Cross Market Modeling for Query-Entity Matching

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

ABSTRACT

Given a query, the query-entity (QE) matching task involves identifying the best matching entity for the query. When modeling this task as a binary classification problem, two issues arise: (1) features in specific global markets (like $\overline{de}$-at: $\overline{German}$ users in Austria) are quite sparse compared to other markets like en-us, and (2) training data is expensive to obtain in multiple markets and hence limited. Can we leverage some form of cross market data/features for effective query-entity matching in sparse markets? Our solution consists of three main modules: (1) Cross Market Training Data Leverage (CMTDL) (2) Cross Market Feature Leverage (CMFL), and (3) Cross Market Output Data Leverage (CMODL). Each of these parts perform "signal" sharing at different points during the classification process. Using a combination of these strategies, we show significant improvements in query-impression weighted coverage for the query-entity matching task.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and Principles; H.4.0 [**Information Systems Applications**]: General

Keywords

Cross Market Feature Leverage, Cross Market Modeling, Cross Market Output Data Leverage, Cross Market Training Data Leverage, Multi-Task Learning, Query-Entity Matching

1. INTRODUCTION

Given a query, the query-entity (QE) matching task involves identifying the best matching entity for the query. This task is crucial for applications like the entity panes in search engines, requiring dominant entity association for a query to display information to user in more structured way. The task is challenging because (1) it is not trivial to map queries like "barack obama's wife" and "lead actor of inception" to entities, (2) the mapping is market-specific for many queries like "president" and "amir khan" (boxer in US, actor in India). We model this problem as a classification problem where market-specific features are used.

Classifier Solution: Entity-URL mapping from an Entity Database is joined with the Query-URL mapping from Click Logs to obtain candidate QE pairs. A large number of features (400+) are then extracted for each of these query-entity candidates. These features include (1) Click features like P(entity|query), P(query|entity), query percentile, etc.; (2) Query-entity features like match of query with entity name (and alias, description, related entity names); (3) Segment features which capture the segment distribution for the query, and the segment distribution for the entity; (4) Ratio features like ratio of the query clicks captured by this entity, ratio of query clicks that can be attributed to entities within the database, rank of this entity among other entities related to this query; (5) Query features like query length, peakedness of entity ratio distribution for query.

Based on these feature values, classifiers are trained separately for each of the markets and query-entity pairs are obtained separately for each market. Though the candidate generation is market independent, the classifier needs to be learned in a market specific way. Thus, the labeled data and hence the classifier learned for en-ca market is different from en-in or pt-br market. This is essential because the feature values and their importance (or correlation with the relevance class label) could be very different across markets. Also the relevance label for a (query, entity) pair could be different across multiple markets.

Challenges: When modeling this task as a binary classification problem, two issues arise: (1) features in specific global markets are quite sparse, e.g., clicks in low query volume markets, and (2) training data is expensive and hence limited to obtain in multiple markets due to lack of accessible skilled workers and large number of markets. To solve these challenges, can we leverage some form of cross market data/features for effective query-entity matching in sparse markets?

Comparison with Related Work: The proposed problem is closely related to multi-task learning [1, 2]. Recently multi-task learning has been used for web related tasks, especially for web ranking [3, 4]. We deal with the specific setting where the task as well as the data across multiple markets is highly related. (1) Previous work focused only on learning *models* across related tasks together while we focus on multi-task learning by a combination of sharing *training data, features and output data* across markets. (2) Previous work focused on improving precision; our goal is to improve recall (or coverage) for an already highly-precise system. (3) Unlike any previous work, we focus on the task of QE matching.

The Proposed Solution: Our solution consists of three main parts: (1) Cross Market Training Data Leverage (CMTDL) (2) Cross Market Feature Leverage (CMFL), and (3) Cross Market Output Data Leverage (CMODL). Each of these parts perform "signal" sharing at different points during the classification process.

2. OUR APPROACH

Cross Market Feature Leverage (CMFL): CMFL deals with borrowing feature values from other markets to obtain a richer training data for current market, e.g., borrowing click features from de-de market into de-at market. For every market M, we pick up all features with non-zero information gain. We also pick up top k information gain features from other markets. We collect all these features (which now capture signals for the same QE pair across markets) and then create an augmented feature vector for every QE instance for market M.

Cross Market Training Data Leverage (CMTDL): Small # of training data samples per market can result in overfitting. To solve this,

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577277.

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

Figure 1: Dataset for Market 1 using Augmented Feature Vector and Augmented Training Data

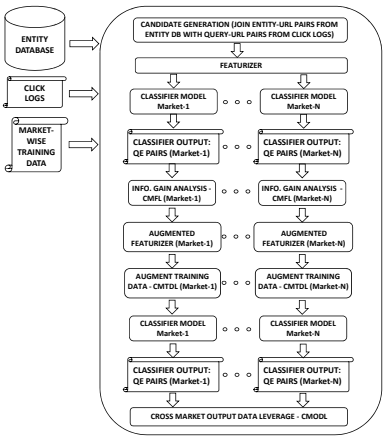

Figure 2: Cross Market Analysis for QE Matching

we can augment training data horizontally across markets by taking the QE pairs from other markets and generating the augmented vector for the QE pair for current market M. But, we do not have the label for these QE pairs (borrowed from training data of other markets) for market M. To avoid extra labeling effort, we use the following heuristics. We include a QE pair with positive label in training data of market M' into the training data for market M, only if the query is associated with one and only one entity and the QE pair is not labeled negative across training data and classifier outputs of all markets. That is, if the query has a single intent globally. For QE pairs with negative label in training data of market M', we include it in training data of market M, only if the query is not associated with the same entity with a positive label in the training data or the classifier output of any market.

The new dataset with augmented feature vector (using CMFL) and augmented training data (using CMTDL) is shown in Figure 1.

Cross Market Output Data Leverage (CMODL): CMODL deals with borrowing QE pairs from output of other markets into the output of current market. Following are the criteria used to borrow QE pairs from market M' into market M. (1) The QE pair exists in classifier output of at least two markets and in candidate QE pairs for market M. (2) No other market output has a QE' where $E \neq E'$. (3) Ratio of clicks for a query that can be attributed to entities versus total clicks for the query is >0.05. (4) Entity E is the top clicked entity in market M for the query Q. The block diagram for cross market QE matching is shown in Figure 2.

3. EXPERIMENTS AND RESULTS

Dataset: We experimented with data for 24 markets including en-au, en-ca, en-gb, en-in, es-es, es-mx, es-us, etc. Total training instances: ∼250K covering ∼168K queries; ∼40% training instances are labeled 0. For measuring precision, we use a test set of ∼62K instances covering ∼42K queries with ∼40% instances labeled 0. Coverage was measured across millions of queries using Bing Query Log for Aug 2012-Oct 2013.

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

Table 1: 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-in	en-ca-P(Entity|Query), en-gb-ratioAmongEntity, en-ca-ratioAmongEntity, 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

Table 2: Other Market Features in Top 10 by Information Gain

Results: Table 1 shows the coverage relative to the baseline. Precision for all approaches remains within 0.5% of the baseline. (1) "Individual" represents the baseline when none of the cross market heuristics are used. (2) CMFLk indicates that top k information gain based features were used. (3) "Aggregate" is the setting where training data across all markets is combined to learn a single global model.

Analysis: 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 much (k=10 provides best results). CMODL improves coverage even after CMFL and CMTDL have been used. Further, we observed that for many markets, features from other markets turn out to be within top 10. Table 2 shows a few of such instances. Analyzing the information gain of the cross market features, we observed that intuitively related markets contribute signals to each other. Some such related markets are as follows. (1) en-au: en-gb (2) en-in: en-ca, en-gb (3) es-us: es-es (4) fr-be: fr-fr (5) fr-ch: fr-fr.

4. CONCLUSION

Query-entity matching is an interesting problem with applications in Entity Pane on search result pages, displaying structured information about most dominant matching entity for user query, etc. Classifier based solutions for multiple markets for the QE matching problem face feature and training data sparsity issues. We proposed a solution which 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.

5. REFERENCES

[1] Theodoros Evgeniou and Massimiliano Pontil. Regularized Multi-task Learning. KDD, pages 109–117, 2004.

[2] A Evgeniou and Massimiliano Pontil. Multi-task Feature Learning. NIPS, volume 19, pages 41–48, 2007.

[3] Jing Bai, Ke Zhou, Guirong Xue, Hongyuan Zha, Gordon Sun, Belle Tseng, Zhaohui Zheng, and Yi Chang. Multi-task Learning for Learning to Rank in Web Search. CIKM, pages 1549–1552, 2009.

[4] Olivier Chapelle, Pannagadatta Shivaswamy, Srinivas Vadrevu, Kilian Weinberger, Ya Zhang, and Belle Tseng. Multi-task Learning for Boosting with Application to Web Search Ranking. KDD, pages 1189–1198, 2010.