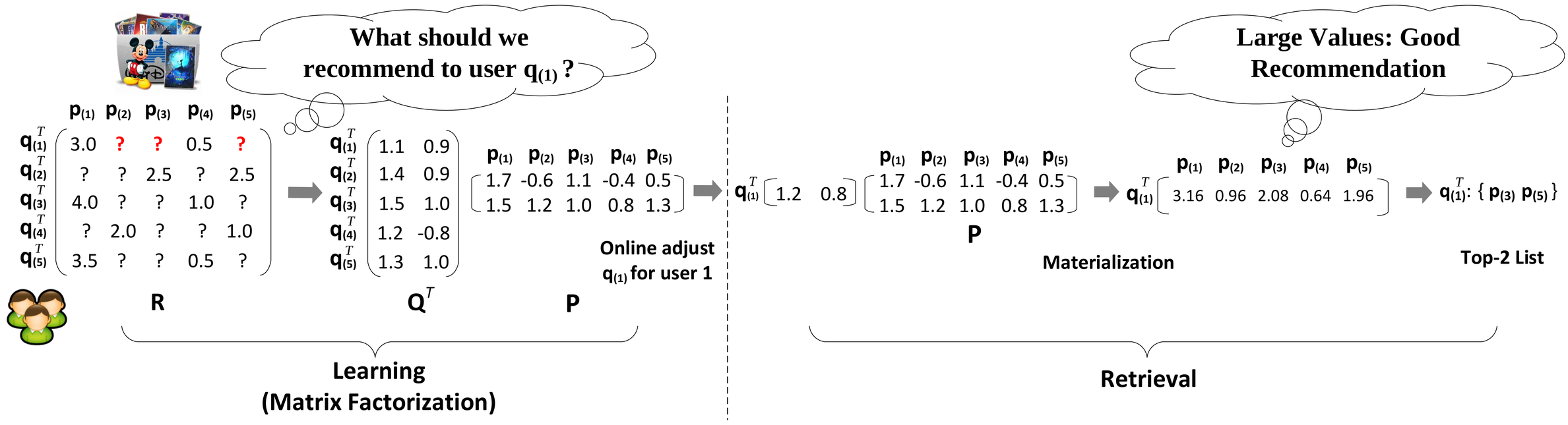
FEXIPRO: Fast and Exact Inner Product Retrieval in Recommender Systems

Hui Li[†], Tsz Nam Chan[§], Man Lung Yiu[§], Nikos Mamoulis[†]

[†]The University of Hong Kong [§]Hong Kong Polytechnic University

[†]{hli2, nikos}@cs.hku.hk [§]{cstnchan, csmlyiu}@comp.polyu.edu.hk

Two-phase Recommender Systems



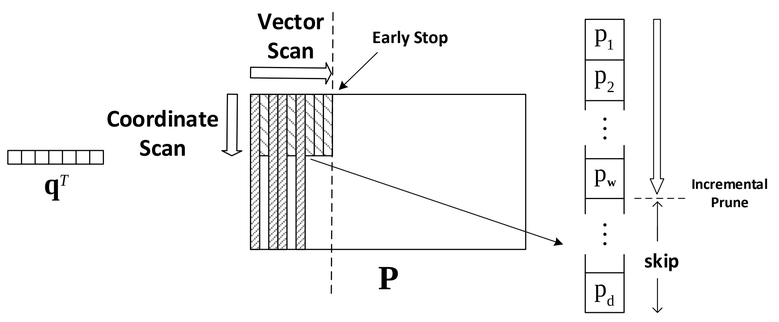
Inner Product Retrieval

- Given a query vector $\mathbf{q}$ from matrix $\mathbf{Q}$, find the k vectors $\mathbf{p}$ from matrix $\mathbf{P}$, for which the $\mathbf{q}^T\mathbf{p}$ values are the largest in $\mathbf{q}^T\mathbf{P}$. Ties are broken arbitrarily.
- Naive method is expensive. Materializing the entire ratings prediction list for all users is practically infeasible.

Sequential Scan

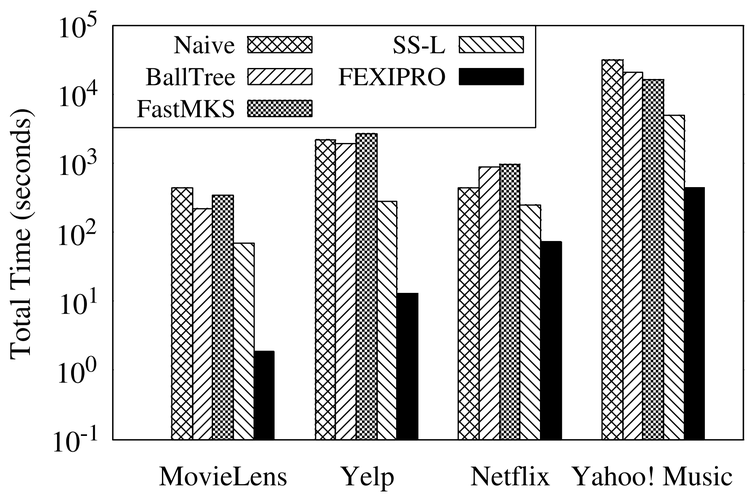
Due to **dimensionality curse**, sequential scan is the most efficient way for **exact** top-k inner product retrieval:

- Vector Scan: $\mathbf{q}^T\mathbf{p} \leq ||\mathbf{q}||\,||\mathbf{p}||$.
- Coordinate Scan: $\mathbf{q}^T\mathbf{p} \leq \mathbf{q}^{\ell T}\mathbf{p}^{\ell} + ||\mathbf{q}^h||\,||\mathbf{p}^h||$.



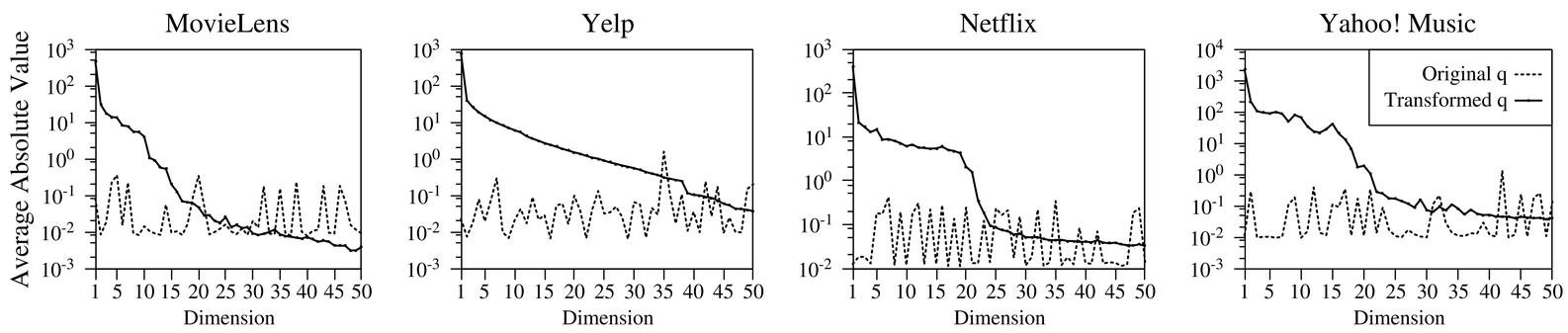
Results for Top-1 Retrieval

FEXIPRO outperforms current exact inner product retrieval approaches typically by an order of magnitude. **Single-threaded** FEXIPRO is even faster than Intel MKL, a high-performance **multithreaded** matrix kernel library.

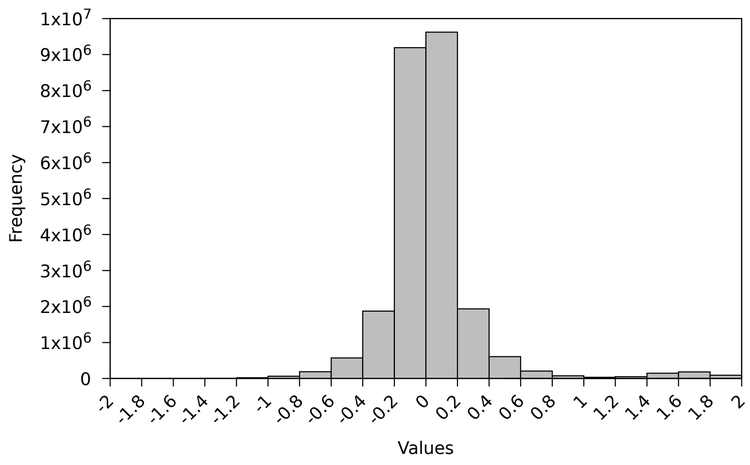


FEXIPRO Framework

- **SVD transformation**: $\mathbf{q}^T\mathbf{P} = \bar{\mathbf{q}}^T\bar{\mathbf{P}}$, where $\bar{\mathbf{q}} = \mathbf{\Sigma_d}\mathbf{U}^T\mathbf{q}$ and $\bar{\mathbf{P}} = \mathbf{V_1}^T$.

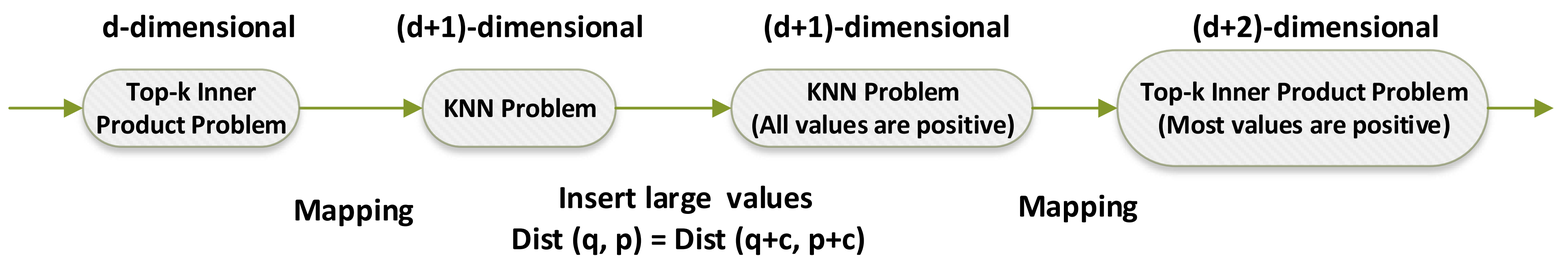


- **Integer-based pruning**: $\mathbf{q}^T\mathbf{p} < \sum_{s=1}^{d}\left(\lfloor q_s \rfloor \cdot \lfloor p_s \rfloor + \left|\lfloor q_s \rfloor\right| + \left|\lfloor p_s \rfloor\right| + 1\right)$. It is a fast-to-compute upper bound. In order for it to be effective in recommender systems, we should scale the vector values before application.



$\mathbf{q}^T$	1.21	0.85	-1.2	0.2	-1.1
$\mathbf{p}^T$	0.25	1.13	0.42	-0.23	0.1
					$\mathbf{q}^T\mathbf{p} = 0.603$
$\hat{\mathbf{q}}^T$	100	70.25	-99.17	16.53	-90.91
$\hat{\mathbf{p}}^T$	22.12	100	37.17	-20.35	0.09
					$\hat{\mathbf{q}}^T\hat{\mathbf{p}} = 5206.28$
$\lfloor\hat{q}\rfloor + \Delta\hat{q}$	100 + 0	70 + 0.25	(-100) + 0.83	16 + 0.53	(-91) + 0.09
$\lfloor\hat{p}\rfloor + \Delta\hat{p}$	22 + 0.12	100 + 0	37 + 0.17	(-21) + 0.65	0 + 0.09
					$\hat{\mathbf{q}}^T\hat{\mathbf{p}} \leq IU(\hat{\mathbf{q}},\hat{\mathbf{p}}) = 5726$

- **Monotonicity reduction** transforms most values to positive ones, hence rendering inner products to monotonically increase as more dimensions are processed. As a result, $\mathbf{q}^{\ell T}\mathbf{p}^{\ell}$ occupies the most of $\mathbf{q}^T\mathbf{p}$ in incremental pruning.



Results for Top-k Retrieval