Predicting Diverse Subsets Using Structural SVMs

Yisong Yue, Thorsten Joachims

Cornell University
Department of Computer Science
Diversified Retrieval

- **Ambiguous queries:**
  - Example query: “SVM”
    - ML method
    - Service Master Company
    - Magazine
    - School of veterinary medicine
    - Sport Verein Meppen e.V.
    - SVM software
    - SVM books
  - “submodular” performance measure
    - make sure each user gets at least one relevant result

- **Learning Queries:**
  - Find all information about a topic
  - Eliminate redundant information

Query: SVM
1. Kernel Machines
2. SVM book
3. SVM-light
4. SVM
5. Service Master Co
6. SV Meppen
8. SVM-light
9. Intro to SVM
10. ...

[YueJo08]
Generic Structural SVM

• **Application Specific Design of Model**
  
  – Loss function $\Delta(y_i, y)$
  
  – Representation $\Phi(x, y)$

• **Prediction:**

  \[
  \hat{y} = \arg\max_{y \in Y} \{ \mathbf{w}^T \Phi(x, y) \}
  \]

• **Training:**

  \[
  \min_{\mathbf{w}, \xi \geq 0} \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{C}{n} \sum_{i=1}^{n} \xi_i \\
  \text{s.t.} \quad \forall y \in Y \setminus y_1 : \mathbf{w}^T \Phi(x_1, y_1) \geq \mathbf{w}^T \Phi(x_1, y) + \Delta(y_1, y) - \xi_1 \\
  \ldots \\
  \forall y \in Y \setminus y_n : \mathbf{w}^T \Phi(x_n, y_n) \geq \mathbf{w}^T \Phi(x_n, y) + \Delta(y_n, y) - \xi_n
  \]

• **Applications:** Parsing, Sequence Alignment, Clustering, etc.
Applying StructSVM to New Problem

• **General**
  – SVM-struct algorithm and implementation
  – Theory (e.g. number of iterations independent of n)

• **Application specific**
  – Loss function $\Delta(y_i, y)$
  – Representation $\Phi(x, y)$
  – Algorithms to compute
    $$\tilde{y} = \arg\max_{y \in Y} \{w^T \Phi(x_i, y)\}$$
    $$\tilde{y} = \arg\max_{y \in Y} \{\Delta(y_i, y) + w^T \Phi(x_i, y)\}$$

• **Properties**
  – General framework for discriminative learning
  – Direct modeling, not reduction to classification/regression
  – “Plug-and-play”
Approach

• **Prediction Problem:**
  - Given set $x$, predict size $k$ subset $y$ that satisfies most users.

• **Approach: Topic Red. $\approx$ Word Red. [SwMaKi08]**
  
  $\Rightarrow$ $y = \{ D1, D2, D3, D4 \}$

  - Weighted Max Coverage: $y = \arg\max_{y \subset x, |y|=k} \left\{ \sum_{w \in \mathcal{U}(y)} \text{score}(w) \right\}$

  - Greedy algorithm is 1-1/e approximation [Khuller et al 97]

  $\Rightarrow$ **Learn the benefit weights:** $\text{score}(w) = w^T \phi(w, x)$

[YueJo08]
Features Describing Word Importance

• **How important is it to cover word w**
  - w occurs in at least X% of the documents in x
  - w occurs in at least X% of the titles of the documents in x
  - w is among the top 3 TFIDF words of X% of the documents in x
  - w is a verb
    → Each defines a feature in $\phi(w, x)$

• **How well a document d covers word w**
  - w occurs in d
  - w occurs at least k times in d
  - w occurs in the title of d
  - w is among the top k TFIDF words in d
    → Each defines a separate vocabulary and scoring function

[D1 D2 D3 D4 D6 D7] + [D1 D2 D3 D4 D6 D7] + … + [D1 D2 D3 D4 D6 D7] [YueJo08]
Loss Function and Separation Oracle

- **Loss function:** $\Delta(y_i, y)$
  - Popularity-weighted percentage of subtopics not covered in $y$
    $\Rightarrow$ More costly to miss popular topics
  - Example:

- **Separation oracle:** $\hat{y} = \arg \max_{y \in Y} \{ \Delta(y_i, y) + \bar{w}^T \Phi(x_i, y) \}$
  - Again a weighted max coverage problem
    $\Rightarrow$ add artificial word for each subtopic with percentage weight
  - Use greedy algorithm again

\[ \Delta(y_i, \{D1, D10\}) = \frac{3}{12} \]
\[ \Delta(y_i, \{D2, D7\}) = \frac{10}{12} \]
Experiments

• **Data:**
  - TREC 6-8 Interactive Track
  - Relevant documents manually labeled by subtopic
  - 17 queries (~700 documents), 12/4/1 training/validation/test
  - Subset size k=5, two feature sets (div, div2)

• **Results:**

<table>
<thead>
<tr>
<th>Method</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.469</td>
</tr>
<tr>
<td>Okapi</td>
<td>0.472</td>
</tr>
<tr>
<td>Unweighted Model</td>
<td>0.471</td>
</tr>
<tr>
<td>Essential Pages</td>
<td>0.434</td>
</tr>
<tr>
<td>SVM$^\Delta_{div}$</td>
<td>0.349</td>
</tr>
<tr>
<td>SVM$^\Delta_{div2}$</td>
<td>0.382</td>
</tr>
</tbody>
</table>