A Generalized Framework of Exploring Category Information for Question Retrieval in Community Question Answer Archives

Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen

In Proceedings of the 19th international conference on world wide web
(WWW 2010)

Prepared and Presented by Baichuan Li
Outline

- Introduction & Motivation
- Category-Enhanced Question Retrieval Models
- Experiments
- Conclusion
Introduction

- Community Question-Answering (CQA) Services
Question Retrieval

Query

Search for questions: Should I buy Mac or PC?

Existed similar questions and their answers
Motivation

Search

Can you recommend a good restaurant in Shanghai

Sort by: Relevance | Newest | Most Answers

Can you recommend a good restaurant in Shanghai, China?
I'm going to be here. Can you recommend a good restaurant in Shanghai, no matter how much it costs ★ In China - Asked by Qindy - 3 answers - 4 months ago

Where can I find a good restaurant in the UK that has a Dim Sum menu?
...asian food, and I would love to try Shanghai Dumplings. Unfortunately, the restaurant that I'm going to in London [The See Cafe] No longer has a Dim Sum menu. Can anyone recommend a good Asian restaurant that offers. Preferably in or around London. Thank you :) ★ In London - Asked by Kaptain Kimbers =] - 2 answers - 1 year ago

Try Yahoo! Search

Can you recommend a good res

Query Category
CATEGORY-ENHANCED QUESTION RETRIEVAL MODELS
Exploiting Categories in Question Retrieval

- Given a query $q$, a historical question $d$, and the category $\text{cat}(d)$ that contains $d$:

$$RS_{q,d} = (1 - \alpha)N(S_{q,d}) + \alpha N(S_{q,\text{cat}(d)})$$

where $S_{q,d}$ is the local relevance score and $S_{q,\text{cat}(d)}$ is the global relevance score, $N()$ is the normalization function and $\alpha$ is a weighting parameter.

- Words play different roles in computing local and global relevance scores.
Retrieval Models

- Vector Space Model
- Okapi BM25 Model
- Language Model
- Translation Model
- Translation-Based Language Model
Vector Space Model

\[
S_{q,d} = \frac{\sum_{t \in q \cap d} w_{q,t} w_{d,t}}{W_q W_d}, \text{ where }
\]

\[
w_{q,t} = \ln(1 + \frac{N}{f_t}), \quad w_{d,t} = 1 + \ln(t f_{t,d})
\]

\[
W_q = \sqrt{\sum_t w_{q,t}^2}, \quad W_d = \sqrt{\sum_t w_{d,t}^2}
\]

Here \( N \) is the number of questions in the whole collection, \( f_t \) is the number of questions containing the term \( t \), and \( t f_{t,d} \) is the frequency of term \( t \) in \( d \).
Vector Space Model

Global relevance score

\[ S_{q, \text{cat}(d)} = \frac{\sum_{t \in q \cap \text{cat}(d)} w_{q,t} w_{\text{cat}(d),t}}{W_q}, \text{ where} \]

\[ w_{q,t} = \ln(1 + \frac{M}{f_{C_t}}), \ w_{\text{cat}(d),t} = 1 + \frac{1}{\ln(\frac{W_{\text{cat}(d)}}{t_{f_t, \text{cat}(d)}})} \]

Here \( M \) is the total number of leaf categories, \( f_{C_t} \) is the number of categories that contain the term \( t \), \( t_{f_t, \text{cat}(d)} \) is the frequency of \( t \) in the category \( \text{cat}(d) \), \( W_{\text{cat}(d)} \) is the length of \( \text{cat}(d) \) (number of words contained in \( \text{cat}(d) \)), and \( w_{q,t} \) captures the IDF of word \( t \) with regard to categories.

Local relevance score

\[ w_{q,t} = \ln(1 + \frac{N_{\text{cat}(d)}}{f_{t, \text{cat}(d)}}) \]
Okapi BM25 Model

\[ S_{q,d} = \sum_{t \in q \cap d} w_{q,t} w_{d,t}, \text{ where} \]

\[ w_{q,t} = \ln\left( \frac{N - f_t + 0.5}{f_t + 0.5} \right) \frac{(k_3 + 1) t f_{t,q}}{k_3 + t f_{t,q}} \]

\[ w_{d,t} = \frac{(k_1 + 1) t f_{t,d}}{K_d + t f_{t,d}} \]

\[ K_d = k_1 \left( (1 - b) + b \frac{W_d}{W_A} \right) \]

Here \( N \) is the number of questions in the collection; \( f_t \) is the number of questions containing the term \( t \); \( t f_{t,d} \) is the frequency of term \( t \) in \( d \); \( k_1, b, \) and \( k_3 \) are parameters.
Okapi BM25 Model

Global relevance score

\[ S_{q,d} = \sum_{t \in q \cap \text{cat}(d)} w_{q,t} w_{\text{cat}(d),t}, \text{ where} \]

\[ w_{q,t} = \ln \left( \frac{N - f_t + 0.5}{f_t + 0.5} \right) \frac{(k_3 + 1)t_f, q}{k_3 + t_f, q} \]

\[ w_{\text{cat}(d),t} = \frac{(k_1 + 1)t_f, \text{cat}(d)}{K_d + t_f, \text{cat}(d)} \]

\[ K_d = k_1((1 - b) + b \frac{W_d}{W_A}) \]

Local relevance score

\[ w_{q,t} = \ln \left( \frac{N_{\text{cat}(d)} - f_{t, \text{cat}(d)} + 0.5}{f_{t, \text{cat}(d)} + 0.5} \right) \frac{(k_3 + 1)t_f, q}{k_3 + t_f, q} \]

\[ K_d = k_1((1 - b) + b \frac{W_d}{W_{A, \text{cat}(d)}}) \]
Language Model

\[ S_{q, d} = \prod_{t \in q} \left( (1 - \lambda) P_{ml}(t | d) + \lambda P_{ml}(t | Coll) \right), \text{ where} \]

\[ P_{ml}(t | d) = \frac{t f_{t, d}}{\sum_{t' \in d} t f_{t', d}} \]

\[ P_{ml}(t | Coll) = \frac{t f_{t, Coll}}{\sum_{t' \in Coll} t f_{t', Coll}} \]

Here \( P_{ml}(t | d) \) is the maximum likelihood estimate of word \( t \) in \( d \); \( P_{ml}(t | Coll) \) is the maximum likelihood estimate of word \( t \) in the collection \( Coll \); and \( \lambda \) is the smoothing parameter.
Language Model

\[ S_{q,d} = \prod_{t \in q} ((1 - \lambda)P_{ml}(t|d) + \lambda P_{ml}(t|\text{Coll})) \], where

\[ P_{ml}(t|d) = \frac{tf_{t,d}}{\sum_{t' \in d} tf_{t',d}} \]

\[ P_{ml}(t|\text{Coll}) = \frac{tf_{t,\text{Coll}}}{\sum_{t' \in \text{Coll}} tf_{t',\text{Coll}}} \]

**Global relevance score**

\[ d \rightarrow \text{Cat}(d) \]

**Local relevance score**

\[ \text{Coll} \rightarrow \text{Cat}(d) \]
Translation Model

\[
S_{q,d} = \prod_{t \in q} (1 - \lambda) \sum_{w \in d} T(t|w) P_{ml}(w|d) + \lambda P_{ml}(t|\text{Coll})
\]

\(T(t|w)\) denotes the probability that word \(w\) is the translation of word \(t\).

IBM translation models:
Translation Model

\[ S_{q,d} = \prod_{t \in q} ((1 - \lambda) \sum_{w \in d} T(t|w)P_{ml}(w|d) + \lambda P_{ml}(t|\text{Coll})) \]

Global relevance score

\[ d \rightarrow \text{Cat}(d) \]

Local relevance score

\[ \text{Coll} \rightarrow \text{Cat}(d) \]
Translation-Based Language Model

\[ S_{q,d} = \prod_{t \in q} ((1 - \lambda) (\beta \sum_{w \in d} T(t|w)P_{ml}(w|d)) + (1 - \beta)P_{ml}(t|d)) + \lambda P_{ml}(t|\text{Coll}) \]

\( \beta \) controls the translation component’s impact.

Global relevance score

\[ d \to \text{Cat}(d) \]

Local relevance score

\[ \text{Coll} \to \text{Cat}(d) \]
EXPERIMENTS
Data Set

- **Question Repository**

<table>
<thead>
<tr>
<th>Category</th>
<th>Question#</th>
<th>Category</th>
<th>Question#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts &amp; Humanities</td>
<td>114737</td>
<td>Health</td>
<td>183181</td>
</tr>
<tr>
<td>Beauty &amp; Style</td>
<td>49532</td>
<td>Home &amp; Garden</td>
<td>50773</td>
</tr>
<tr>
<td>Business &amp; Finance</td>
<td>154714</td>
<td>Local Businesses</td>
<td>69581</td>
</tr>
<tr>
<td>Cars &amp; Transportation</td>
<td>208363</td>
<td>News &amp; Events</td>
<td>27884</td>
</tr>
<tr>
<td>Computers &amp; Internet</td>
<td>129472</td>
<td>Pets</td>
<td>72265</td>
</tr>
<tr>
<td>Consumer Electronics</td>
<td>126253</td>
<td>Politics &amp; Government</td>
<td>85392</td>
</tr>
<tr>
<td>Dining Out</td>
<td>58980</td>
<td>Pregnancy &amp; Parenting</td>
<td>63228</td>
</tr>
<tr>
<td>Education &amp; Reference</td>
<td>107337</td>
<td>Science &amp; Mathematics</td>
<td>116047</td>
</tr>
<tr>
<td>Entertainment &amp; Music</td>
<td>196100</td>
<td>Social Science</td>
<td>61011</td>
</tr>
<tr>
<td>Environment</td>
<td>28476</td>
<td>Society &amp; Culture</td>
<td>122358</td>
</tr>
<tr>
<td>Family &amp; Relationships</td>
<td>53687</td>
<td>Sports</td>
<td>275893</td>
</tr>
<tr>
<td>Food &amp; Drink</td>
<td>55955</td>
<td>Travel</td>
<td>403926</td>
</tr>
<tr>
<td>Games &amp; Recreation</td>
<td>72634</td>
<td>Yahoo! Products</td>
<td>228368</td>
</tr>
</tbody>
</table>

- **Query Set**
## Results

<table>
<thead>
<tr>
<th></th>
<th>VSM</th>
<th>OptC</th>
<th>QC</th>
<th>VSM+VSM</th>
<th>%chg</th>
<th>Okapi+VSM</th>
<th>%chg</th>
<th>LM+VSM</th>
<th>%chg</th>
<th>TR+VSM</th>
<th>%chg</th>
<th>TRLM+VSM</th>
<th>%chg</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.2407</td>
<td>0.2414</td>
<td>0.2779</td>
<td>0.3711</td>
<td>54.2*</td>
<td>0.3299</td>
<td>37.1*</td>
<td>0.3632</td>
<td>50.9*</td>
<td>0.3629</td>
<td>50.8*</td>
<td>0.3628</td>
<td>50.7*</td>
</tr>
<tr>
<td>MRR</td>
<td>0.4453</td>
<td>0.4534</td>
<td>0.4752</td>
<td>0.5637</td>
<td>26.6*</td>
<td>0.5314</td>
<td>19.3*</td>
<td>0.5596</td>
<td>25.7*</td>
<td>0.5569</td>
<td>25.1*</td>
<td>0.5585</td>
<td>25.4*</td>
</tr>
<tr>
<td>R-Prec</td>
<td>0.2311</td>
<td>0.2298</td>
<td>0.2568</td>
<td>0.3419</td>
<td>48.0*</td>
<td>0.3094</td>
<td>33.9*</td>
<td>0.3366</td>
<td>45.7*</td>
<td>0.3346</td>
<td>44.8*</td>
<td>0.3357</td>
<td>45.3*</td>
</tr>
<tr>
<td>P@5</td>
<td>0.2222</td>
<td>0.2289</td>
<td>0.2436</td>
<td>0.2789</td>
<td>25.5*</td>
<td>0.2559</td>
<td>15.2*</td>
<td>0.2746</td>
<td>23.6*</td>
<td>0.2746</td>
<td>23.6*</td>
<td>0.2753</td>
<td>23.9*</td>
</tr>
</tbody>
</table>

Table 1: VSM vs. CE with VSM for computing local relevance (%chg denotes the performance improvement in percent of each model in CE; * indicates a statistically significant improvement over the baseline using the t-test, p-value < 0.05)

<table>
<thead>
<tr>
<th></th>
<th>Okapi</th>
<th>OptC</th>
<th>QC</th>
<th>VSM+Okapi</th>
<th>%chg</th>
<th>Okapi+Okapi</th>
<th>%chg</th>
<th>LM+Okapi</th>
<th>%chg</th>
<th>TR+Okapi</th>
<th>%chg</th>
<th>TRLM+Okapi</th>
<th>%chg</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.3401</td>
<td>0.2862</td>
<td>0.3622</td>
<td>0.4007</td>
<td>17.8*</td>
<td>0.3977</td>
<td>16.9*</td>
<td>0.4138</td>
<td>21.7*</td>
<td>0.4082</td>
<td>20.0*</td>
<td>0.4132</td>
<td>21.5*</td>
</tr>
<tr>
<td>MRR</td>
<td>0.5406</td>
<td>0.4887</td>
<td>0.5713</td>
<td>0.6131</td>
<td>13.4*</td>
<td>0.5884</td>
<td>8.8</td>
<td>0.6214</td>
<td>15.0*</td>
<td>0.6172</td>
<td>14.2*</td>
<td>0.6215</td>
<td>15.0*</td>
</tr>
<tr>
<td>R-Prec</td>
<td>0.3178</td>
<td>0.2625</td>
<td>0.3345</td>
<td>0.3648</td>
<td>14.8*</td>
<td>0.3613</td>
<td>13.7*</td>
<td>0.3758</td>
<td>18.3*</td>
<td>0.3677</td>
<td>15.7*</td>
<td>0.3762</td>
<td>18.4*</td>
</tr>
<tr>
<td>P@5</td>
<td>0.2857</td>
<td>0.2824</td>
<td>0.2998</td>
<td>0.3140</td>
<td>9.9*</td>
<td>0.3176</td>
<td>11.2*</td>
<td>0.3161</td>
<td>10.6*</td>
<td>0.3111</td>
<td>8.8</td>
<td>0.3147</td>
<td>10.2*</td>
</tr>
</tbody>
</table>

Table 2: Okapi vs. CE with Okapi for computing local relevance (%chg denotes the performance improvement in percent of each model in CE; * indicates a statistically significant improvement over the baseline using the t-test, p-value < 0.05)
Conclusion

• Exploiting category information associated with questions for improving question retrieval
• Conducting experiments with large scale CQA data
• Improvements
  ◦ Considering answers
  ◦ Utilizing hierarchical category structures
  ◦ …