1. Motivation

- **Domain search**: A domain is a set of values in a dataset (e.g., a column). We use containment score, \(|Q \cap X|/|Q|\), as the measure of relevance of a domain \(X\) to a query domain \(Q\).
- **Application**: Domain search is often essential to finding joinable datasets. As the landscape of Open Data and Web Data is growing fast, a domain search index can be useful to data scientists.

2. Major Challenges

- **Massive scale**: For example, 263 million domains from attributes in 2015 English Relational WDC Web Table [1].
- **Small memory**: Index must be compact to handle hundreds of millions of domains, and search query must have small memory footprint as it needs to be exchanged over the Web.
- **Open world domains**: A fixed-vocabulary cannot be assumed – unseen values may occur in query/new domain.

3. LSH Ensemble - Containment Search

LSH Ensemble solves the domain search problem with a containment threshold. Given a collection of domains \(\mathcal{D}\), a query domain \(Q\), and a threshold \(\tau^* \in [0, 1]\) on the containment score, find a set of relevant domains from \(\mathcal{D}\) defined as:

\[
\{ X : \text{Containment}(Q, X) \geq \tau^*, X \in \mathcal{D} \}
\]

Containment can be converted to Jaccard using the following equation:

\[
\text{Jaccard}(Q, X) = \frac{|Q \cap X|}{|Q| + |X| - |Q \cap X|} + 1 - \text{Containment}(Q, X)
\]

Given domain sizes in the range \([l, u]\), we use the new Jaccard threshold \(s^* = \frac{2 \tau^*}{\tau^* + 1}\) on a MinHash LSH index [3] to approximate containment search, creating false positive domains. Removing the false positives incurs additional cost.

The query cost of containment search can be characterized as:

\[
T_{\text{Containment}} = T_{\text{Jaccard}} + \Theta(N_{l,a}^C) + \Theta(N_{l,a}^{FP})
\]

where \(N_{l,a}^C\) is the number of correct domains and \(N_{l,a}^{FP}\) is the number of false positive domains.

3.1. Domain Size

- **NVIDIA**: C0112
- **Imperial Oil Ltd**: C1254
- **IBM Canada Ltd**: C5678

3.2. Industry Partners

- **NVIDIA**: Ontario
- **Imperial Oil Ltd**: Alberta
- **IBM Canada Ltd**: Quebec

4. LSH Ensemble - Partitioning

Query cost can be reduced by reducing the number of false positives, which is bounded by \(N_{l,a}^{FP} \leq N_{l,a} \cdot \frac{u - l + 1}{2u}\). We achieve this by domain partitioning.

By executing query in parallel, query cost is determined by the partition with the most false positive domains.

This led us to formulate an optimization problem for partitioning using the upper bound of \(N_{l,a}^{FP}\) on each partition.

\[
\Pi^* = \arg\min_{\Pi} \max_{1 \leq i \leq n} M_i, \quad M_i = N_{l,a} \cdot \frac{u - l + 1}{2u}
\]

This is equivalent to finding a partitioning such that all partitions have the same \(M_i\).

We proved that **equi-depth** partitioning is optimal for domains following power-law distribution.

We also proposed a query-time tuning strategy for MinHash LSH given any containment threshold. See the paper for details.

5. Evaluation


Scalability experiment used the complete 2015 English Relational WDC Web Table [1] (263 million domains), and \(\tau^* = 0.5\).

<table>
<thead>
<tr>
<th></th>
<th>Mean Query (sec)</th>
<th>Precision Before Pruning</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinHash LSH</td>
<td>45.13</td>
<td>0.27</td>
</tr>
<tr>
<td>LSH Ensemble  (8)</td>
<td>7.55</td>
<td>0.48</td>
</tr>
<tr>
<td>LSH Ensemble  (16)</td>
<td>4.26</td>
<td>0.53</td>
</tr>
<tr>
<td>LSH Ensemble  (32)</td>
<td>3.12</td>
<td>0.58</td>
</tr>
</tbody>
</table>

References