Improving Context-Aware Query Classification via Adaptive Self-Training

Minmin Chen¹, Jian-Tao Sun², Xiaochuan Ni², Yixin Chen¹

1. Washington University, Saint Louis, MO
2. Microsoft Research Asia
Background

• Query classification: classify query to predefined categories
  – E.g., query: “cikm 2011” → category: “Computer Science Conferences”

• Challenges
  – Lack of training data (human labeling is expensive & time consuming)
  – Sparseness of query features (e.g., query: “up” → category: “movie”)
    • Query enrichment (Shen et. al 2006, Broder et.al 2007)
Context-aware query classification

- Context is users action within one session
  - E.g., submitting query, viewing result pages, click on URLs

- Search session: a series of interactions by the user toward addressing a single information need (Jansen, 2007)

  query = “giant” → category = “Sport”? (San Francisco baseball team)
  category = “Product”? (Giant bicycle)

Session A: “Major League Baseball” | “giant”
Session B: “bike review” | “mountain bikes” | “giant”
Model search context using Conditional Random Fields (CRF)

- \( x = \langle x_1, x_2, \cdots, x_T \rangle \), observations of query session (queries, clicked URLs, etc.)
- \( y = \langle y_1, y_2, \cdots, y_T \rangle \), category sequence

- Maximum likelihood training

\[
\max_w \sum_{n=1}^{N} \log p(y^{(n)}|x^{(n)}) , \quad \mathcal{D} = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N}
\]

H. Cao, et. al. SIGIR, 2009
Our work: utilize unlabeled query sessions for context-aware query classification

- Combine self-training with Conditional Random Fields

Start with a small set of labeled query sessions

Train a conditional random field model

How to train accurate model with noisy data included?

Adjust the hyper-parameters for the model

How to adapt model to auto-labeled training data?

How to measure confidence of unlabeled query sessions?

Apply the trained model on the unlabeled set

Extract confident data as labeling data
Adaptive self-training CRF (ASCRF)

1. Start with a small set of labeled query sessions
2. Train a conditional random field model
3. Adjust the hyper-parameters for the model
4. Control margin to adapt models
5. Max-Margin CRF
6. Study different confidence measurements
7. Apply the trained model on the unlabeled set
8. Extract confident data as labeling data
SoftMax margin Conditional Random Field (SMMCRF)

• Maximum likelihood training: maximizes the likelihood of observing the labeling sequences given input sequences

• Margin maximization training: finds a hyperplane that not only separates the training data, but also maximizes the score difference between the true label and other labels (Taskar 2003, Tsochantaridis 2004, Gimpel 2010)

\[
\begin{align*}
\min_w & \quad \frac{1}{2} \|w\|^2 + C \sum \xi_n \\
\text{s.t.} & \quad -w^T f_{x(n)}(y^{(n)}) + \max_y \left( w^T f_{x(n)}(y) + \Delta_{y(n)}(y) \right) \leq \xi_n, \forall n
\end{align*}
\]

– More robust to noise than maximum likelihood training
– Better generalization performance
SMMCRF generalization performance

\[ R(w) \leq R_{emp}(w) + \Omega(N, d, \eta) \]

\[
\min_w \quad \frac{1}{2} \|w\|^2 + C \sum \xi_n
\]

- **Expected risk**
- **Empirical risk**
- **VC confidence**
- **Trade-off parameter**

**Increase C**
- Minimize empirical risk
- Grow VC confidence
- Complicated model

**Decrease C**
- Minimize VC confidence
- Increase empirical risk
- Simplified model
Model adaption in self-training

• Infeasible to tune C parameter of SMMCRF by validation data
  – Training data is limited
  – Auto-labeled data contains noise

• Dynamically update C in self-training
  – Decrease C as unlabeled query sessions are included as training data
  – Avoid complicated model and over-fitting to noises
Model adjustment in self-training (Cont’d)

• Control the margin according to the noise rate estimated using new auto-labeled data

\[ \epsilon \approx \frac{\sum_{num} (1 - p_w(\hat{y}|x))}{num} \]

\[ C' \leftarrow C/(1 + \epsilon) \]

• Drop unconfident instances included in previous runs

\[ \tau_w(x, \hat{y}) = w^T f_x(\hat{y}) - \max_{y'} \left( w^T f_x(y') + \Delta_{\hat{y}}(y') \right) < 0. \]
Confidence measurement

- \( c_1: \) margin based confidence measurement

\[
c^{(1)}_w(x, \hat{y}) = w^T f_x(\hat{y}) - \log \sum_{y' \neq \hat{y}} e^{w^T f_x(y') + \Delta y(y')}
\]

- \( c_2: \) conditional probability

\[
c^{(2)}_w(x, \hat{y}) = p(\hat{y} | x) = \frac{e^{w^T f_x(\hat{y})}}{\sum_{y'} e^{w^T f_x(y')}}
\]

- \( c_3: \) distance to decision boundary

\[
c^{(3)}_w(x, \hat{y}) = -c^{(1)}_w(x, \hat{y}) \cdot I \left( c^{(1)}_w(x, \hat{y}) \geq 0 \right)
\]
Confidence measurement (Cont’d)

<table>
<thead>
<tr>
<th></th>
<th>c1: margin gap</th>
<th>c2: conditional probability</th>
<th>c3: support vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Useful</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
</tbody>
</table>
Experiment datasets

• Dataset 1
  – 3,500 query sessions (8,988 queries)
  – Manually labeled using KDD Cup’05 categories
  – Used by Cao, SIGIR 2009

• Dataset 2
  – 1,727 query sessions (39,565 queries)
  – Manually labeled by search intent type, e.g., “Compare products, services, or activities for use”
Experiment 1: algorithm robustness to training data noise

- Run experiments with Dataset 1
- 80% data used for training, 20% for testing
- Average results of 5 fold experiments

CRF: maximum likelihood
CRF-R: maximum likelihood with regularization
SMMCRF: margin maximization
Experiment 2: confidence measurement

- Run experiment on Dataset 1
- Adopt adaptive self-training with SMMCRF
- 80% as training (10% as labeled data, others as unlabeled), 20% for testing
Experiment 3: margin control methods

- C\rightarrow C : no change to C value
- C\rightarrow C/(1+e): change C according to error rate
- C\rightarrow C/1.05, C\rightarrow C/2: constant update
Overall Results

- SMMCRF used as baseline algorithm
- Dataset 1: 80% as training (10% as labeled data, others as unlabeled), 20% for testing
- Dataset 2: 80% as training (5% as labeled data, others as unlabeled), 20% for testing
- Average results of 10 runs experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>ASCRF</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>44.67% (1.48 × 10^0)</td>
<td>49.94% (8.07 × 10^-1)</td>
<td>11.7%</td>
</tr>
<tr>
<td>2</td>
<td>35.25% (4.95 × 10^-2)</td>
<td>43.53% (4.44 × 10^-3)</td>
<td>23.4%</td>
</tr>
</tbody>
</table>
Conclusion and future work

• Conclusion
  – Search context can be used to improve query classification
  – Adaptive self-training is able to utilize unlabeled data
    • Max margin CRF
    • Control margin to adapt model training
    • Effective confidence measurement method

• Future Work
  – Include more information as search contexts to further improve performance
Thanks!