Actively Learning Blocking Schemes For Entity Resolution

Introduction

Entity resolution refers to the process of identifying records which represent the same real-world entity from one or more datasets.
Blocking is an important part of entity resolution. It aims to improve the time efficiency of entity resolution by grouping potentially matched records into the same block.
Limitations of existing approaches:

1. Supervised blocking scheme learning approaches require a large number of labels, but it is an expensive task to obtain labels for entity resolution;
2. Existing unsupervised blocking scheme learning approaches, generate training sets based on the similarity of record pairs, instead of their true labels, thus the blocking quality can not be guaranteed.

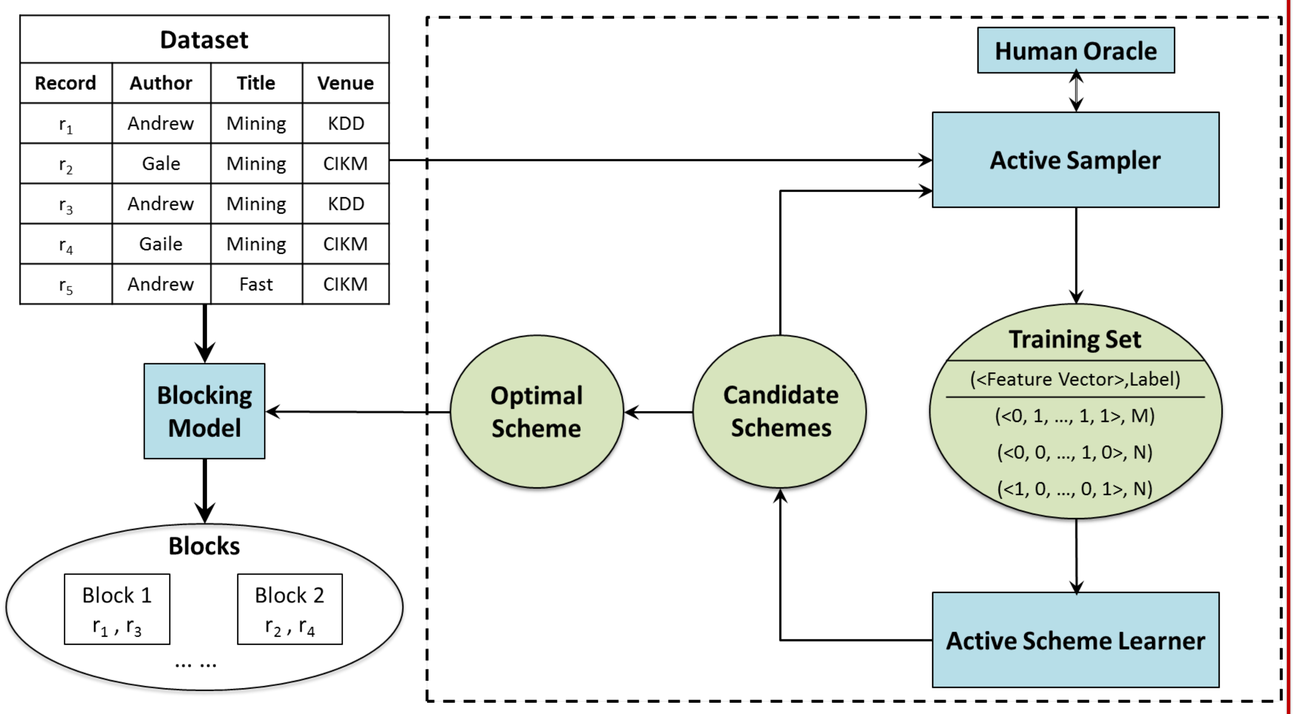
Problem Statement

Given a human oracle ζ, and an error rate $\varepsilon \in [0, 1]$, the *active blocking problem* is to learn a blocking scheme s for a set of blocks B_s within the budget label cost of $budget(\zeta)$:

$$\text{Minimize } |fp(B_s)|$$
$$\text{Subject to } \frac{|fn(B_s)|}{|tp(B_s)|} \leq \varepsilon$$
$$\text{And } |T| \leq budget(\zeta)$$

where, fp: false positive; fn: false negative; tp: true positive

Learning Approach Overview



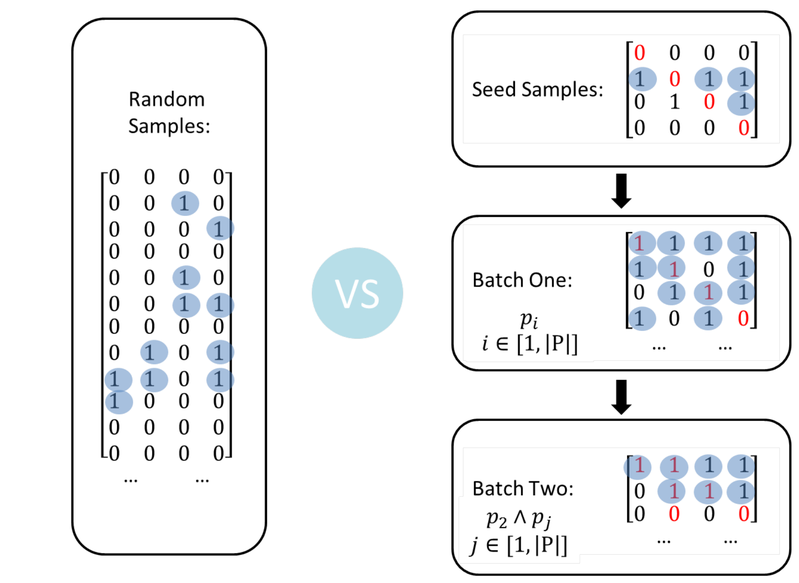
Blocking Process — Our approach to learn a blocking scheme for blocking model

Active Scheme Learning Framework

We develop two complementary and integrated strategies to adaptively learn the blocking scheme.

Active Sampling

To deal with the active blocking problem, we need both match and non-match samples for training. However, one of the well-known challenges in entity resolution is the *class imbalance problem*. That is, if samples are selected randomly, there are usually much more non-matches than matches. We convert the class imbalance problem into the *balanced sampling problem* based on the observation that: the more similar two records are, the higher probability they can be a match.

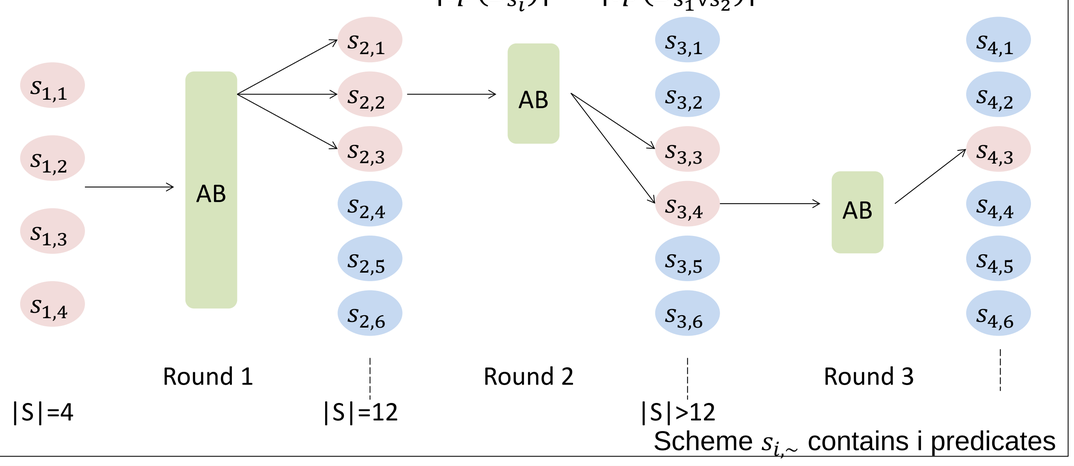


A training set of random sampling — A training set of active sampling

Active Branching

An active branching (AB avoids enumerating all possible blocking schemes and reduce the number of candidate blocking schemes by deciding whether conjunction or disjunction of two candidate schemes will be used in terms of two lemmas.

$$|fp(B_{s_i})| \geq |fp(B_{s_1 \wedge s_2})| \text{ and } \frac{|fn(B_{s_i})|}{|tp(B_{s_i})|} \geq \frac{|fn(B_{s_1 \vee s_2})|}{|tp(B_{s_1 \vee s_2})|}, \; where\ i = 1, 2$$



Scheme $s_{i,\sim}$ contains i predicates

Experimental Results

Characteristics of datasets

Dataset	# Attribute	# Records	Class Imbalance Ratio
Cora	4	1,295	1:49
DBLP-Scholar	4	2,616 / 64,263	1:31,440
DBLP-ACM	4	2,616 / 2,294	1:1,117
NCVR	18	267,716 / 278,262	1:2,692

The number of record pairs generated

	TBlo	Fisher	ASL	RSL
Cora	**2,945**	67,290	29,306	17,974
DBLP-Scholar	6,163	1,039,242	**3,328**	**3,328**
DBLP-ACM	25,279	69,037	**3,043**	17,446
NCVR	932,239	7,902,910	**634,121**	**634,121**

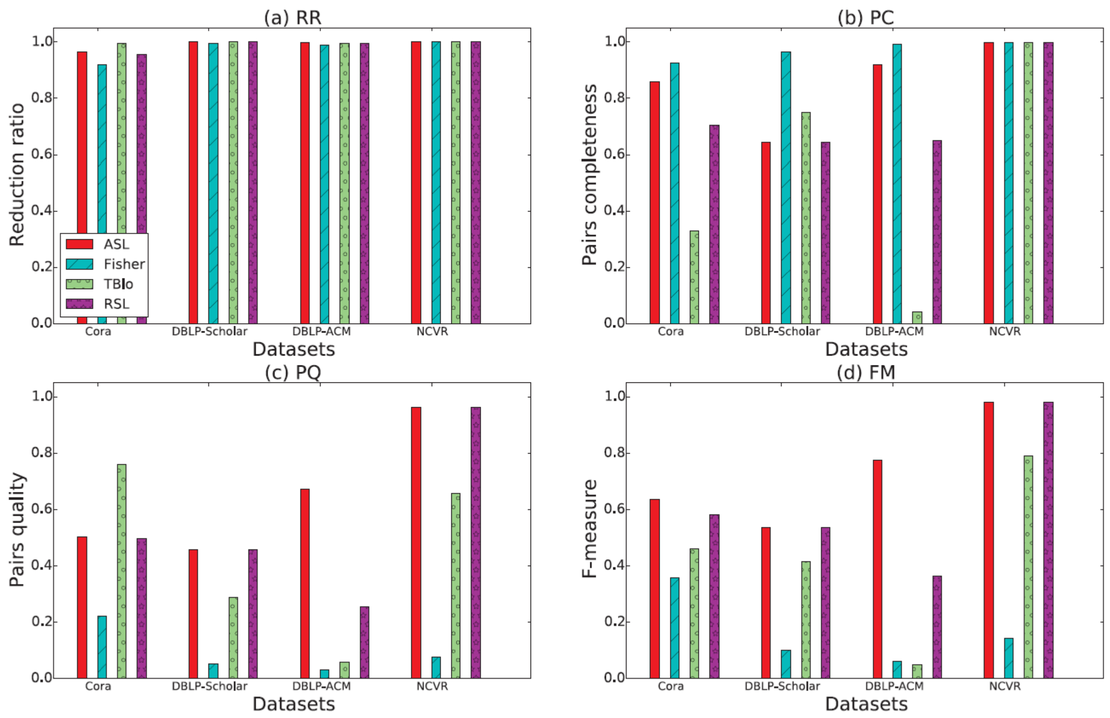
Label cost

	ASL					RSL
Error Rate	0.8	0.6	0.4	0.2	0.1	
Cora	600	400	450	550	500	8,000
DBLP-Scholar	500	350	250	300	250	10,000+
DBLP-ACM	300	200	150	200	300	2,500
NCVR	300	350	250	150	200	10,000+

Baseline Methods:

- Fisher: Mayank Kejriwal & Daniel P. Miranker, ICDM 2013
- Tblo: Ivan P. Fellegi & Alan B. Sunter, 1969
- RSL: uses random sampling technique instead of active sampling

Comparison on blocking quality by different blocking approaches over four real-world datasets using the measures: (a) RR, (b) PC, (c) PQ, and (d) FM



Jingyu Shao and Qing Wang

Research School of Computer Science, College of Engineering and Computer Science,
The Australian National University, Canberra ACT 2000, Australia
{Jingyu.shao ,Qing.wang}@anu.edu.au

Australian National University

ANU College of Engineering & Computer Science