

Cross Market Modeling for Query-Entity Matching

Manish Gupta Prashant Borole Praful Hebbar Rupesh Mehta Niranjan Nayak
{gmanish, prashant.borole, prafulh, rupeshme, niranjan} @microsoft.com

Problem: Given a query, the query-entity (QE) matching task involves identifying the best matching entity for the query in a particular market.

Examples: Map queries like “Barack Obama’s wife” to “Michelle Obama” and “lead actor of inception” to “Leonardo DiCaprio.” Map "amir khan" to boxer in US and actor in India.

Classifier Solution: Join Entity-URL mapping from an Entity Database with the Query-URL mapping from Click Logs to obtain candidate QE pairs.

Filter candidates using a classifier.

Features

- Click features (P(E|Q), P(Q|E), etc).
- Query-entity features (n-gram match of query with entity).
- Segment distribution for the query and entity.
- Ratio features.
- Query features (query length, peaked-ness of entity ratio distribution for query, etc.)

Challenges

- Features in specific global markets are quite sparse, e.g., clicks in low query volume markets.
- Training data is expensive and hence limited to obtain in multiple markets due to lack of accessible skilled workers and large number of markets.

Main Idea: Leverage cross market data/features for effective query-entity matching in sparse markets.

Related Work

- Multi-task learning [1, 2], and Multi-task learning for web related tasks like web ranking [3, 4].
- Learning models only across related tasks together vs multi-task learning by a combination of sharing training data, features & output data across markets.
- Improving precision vs improving recall (or coverage) for an already highly-precise system.
- Focus on the task of QE matching.

The Proposed Solution

Cross Market Feature Leverage (CMFL)

- For every market M
 - Pick up all features with non-zero information gain.
 - Pick up top k information gain features from other markets.
- Collect all these features and create an augmented feature vector (which now captures signals for same QE pair across markets).

Cross Market Training Data Leverage (CMTDL)

- For QE pair with positive label in training data of market M’, include it into the training data for market M, only if
 - Q is associated with one and only one entity.
 - QE pair is not labeled negative across training data and classifier outputs of all markets.
- For QE pairs with negative label in training data of market M’, include it in training data of market M, only if Q is not associated with the same entity with a positive label in the training data or the classifier output of any market.

Cross Market Output Data Leverage (CMODL)

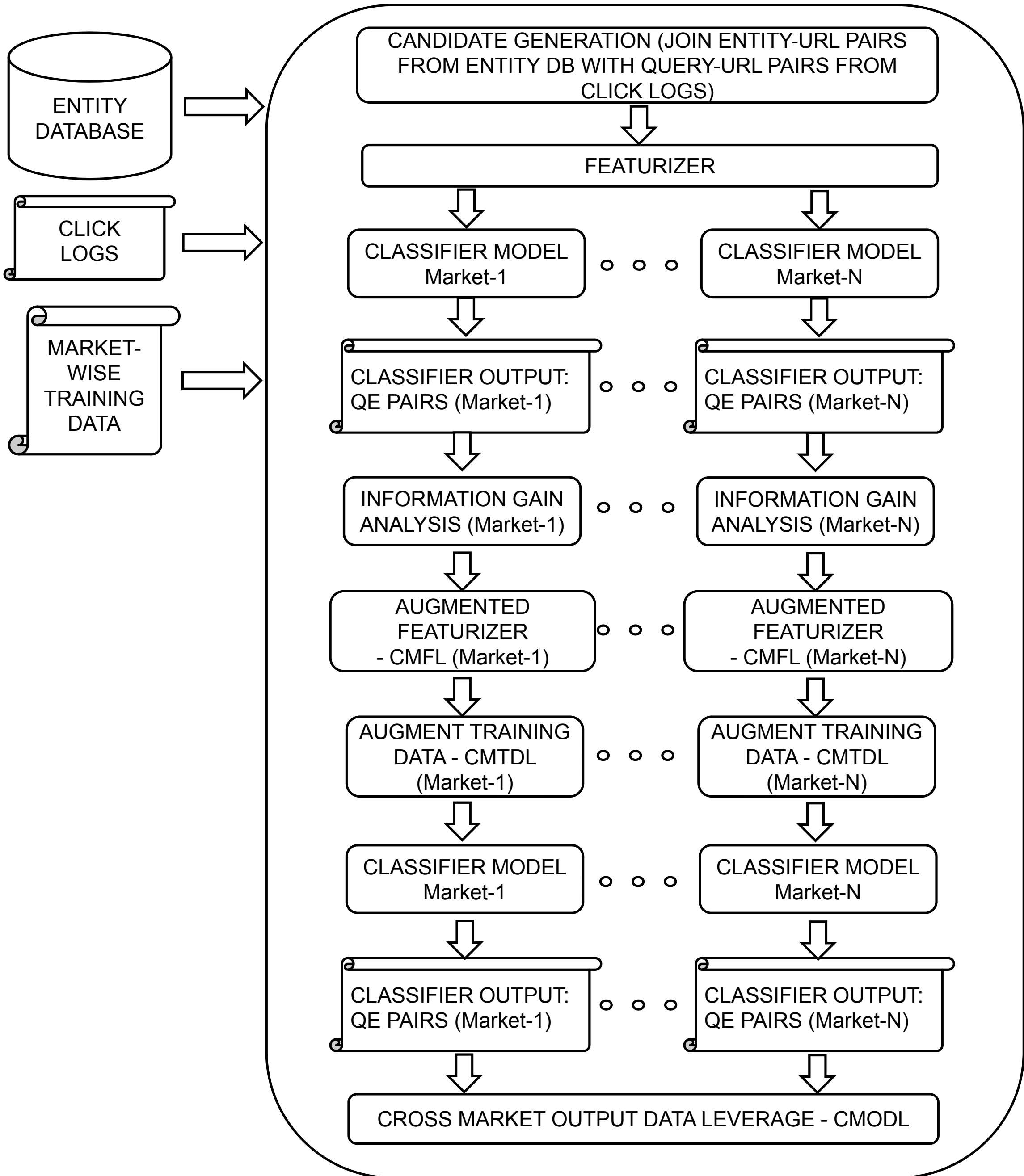
- Borrow QE pairs from market M’ into market M if
 - The QE pair exists in classifier output of at least two markets and in candidate QE pairs for market M, but not in classifier output of market M.
 - No other market output has a QE’ where $E' \neq E$.
 - #clicks for query to entity URLs vs total clicks is >0.05 .
 - Entity E is top clicked entity in market M for query Q.

Features			Class Label
Features from Market 1	...	Features from Market n	0/1
Training Instances from Market 1		Training Instances from Market 1	⋮
⋮	⋮	⋮	⋮
Features from Market 1	...	Features from Market n	⋮
Training Instances from Market n		Training Instances from Market n	0/1

Dataset for Market 1 with Augmented Feature Vector and Augmented Training Data

Summary

- **Query-entity matching** is an interesting problem with applications in Entity Pane on search result pages, etc.
- Classifier based solutions for the QE matching problem face **feature and training data sparsity issues**.
- Our solution helps **effective sharing of data and features across markets** to handle the sparsity issue.
- The solution consisting of **cross market training/output data leverage and cross market feature leverage** shows huge query impression weighted coverage gains with comparable precision.



System Diagram: Cross Market Analysis for QE Matching

Dataset

- Data for 24 markets including en-au, en-ca, en-gb, en-in, es-es, es-mx, es-us, etc.
- Train set of ~250K instances covering ~168K queries with ~40% instances labeled 0.
- Test set of ~62K instances covering ~42K queries with ~40% instances labeled 0.
- Coverage measured across millions of queries using Bing Query Log for Aug 2012-Oct 2013.

Results

- Precision for all approaches is within 0.5% of the baseline.
- “Individual” baseline: No cross market heuristics are used.
- CMFLk: top k information gain based features were used.
- “Aggregate”: Training data across all markets is combined to learn a single global model.

Model	Relative Coverage
Individual	1.00x
Aggregate	0.31x
Individual+CMODL	1.91x
Aggregate+CMODL	1.28x
Individual+CMFL(10/20/50)	1.35x/1.11x/1.15x
Individual+CMFL(10/20/50)+CMODL	2.39x/2.21x/2.2x
Individual+CMTDL +CMFL(10/20/50)	7.79x/7.71x/7.71x
Individual+CMTDL +CMFL(10/20/50)+CMODL	7.91x/7.85x/7.85x

Relative Average Query Impression Coverage for Various Approaches

Market	Other Market Features in the Top 10 by Information Gain
de-at	de-de-ratioAmongEntity, en-gb-ratioAmongEntity
en-au	en-gb-P(Entity Query), en-gb-ratioAmongEntity, en-ca-ratioAmongEntity
	en-ca-P(Entity Query), en-gb-ratioAmongEntity, en-ca-ratioAmongEntity,
en-in	en-gb-P(Entity Query)
es-us	es-es-ratioAmongEntity
fr-be	fr-fr-P(Entity Query), en-gb-ratioAmongEntity
fr-ca	en-ca-Entity-EntityStaticRank, en-ca-ratioAmongEntity
nl-be	en-gb-P(Entity Query), nl-nl-ratioAmongEntity, en-gb-ratioAmongEntity

Other Market Features in Top 10 by Information Gain

Analysis of Results

- Aggregating training data blindly across markets leads to low coverage.
- CMFL and CMTDL contribute significant gains in coverage with comparable precision.
- Borrowing too many features does not help (k=10 is best)
- CMODL improves coverage even after CMFL and CMTDL have been used.
- For many markets, features from other markets turn out to be within top 10.
- Analyzing the information gain of the cross market features, we observed that intuitively related markets contribute signals to each other.