Motivation

- Ranking: match a query to document (find documents that are most relevant to the query)
- Most of the current methods use hand-coded features
- Our Goal: learning to rank directly from words

- Vector Space Model (query/document as vector)
  \[ q, d = [w_1, \ldots, w_D] \in \mathbb{R}^D \]
  \( D \): vocabulary size
  \( w_i \): normalized weight (tf-idf) of \( i \)-th word

- Cosine similarity (relevance of a document to a query)
  \[ f(q, d) = q^T d \]

- Does not deal with synonyms 😞
- No machine learning 😞
Basic Model

- **Similarity Score (relevance of documents to a query)**
  \[ f(q, d) = q^T W d = \sum_{i,j} W_{ij} (q_i \cdot d_j) \]

  \( W_{ij} \): relationship/correlation between \( q_i \) and \( d_j \)

- **Goal:** Learn \( W \) matrix

- **Generalization ability:** \( W \in \mathbb{R}^{D \times D} \) \( D \): vocabulary size
  \( D = 10,000, \ D^2 = 10^8 \) free parameters to model

- **Memory issues:** \( D = 10,000 \): \( W \) needs 1GB Memory

- **Computational cost:** \( q^T W d \)

\( W \): Sparse Matrix !!!
Training Framework

- **Data:**
  Tuples $\mathcal{R}$: query $q$, related doc. $d^+$, unrelated doc. $d^-$.

- **Learn $W$ such that:**
  \[ f(q, d^+) = q^T W d^+ > f(q, d^-) = q^T W d^- \]

- **Margin rank loss:**
  \[ L_W(q, d^+, d^-) = \max(0, 1 - q^T W d^+ + q^T W d^-) \]

- **Model:**
  \[ W^* = \arg \min_W \frac{1}{|\mathcal{R}|} \sum_{(q,d^+,d^-) \in \mathcal{R}} L_W(q, d^+, d^-) + \lambda \|W\|_1. \]
Training Algorithm

- **Stochastic (sub)Gradient Descent:**
  \[
  \nabla L_{W^t}(q, d^+, d^-) = \begin{cases} 
  -q(d^+ - d^-)^\top & \text{if } q^\top W^t(d^+ - d^-) < 1 \\
  0 & \text{otherwise}
  \end{cases}
  \]
  \[
  W^{t+1} = W^t - \eta_t \nabla L_{W^t}(q, d^+, d^-)
  \]
  \(\eta_t = \frac{C}{\sqrt{t}}\): decaying rate
  \(\eta_t = C\): fixed rate

- **Mini-batch Shrinkage Strategy (every T iterations):**
  \[
  \widehat{W}^t = \text{argmin}_W \frac{1}{2} \| W - W^t \|_F^2 + \lambda \sum_{k=t-T+1}^{t} \eta_t \| W \|_1
  \]
  \[
  \Rightarrow \widehat{W}^t_{i,j} \rightarrow 0
  \]

- **Refitting Step** (Reduce the bias of \(\ell_1\) regularization):
  Fixing zeros elements and training the remaining elements without L1-regularization
Experimental Results

- Ranking Performance (multi-class classification data, relate doc. are in the same class)

![Graphs showing test error rate and density of W over number of data accesses for 20 News Group (20NG) and RCV1 datasets.](image)

\[ D = 10,000 \]
## Experimental Results

### 20NG

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>Test Error (%)</th>
<th>Memory (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
<td>0.185</td>
<td>32.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Diagonal</td>
<td>0.190</td>
<td>31.8</td>
<td>0.2</td>
</tr>
<tr>
<td>SGD (fixed learning rate)</td>
<td>0.258</td>
<td>19.7</td>
<td>1294</td>
</tr>
<tr>
<td>SGD (decaying learning rate)</td>
<td>0.399</td>
<td>9.9</td>
<td>943.1</td>
</tr>
<tr>
<td>Sparse</td>
<td>0.360</td>
<td>11.4</td>
<td>154.2</td>
</tr>
<tr>
<td>Sparse (refitting)</td>
<td>0.426</td>
<td>9.0</td>
<td>154.2</td>
</tr>
</tbody>
</table>

### RCV1

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>Test Error (%)</th>
<th>Memory (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
<td>0.380</td>
<td>23.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Diagonal</td>
<td>0.390</td>
<td>22.3</td>
<td>0.2</td>
</tr>
<tr>
<td>SGD (fixed learning rate)</td>
<td>0.451</td>
<td>8.7</td>
<td>717.2</td>
</tr>
<tr>
<td>SGD (decaying learning rate)</td>
<td>0.453</td>
<td>4.6</td>
<td>360.2</td>
</tr>
<tr>
<td>Sparse</td>
<td>0.463</td>
<td>3.6</td>
<td>105.4</td>
</tr>
<tr>
<td>Sparse (refitting)</td>
<td>0.501</td>
<td>2.9</td>
<td>105.4</td>
</tr>
</tbody>
</table>

$\mathcal{D} = 10,000$
## Experimental Results

- **Learned word-relationship from Sparse W (20 News Group)**

<table>
<thead>
<tr>
<th>Query word</th>
<th>Most related document words</th>
</tr>
</thead>
<tbody>
<tr>
<td>atheism</td>
<td>keith atheists god caltech</td>
</tr>
<tr>
<td>clinton</td>
<td>government health people gay</td>
</tr>
<tr>
<td>cpu</td>
<td>mac drive scsi card</td>
</tr>
<tr>
<td>graphics</td>
<td>tiff image color polygon</td>
</tr>
<tr>
<td>handgun</td>
<td>gun weapons militia fbi</td>
</tr>
<tr>
<td>hockey</td>
<td>game espn colorado team</td>
</tr>
<tr>
<td>religions</td>
<td>god bible christian jesus</td>
</tr>
</tbody>
</table>