**SYSTEM ANALYSIS**

**Existing System**

* These rules can be discovered from existing high quality data such as master data or manually identified data. Inspired by the swoosh method, each cluster is then merged into a composite record via a merge function. Finally a traditional ER method, denoted by T-ER, can be applied to identify the new data set. Moreover, in order to identify more records, the current ER result can be used as the training data to discover new ER-rules. The training data can also be obtained by using techniques, such as relevant feedback, crowd sourcing and knowledge extraction from the web. Therefore, with the accumulated information, ER-rules for more entities can be discovered.
* **Invalid rule.** A rule r is invalid if there exist records that match LHS(r) but do not refer to RHS(r) . Invalid rules might be discovered when the information of entities is not comprehensive. For example, suppose the training data set involves the records. The rule r: (name ¼“wei wang”)^(coa 2“zhang”)) e1 can be generated. For o31, it matches LHS(r) but does not refer to e1. Therefore, r is an invalid rule.
* **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. Both the incomprehensive information of entities and continuous changes of entity features would cause a rule set become incomplete. To solve these problems, we develop some methods to identify candidate invalid rules and candidate useless rules and discover new effective ER-rules.

**PROPOSED SYSTEM**

1. The syntax and semantics of the rules for ER are designed, and the independence, consistency, completeness and validity of the rules are defined and analyzed.

2. An efficient rule discovery algorithm based on training data is proposed and analyzed.

3. An efficient rule-based algorithm for solving entity resolution problem is proposed and analyzed.

4. A rule maintaining method is proposed when entity information is changed.

5. Experiments are performed on real data to verify the effectiveness and efficiency of the proposed algorithms.

**RULES FOR ENTITY RESOLUTION**

A rule system for entity resolution, called ER rule, is defined. We can see that each rule consists of two clauses.

1. The If clause includes constraints on attributes of records, such as “including zhang in coauthors”.

2. The real world entity referred by the records that satisfy the first clause of the rule, such as “refers to entity e1”. Thus, we use A) B to express the rules “8o, If o satisfies A Then o refers to B” for ER. We denote the left-hand side and the right-hand side of a rule r as LHS(r) and RHS(r) respectively.

**RULE-BASED ENTITY RESOLUTION**

The algorithm of entity resolution by leveraging ER-rules. We first define the rule-based ER problem. Next we develop an online algorithm for rule based ER problem. Finally, we describe how to incorporate this algorithm into a generalized ER framework. (Rule-based ER). Rule-based ER takes U and RE as input, and outputs U. U is a data set, RE is an ER-rule set of entity set E ¼ fe1; . . . ; emg, U ¼ fU1; . . . ; Umg is a partition of records where each group Uj(1 \_ j \_ m) is a subset of U which are determined to refer to the entity ej and [1\_j\_mUj is a subset of U. Our rule-based ER algorithm R-ER scans records one by one and determines the entity for each record. The determination process can be divided into the following steps. First, we find all the rules satisfied by o (FINDRULES). Second, for each entity e to which o might refer, we compute the confidence that o refers to e according to the rules of e that are satisfied by o (COMPCONF). Third, we select the entity e with the largest confidence to which o might refer, and if this confidence is larger than a confidence threshold, it is determined that o refers to e (SELENTITY). These procedures are described as follows.

**PAIR WISE ER**

Most works on ER focus on record matching, which involves comparing record pairs and identifying whether they match. A major part of work on record matching focuses on similarity functions. To capture string variations, proposed a transformation-based framework for record matching. Some machine-learning based approaches can identify matching strings which are syntactically far apart. Similarity based on record relationships are also proposed to solve the people identification problem.