Similarity Spreading based Pay-as-you-go Record Linkage

Chenchen SUN, Derong SHEN

Frontiers of Computer Science, DOI: 10.1007/s11704-021-1130-1

Problems & Ideas

- Problems of pay-as-you-go record linkage (PRL):
  - One key issue is how to globally select pairs by match probability for high quality PRL.
  - Another key issue is PRL in disk based setting.
- Ideas: Similarity spreading based PRL
  - We define similarity spreading based pair similarity estimation (PSE) for global pair ranking; we design a record scheduling method (between RAM and disks) for disk based global pairs selection.

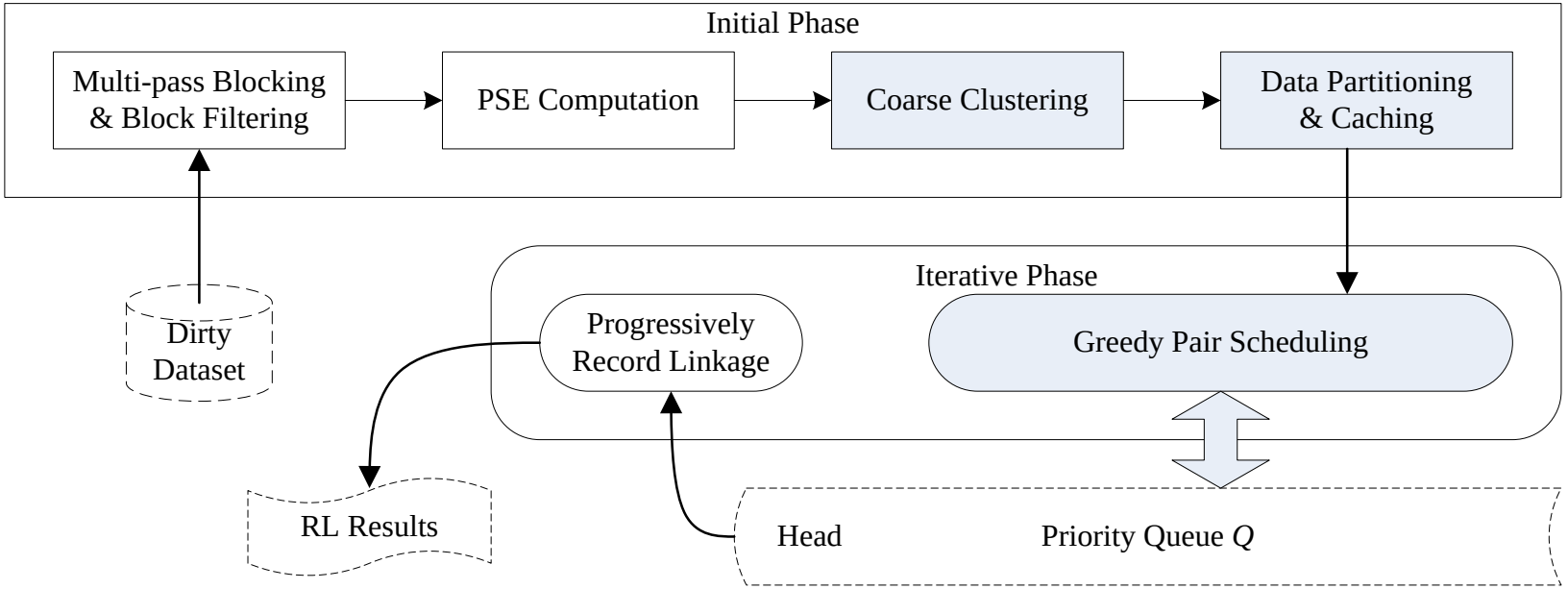

Fig. Disk based pay-as-you-go record linkage workflow

Main Contributions

- Contributions:
  - First, we propose a memory based solution to PRL (M-PRL). As the core of M-PRL, PSE is computed by similarity spreading over interactive bipartite graphs consisted of record pairs and blocks. In this way, we mine overall information intra and inter blocks for global pair similarity estimation.
  - Second, we extend M-PRL to a disk based solution to PRL (D-PRL). We design a cost benefit model to schedule pairs between RAM and disks for PRL. Data are partitioned according to clustering based proximities, and then are greedily exchanged for progressive resolution iteratively.
  - Third, the evaluation results are presented in two tables below. Both M-PRL and D-PRL outperform previous approaches in their respective settings.

Table 1 Memory based Comparison (in Progressive Quality)

	M-PRL	PSNM	PB
Citation data 1	0.71	0.43	0.39
Person data 1	0.62	0.41	0.37

Table 2 Disk based Comparison (in Progressive Quality)

	D-PRL	PSNM	PB
Citation data 2	0.68	0.46	0.34
Person data 2	0.61	0.42	0.33