Learning to Rank for Quantity Consensus Queries

Somnath Banerjee    Soumen Chakrabarti
Ganesh Ramakrishnan

HP Labs India and IIT Bombay
www.cse.iitb.ac.in/~soumen/doc/QCQ
Quantity queries

- Physical quantity with units or unitless count
- Price, weight, battery life, driving time, mileage
- Frequent, commercially important query class
- For effective quantity search, must support
  - Expressing the target quantity type
  - Extracting typed quantities from text snippets
  - Assembling evidence in favor of numeric answers

microsoft earnings
driving time between paris and nice
battery life of lenovo x300
number of people infected by hiv worldwide
top speed of mclaren f1 car
price canon powershot sx10is
Sources of uncertainty

**Sampling/measurement**
- Height of giraffe
- Driving time from A to B
- Speed of light, value of $\pi$

**Temporal:**
- Number of planets
- Pluto to Sun distance
- Microsoft revenue

**Ambiguity:**
- $1 \text{ ton} = ? \text{ kg}$
- Plutonium half-life

Snippet with incorrect quantity
Detecting consensus is nontrivial

La Giraffe was small (approx. 11 feet tall) because she was still young, a full grown giraffe can reach a height of 18 feet.

Giraffe Photography uses a telescopic mast to elevate an 8 megapixel digital camera to a height of approximately 50 feet.

The record height for a Giraffe unicycle is about 100 ft (30.5m).

Since the Airbus A380 weighs approximately 1,300,000 pounds when fully loaded with passengers ...

The new mega-liner A380 needs the enormous thrust of four times 70,000 pounds in order to take off.

According to Teal, the 319-ton A380 would weigh in at 1,153 pounds per passenger

It also says – unnervingly – that relocated raccoons have been known to return from as far away as 75 miles.

Sixteen deer, 2 foxes, one skunk, and 2 raccoons are sighted during one 35 mile drive.

One study found that raccoons could move over 20 miles from the drop-off point in a short period of time.

- Confounding candidates with correct units
  - (four times) 70,000 pounds
  - 35 mile (drive)
  - telescopic mast . . . 50 feet

- Query token proximity = noisy relevance indicator

- Unit variation: 1.3 million pounds, 319 tons; 100 feet, 30.5 m
Snippet feature vector and scoring

- Snippet = window of tokens centered around quantity of desired unit/type
- Query + snippet $\rightarrow$ feature vector $z_i$
- Standard TFIDF features over different fields
  + Proximity features as used in entity ranking
    - Proximity between query token and candidate quantity = reciprocal of number of tokens between them
    - Max proximity to any query token
    - Proximity to rarest (max IDF) query token
    - IDF-weighted average proximity to all query tokens
- Relevance judgment $y_i \in \{-1, +1\}$
- From training snippets $\{(z_i, y_i)\}$ learn $w$
- Sort by decreasing snippet score $s_i = w^\top z_i$
**QCQ system architecture**

1: Query = Unit, words/phrases, interval width

2–4: Snippet construction
- Get URL using search API, fetch pages
- Annotate quantity tokens, extract snippets
- Filter to ensure candidate quantity and ≥ 1 query tokens

5,6: Training snippet ranking model \( w \)
- Manually label snippets (ir)relevant
- Run Joachim’s \( \text{RANKSVM} \)
$w^T z_i$ vs. $x_i$ scatter plots

- Both axes scaled to $[0, 1]$ for clarity
- Relevant/good snippets $= +$, irrelevant/bad $= \circ$
- Ideal $w \implies$ horizontal line separating $+$ from $\circ$
- No such $w$ for any query in our experiments
- Rectangles densely packed with many $+$, few $\circ$
  - Possibly $> 1$ rectangles for some queries
Consensus rectangles and intervals

- Relevant rectangle/s in sea of irrelevant snippets
- If there is any signal in $w^T z_i$, relevant rectangles should have decent typical/average score
- But there are many low-scoring relevant snippets
- How to detect and rank consensus rectangles?
- Position and shape varies across queries
- Turns out top, bottom boundaries can be ignored
Laplacian consensus (Qin et al.)

- Graph with node \( i \) ⇔ snippet \( i \)
- Edge \( (i, j) \) ⇔ similarity between quantities \( x_i, x_j \)
- Edge weight \( R(i, j) \) inversely related to \( |x_i - x_j| \)
  - Decay: \( R(i, j) = \exp(-s(x_i - x_j)^2) \)
  - Distance: \( R(i, j) = \max\left\{0, 1 - \frac{|x_i - x_j|}{|x_i| + |x_j|}\right\} \)
- Final score of node \( i \) is \( f_i \)

\[
\text{Distortion} = \sum_i (f_i - w^\top z_i)^2
\]
\[
\text{Roughness} = \sum_{(i,j) \in E} R(i, j)(f_i - f_j)^2
\]
\[
\text{Violation} = \sum_{g,b} \max\{0, 1 + f_b - f_g\} \geq \sum_{g,b} \mathbb{1}[f_g \leq f_b]
\]

where \( g, b \) are good, bad snippet indexes

\( \arg\min_f \text{Distortion} + \text{Roughness} + \text{Violation} \)
Wu and Marian (W&M)

- Accumulator $A_x$ for each **distinct** quantity $x$
- Snippet $(z_i, x_i)$ contributes score to $A_{x_i}$
- Snippet score decreases . . .
  - Geometrically with search engine rank of containing page
  - Reciprocally with number of candidate quantities on page
  - Exponentially with number of near-duplicate pages
  - Reciprocally with distance between $x_i$ and query tokens

- Whole-page search engine rank signal inappropriate
- No reinforcement between nearby quantities
- Ad-hoc snippet scoring
## Preliminary bake-off

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>NDCG@1</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web1</td>
<td>0.375</td>
<td>0.338</td>
<td>0.362</td>
<td>0.380</td>
</tr>
<tr>
<td>Web2</td>
<td>0.350</td>
<td>0.413</td>
<td>0.357</td>
<td>0.377</td>
</tr>
<tr>
<td>RankSVM</td>
<td>0.369</td>
<td>0.450</td>
<td>0.412</td>
<td>0.406</td>
</tr>
<tr>
<td>W&amp;M</td>
<td>0.306</td>
<td>0.247</td>
<td>0.303</td>
<td>0.322</td>
</tr>
<tr>
<td>Equality</td>
<td>0.384</td>
<td>0.369</td>
<td>0.353</td>
<td>0.382</td>
</tr>
<tr>
<td>Distance</td>
<td>0.407</td>
<td>0.413</td>
<td>0.401</td>
<td>0.420</td>
</tr>
<tr>
<td>Decay</td>
<td>0.421</td>
<td>0.433</td>
<td>0.422</td>
<td>0.435</td>
</tr>
<tr>
<td>Cosine</td>
<td>0.375</td>
<td>0.438</td>
<td>0.396</td>
<td>0.405</td>
</tr>
</tbody>
</table>

- **Web1, Web2**: Public search engines
  - Rewarded for correct quantity anywhere on whole page
  - Very generous upper bound to accuracy

- **RankSVM**: Fit $w$ from manual per-snippet relevance judgment

- Various choices of $R(i, j)$ in Laplacian

- $\{\text{RankSVM, Laplacian-Decay}\} \succ W&M, \text{others}$
Scanning and scoring intervals

1. **inputs**: snippets $S$, interval width tolerance parameter $r$
2. sort snippets $S$ in increasing $x_i$ order
3. **for** $i = 1, \ldots, n$ **do**
4. \hspace{1em} **for** $j = i, \ldots, n$ **do**
5. \hspace{2em} **if** $x_j < \left(1 + \frac{r}{100}\right) x_i$ **then**
6. \hspace{3em} let $I = [x_i, x_j]$
7. \hspace{1em} merit ← $GetIntervalMerit(S, I)$
8. maintain intervals with top-$k$ merit values
9. **for** surviving intervals $I$ in decreasing merit order **do**
10. present snippets in $I$ in decreasing $w^T z_i$ order

- Key question: how to define $GetIntervalMerit$
- $r > 0$ helps, but system robust to $r$
Interval merit score

Snippet set $S$, quantity interval $I$

Sum: $\sum_{i : x_i \in I} w^T z_i$ — Could have scaling problems across queries

Hinge gain: $\sum_{i : x_i \in I} \sum_{j : x_j \not\in I} \max \{0, w^T z_i - w^T z_j\}$ — Inspired by RankSVM

Diff: $\sum_{i : x_i \in I} \sum_{j : x_j \not\in I} (w^T z_i - w^T z_j)$ — Averaged pairwise moment
Interval merit beats all baselines

- Best for both MAP and NDCG
- Picking $r > 0$ improves accuracy
- $\text{Diff} \succ \text{Hinge} \succ \text{Laplacian-Decay}$
- Laplacian pays for $R(i,j)$ even if $i,j$ bad
- Single width param $s$ in $\exp(-s(x_i - x_j)^2)$ problematic
Ranking intervals directly

- \( I \) contains snippet \( i \) if \( x_i \in I \)
- \( I \succ I' \) if \( I \) contains a larger fraction of good snippets than \( I' \)
- Invent interval features
  - All snippets in \( I \) contain \{some, rarest\} query word?
  - Number of distinct quantities mentioned in snippets contained in \( I \)
- Train \textsc{RankSVM} to order intervals
- Result: Further improvements in MAP and NDCG

\[
\begin{array}{cccccc}
  & 0 & 1 & 2 & 3 & 4 \\
\hline
\text{MAP} & 0.45 & 0.47 & 0.49 & 0.51 & 0.53 \\
\hline
\text{IntervalRank} & \ast & \star & \dagger & \ddagger & \Upsilon \\
\text{Diff} & \diamondsuit & \heartsuit & \spadesuit & \clubsuit & \flat \\
\end{array}
\]
Interval-oriented evaluation

► For snippet-oriented MAP and NDCG evaluation
  ▶ First sorted intervals by decreasing merit score
  ▶ Then reported snippets in interval by decreasing $w^T z_i$
► Suppose we output interval list $I_1, \ldots, I_m$
  ▶ Say $I_j$ contains $n_j$ snippet (quantities), $k_j$ relevant

\[
\text{IntervalPrecision}@j = \frac{k_1 + \cdots + k_j}{n_1 + \cdots + n_j}
\]
\[
\text{IntervalRecall}@j = \frac{k_1 + \cdots + k_j}{\text{numGoodSnippets}}
\]

► Intuitive R-P-F1 tradeoff with interval width

<table>
<thead>
<tr>
<th>Algo, measure</th>
<th>#Intervals $j \rightarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>IntervalRank recall</td>
<td>0.521</td>
</tr>
<tr>
<td>Laplacian-Decay recall</td>
<td>0.510</td>
</tr>
<tr>
<td>RankSVM recall</td>
<td>0.458</td>
</tr>
<tr>
<td>IntervalRank prec</td>
<td>0.443</td>
</tr>
<tr>
<td>Laplacian-Decay prec</td>
<td>0.382</td>
</tr>
<tr>
<td>RankSVM prec</td>
<td>0.330</td>
</tr>
</tbody>
</table>
Summary

- Introduced and formalized QCQs
- Standard snippet and entity ranking inadequate
- Clue from score-vs.-quantity scatter plots
- Cannot score snippet independent of others
- New collective ranking algorithms for QCQs
- Better snippet- and interval-oriented accuracy

www.cse.iitb.ac.in/~soumen/doc/QCQ

- ~160 queries, ~15000 labeled snippets available
- 500M page Web-scale evaluation in progress
- Soon: New search API with quantity support