УДК 004.6

ENTITY RESOLUTION APPROACHES FOR DATA QUALITY

Chenchen Sun, Derong Shen
College of Information Science and Engineering
Northeastern University, P.R. China
E-mail: dustinchenchen_sun@163.com

Abstract. *Entity resolution is a key aspect of data quality, identifying which records correspond to the same real world entity in data sources. Entity resolution is a hot topic in both research communities and industries. We introduce three approaches to solve different aspects of entity resolution. The first approach learns entity resolution classifiers with genetic algorithm and active learning. The second approach proposes a solution for joint entity resolution. The third approach makes match decision for unsupervised entity resolution by graph clustering. All the three approaches are effective in entity resolution tasks.*

Introduction

In the big data era, one important trait of data is variety. One real world entity may be described by multiple records in data sources. All redundant records are not necessarily the same due to expressive errors and different schemata of data sources. In order to eliminate redundancy and promote data quality, *Entity resolution* (*ER*) indentifies records corresponding to the same entity. There are three challenges in ER tasks: (1) For supervised ER, how to train effective ER classifiers with limited manually labeled data; (2) For related data such as citation data and movie data, how to utilize relations between different records to jointly resolve records; (3) For unsupervised ER, how to effectively make match decision. We focus on the three challenges and introduce three solutions.

GALER: a learning based entity resolution approach

In order to generate effective ER classifiers with less manually labeled data, we propose a novel supervised entity resolution (ER) approach *GALER* with genetic algorithm (GA) and active learning (AL) [1]. Genetic algorithm is able to search over large search spaces and find out near-optimal answers. GALER uses genetic algorithm to learn effective classifiers by choosing proper attributes, effective similarity functions, appropriate thresholds and logical aggregation functions. The approach initializes a randomly generated population of simple ER classifiers. The population iteratively evolves following genetic algorithm. On each iteration, selected classifiers are applied to genetic operations such as reproduction, crossover and mutation with certain probabilities. A set of specialized crossover operators are invented. Each crossover operator is responsible for a special aspect of the ER classifier. In order to reduce manually labeled data, active learning is used to find highly informative pairs. Each time a user is asked to only label a few highly informative pairs. With less labeled training data, GALER is still able to learn effective ER classifiers without accuracy loss.

GB-JER: a joint entity resolution approach

The graph-based joint entity resolution model (GB-JER) [2] iteratively exploits a gradually converged entity record related graph, fully utilizing both direct and indirect relationships among records, to jointly resolve multiple classes of related records. It consists of three core modules (joint match, joint merge and similarity propagation). GB-JER includes three iterative steps: (1) The first pair from candidate queue Q is sent to the joint match. If the pair matches, go to next step; otherwise, repeat the same procedure. The joint match hires a hybrid similarity, combining an attribute-based similarity (ABS) and a structure-based similarity (SBS), to measure the pair's similarity. We propose a schema path based similarity computation algorithm to compute SBS. (2) The joint merge merges the matched pair and contracts the neighborhood. (3)

The contraction change the neighborhood's topology and increases SBSes of some pairs, triggering re-computation. Finally, the graph converges and all the data are resolved. A toy example of joint ER is shown in Fig. 1.

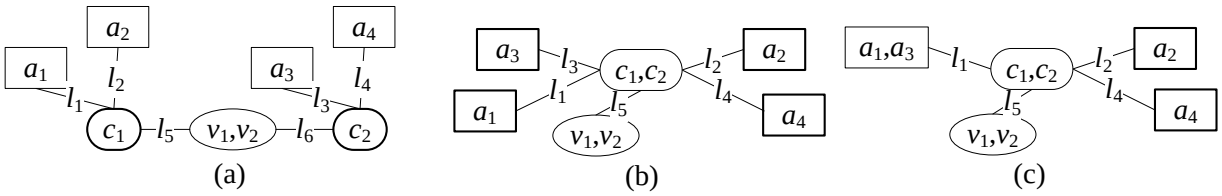

Fig.1 – A toy example of joint entity resolution

ERC: an entity resolution oriented clustering algorithm

We propose a graph-clustering algorithm for ER, named ERC. Build a weighted similarity graph with data objects and their pairwise similarities. During clustering, the similarity between a cluster and a vertex is dynamically computed via random walk with restarts on the similarity graph. The basic logic behind clustering is that a cluster absorbs the nearest neighbor vertex iteratively. A data objects ordering method is proposed to optimize clustering order, promoting clustering accuracy. An improved computation method of random walk's stationary probability distribution is proposed to reduce cost of the clustering algorithm. The whole process is exemplified in Fig.2.

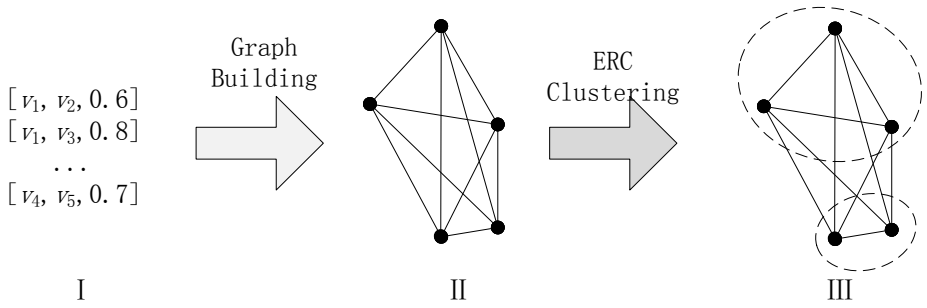

Fig.2 – Workflow of ERC

Conclusion

We introduce three approaches, each of which focuses on a specific aspect of ER. GALER uses genetic algorithm to learn effective ER classifier and reduces overload of manually labeled data via active learning. GB-JER consumes relations between records to resolve multiple records jointly. ERC is a graph-clustering algorithm that makes match decision for unsupervised ER.

References

[1] Sun C, Shen D, et al. A Genetic Algorithm Based Entity Resolution Approach with Active Learning. Accepted by Frontier of Computer Science. 2015.

[2] Sun C, Shen D, et al. GB-JER: A Graph-Based Model for Joint Entity Resolution. Database Systems for Advanced Applications. Springer International Publishing, 2015: 458-473.

[3] Sun C, Shen D, et al. ERC: An Entity Resolution Oriented Clustering Algorithm. Submission to Journal of Software (China). 2015.