Discovering Diverse and Salient Threads in Document Collections
Jennifer Gillenwater, Alex Kulesza, Ben Taskar
University of Pennsylvania

Document Collection Threading – (1) Build a graph from the collection, using measures of importance and relatedness to weight nodes (documents) and build edges (relationships). (2) From this graph, extract a diverse, salient set of threads to represent the collection.

Motivation: current search tools are insufficient
Random Projections for Tractability
Complexity $D^3$ can be prohibitively large, so we project $D$ down to $d$.

Theorem: Given $	ilde{P}(Y)$ distribution after projecting $D$ to $d = O(\max\{k/\epsilon, (\log(1/\delta) + \log N)/\epsilon^2\})$, error is bounded by:

$$||P^k - \tilde{P}^k|| \leq e^{6\kappa} - 1 \approx 6\kappa$$

with probability at least $1 - \delta$.

Random projections on a small threading task where the exact model is tractable: $n = 600$ and $D = 150$. As predicted by the theorem, fidelity to the true model increases rapidly with $d$.

Introduction

Figure: Prior knowledge of document contents is required to construct a query
Figure: Structure indicating relationships among returned documents is missing

Related threading work

- Selecting a single thread (D. Shahaf and C. Gastruin, KDD 2010)
- Constructing diverse topic threads (A. Ahmed and E. Xing, UAI 2010)

Approach: Determinantal Point Processes

- Decompose quality and similarity of a thread $y_i = (y_{i1}, \ldots, y_{iT})$

  \[ q(y_i) = q(y_{i1}) \prod_{t=2}^{T} q(y_{it}|y_{i(t-1)}, y_{it}) \]

  \[ \phi(y_i) = \sum_{t=1}^{T} \phi(y_{it}) \]

- Score a set of threads $Y$ via structured determinantal point process (SDPP)

  (A. Kulesza and B. Taskar, NIPS 2010)

- SDPP: defines a distribution over sets $Y$

  \[ L_k = q(y_i)\phi(y_i)^T\phi(y_i)q(y_i) \]

  \[ P(Y) = \frac{\det(L_Y)}{\sum_{Y \subseteq \{1, \ldots, n\}} \det(L_Y)} = \frac{\det(L + 1)}{\det(L + 1)} \]

  \[ Y = \{i\} \rightarrow P(Y) \propto q(y_i)^2 \]

  \[ Y = \{i, j\} \rightarrow P(Y) \propto q(y_i)^2q(y_j)^2(1 - (\phi(y_i)^T\phi(y_j))^2) \]

  $\det(L_Y)$ is proportional to volume spanned by the vectors $q(y_i)\phi(y_i)$. As quality (length) or diversity (angle) decreases, volume decreases.

$k$-SDPPs: fix $k$ points in $Y$ to $k$ (A. Kulesza and B. Taskar, ICML 2011)

Sampling from $k$-SDPPs can be done in $O(TrnD^2 + D^3)$, where $r = \max$ node degree, $n = \#$ of nodes, $D = \#$ of features

Example New York Times Graph

New York Times Timelines

- Data – six 6-month NYT article sets
- Graph – edges are tf/idf cosine scores
- Baselines – $k$-means clustering on time slices
- Dynamic topic model (DTM) (D. Blei and J. Lafferty, ICML 2006)

<table>
<thead>
<tr>
<th>ROUGE-SU4</th>
<th>Coherence</th>
<th>Interlopers</th>
<th>Secs</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-means</td>
<td>3.76</td>
<td>2.73</td>
<td>0.71</td>
</tr>
<tr>
<td>DTM</td>
<td>3.44</td>
<td>3.19</td>
<td>1.10</td>
</tr>
<tr>
<td>k-SDPP</td>
<td>3.98</td>
<td>3.31</td>
<td>1.15</td>
</tr>
</tbody>
</table>

Table: ROUGE-SU4: comparison to human summaries. Mechanical Turk: thread coherence rating (1-5); average # of random interloper articles identified. Secs: runtime.

Figure: A set of threads from a k-SDPP (left) and a DTM (right). Above, threads are shown with the most salient words superimposed; below, headlines from the last thread are listed.

Figure: Top left: Full graph with 5 DPP threads. Other: Zoom in.