∀oi, if oi[name] is "weiwang" and oi [coauthors] includes "zhang" and excludes "shi",then oi refers to entity e1.

Rule based method for Entity Resolution (ER) is being posed when a user want to retrieve data to find 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 pointing 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.

II. Existing System

A. Record linkage: Similarity measures and algorithms

N. Koudas, S. Sarawagi, and D. Srivastava [1] proposed a pairwise ER. They focused on record matching, which involves comparing record pairs and identifying whether they match. They used Similarity metrics ,effective sub-quadratic approximate join algorithms and Efficient clustering algorithms

B. Adaptive duplicate detection using learnable string similarity measures

M. Bilenko and R. J. Mooney[2] proposed a framework for improving duplicate detection using trainable calculations of textual similarity. They used learnable text distance functions for every database field, and show that such measures are capable of adapting to the specific indication of similarity that is appropriate for the field's domain. They presented two learnable text similarity calculations suitable for this task: a new extended variant of learnable string edit distance, and a novel vector-space based measure that exploits support Vector Machine (SVM) for training.

C. Integration of heterogeneous databases without common domains using queries based on textual similarity

W W Cohen [3]rejected the assumption that global domains can be easily constructed, and he assumed instead that the names are given in natural language text. He then proposed a new logic called WHIRL which reasons explicitly about the similarity of local names, as calculated using the vector-space model commonly used in statistical information retrieval. He described an efficient implementation of WHIRL and evaluate it experimentally on data extracted from the World Wide Web.

D. Text joins in an RDBMS for web data integration

L. Gravano, P. G. Ipeirotis, N. Koudas, and D. Srivastava [4]adopted the widely used and well established cosine similarity metric from the information retrieval in order to find potential string matches across web sources. They then used this similarity metric to characterize this key aspect of data integration as a join between relations on textual attributes, where the similarity of matching records exceeds a specified threshold. Computing an exact result to the text join can be expensive. For query processing efficiency proposed a sampling-based join approximation strategy for execution in a standard, unmodified relational database management system (RDBMS), since more and more web sites are powered by RDBMSs with a web-based front end. They implemented the join inside an RDBMS, using SQL queries, for scalability and robustness reasons.

E. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida

M. A. Jaro [5]described three advances. The first is an enhanced method of string comparison for dealing with typographical variations and scanning errors. It improves upon string comparators in computer science. The second is a linear assignment algorithm that can use only 0.002 as much storage as existing

algorithms in operations research, requires at most an additional 0.03 increase in time, and has less of a tendency to make erroneous matching assignments than existing sparse-array algorithms because of how it deals with most arcs. The third is an expectation-maximization algorithm for estimating parameters in latent class, loglinear models of the type arising in record linkage.

F. Transformation-based framework for record matching

G.Arasu, S. Chaudhuri, and R. Kaushik[6] proposed a programmatic framework of record matching that takes such user-defined string transformations as input. To the best of our knowledge, this is the first proposal for such a framework. This transformational framework, while expressive, poses significant computational challenges which we address. They empirically evaluated their techniques over real data.

G. Example driven design of efficient record matching queries

S. Chaudhuri, B. C. Chen, V. Ganti, and R. Kaushik[7] used the availability of positive and negative examples to search through this space and suggest an initial record matching query. Those queries can be subsequently modified by the programmer as needed. Their approach produces are (i) efficient: these queries can be run on large datasets by leveraging operations which are well-supported by RDBMSs, and (ii) explainable: they are easy to understand so that they may be updated by the programmer with relative ease. They demonstrated the effectiveness of approach on several real-world datasets.

H. Learning string transformations from examples

A. Arasu, S. Chaudhuri, and R. Kaushik [8]formulated an optimization problem where they required to learn a concise set of transformations that explain most of the differences, and they proposed a greedy approximation algorithm for this NP-hard problem.

I. Disambiguating web appearances of people in a social network

R. Bekkerman and A. McCallum[9] proposed two models,(a) Link Structure Model (LS) : Builds a core of interconnected pages Of different people! Add proximate pages to the core.(b) Distributional Clustering Model (DC) Simultaneously cluster pages and their words Double clustering is usually more accurate Pick cluster with most interconnected pages.

J. Learning object identification rules for information integration

S. Tejada, C. Knoblock, and S. Minton[10] proposed a technique as a genetic programming technique which contains the three major operations. Those operations are selection, crossover and mutation. Execution of operations applies the de-duplication function. After removing the duplicate records apply the suggested function.

III. Proposed System

In our system, we use a new algorithm using decision tree concept. Rule Discovery (DISCR) algorithm is defined analysing the syntax and semantics of the Rule Based Entity, which includes few primary requirements .The entity for each record is determined by Rule based ER(R-ER) which is done by scanning the records one by one this is done by constructing trees for each candidates and then discovering new rules.

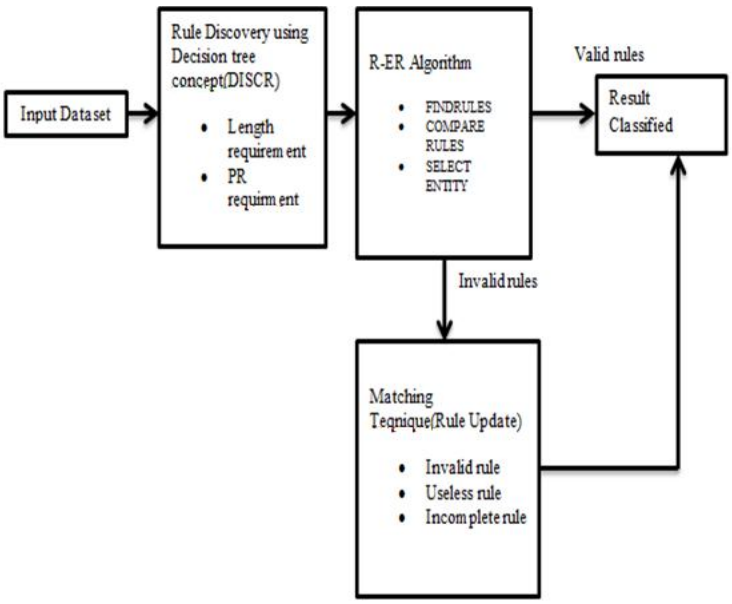

Fig 3.1 : System Flow

IV. Conclusion

In this paper, we show how to solve the Entity Resolution problem in data mining. There are two main algorithms developed, which are Rule Discovery and Rule Based Entity Resolution. We enhance the system by an efficient way to match complex records and entities which comprises of new class of rules based on decision tree method. An efficient Rule Discovery (DISCR) algorithm is defined analysing the syntax and semantics of the Rule Based Entity, which includes few primary requirements .The entity for each record is determined by Rule based ER(R-ER) which is done by scanning the records one by one this is done by constructing trees for each candidates and then discovering new rules. R-ER based on decision tree algorithm achieves good conduct on efficiency, accuracy and also reduces space consumption.

References

[1] N. Koudas, S. Sarawagi, and D. Srivastava, "*Record linkage: Similarity measures and algorithms*," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2006, pp. 802–803.

[2] M. Bilenko and R. J. Mooney, "*Adaptive duplicate detection using learnable string similarity measures*," in Proc. ACM SIGKDD Int. Conf. Knowl . Discovery Data Mining, 2003, pp. 39–48.

[3] W. W. Cohen, "*Integration of heterogeneous databases without common domains using queries based on textual similarity*," ACM SIGMOD Rec., vol. 27, no. 2, pp. 201–212, 1998.

[4] L. Gravano, P. G. Ipeirotis, N. Koudas, and D. Srivastava, "*Text joins in an RDBMS for web data integration*," in Proc. 12th Int. Conf. World Wide Web, 2003, pp. 90–101.

[5] M. A. Jaro, "*Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida*," J. Amer. Statist. Assoc., vol. 84, no. 406, pp. 414–420, 1989.

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

[7] S. Chaudhuri, B. C. Chen, V. Ganti, and R. Kaushik, "*Exampledriven design of efficient record matching queries*," in Proc. 33rd Int. Conf. Very Large Databases, 2007, pp. 327–338.

[8] A. Arasu, S. Chaudhuri, and R. Kaushik, "*Learning string transformations from examples*," Proc. VLDB Endowment, vol. 2, no. 1, pp. 514–525, 2009.

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