Phrase Mining, Key Phrase Identification and Similar Document Search

Survey Report for CS512

Yu Shi, Karan Ghai and Zequn Zhang
• Course Project:

• Capturing synergetic co-occurrence relevance

5 + 1 + 1 should not simply be 7 (synergy)
Course Project:

- Capturing synergetic co-occurrence relevance
- With application to similar research paper search

Survey: (data-driven)

- Phrase mining
- Key phrase identification (as representative concepts)
- Similar document search
Phrase Mining from Corpora

• Constructing Topical Hierarchies in Heterogeneous Information Networks (CathyHIN)

• Scalable Topical Phrase Mining from Text Corpora (ToPMine)

• Mining Quality Phrases from Massive Text Corpora (SegPhrase)
Constructing Topical Hierarchies in Heterogeneous Information Networks (CathyHIN)

- Associated heterogeneous information network facilitates phrase mining
- Generative model
- FP mining + phraseness measures
- Different granularities
- Phrase: orderless set

Figure origin: surveyed paper
Scalable Topical Phrase Mining from Text Corpora (ToPMine)

- (To compare with next method)
- Significant score
- Bottom-up merge
- Simple but effective, highly scalable

\[ \text{sig}(P_1, P_2) \approx \frac{f(P_1 \oplus P_2) - \mu_0(P_1, P_2)}{\sqrt{f(P_1 \oplus P_2)}} \]
Mining Quality Phrases from Massive Text Corpora (SegPhrase)

- Concordance features, informative features
- Random forest
- Rectified frequency
  - Rectified frequency boost phrase quality
  - Updated phrasal segmentation in turn updates rectified frequency
- Start with a small number (300) of labeled phrase
- Dynamic programming

![Table]

<table>
<thead>
<tr>
<th>sequence</th>
<th>frequency</th>
<th>phrase?</th>
<th>rectified</th>
</tr>
</thead>
<tbody>
<tr>
<td>relational</td>
<td>100</td>
<td>yes</td>
<td>70</td>
</tr>
<tr>
<td>database system</td>
<td>150</td>
<td>yes</td>
<td>40</td>
</tr>
<tr>
<td>relational</td>
<td>160</td>
<td>yes</td>
<td>35</td>
</tr>
<tr>
<td>database system</td>
<td>500</td>
<td>N/A</td>
<td>20</td>
</tr>
<tr>
<td>relational</td>
<td>1000</td>
<td>N/A</td>
<td>200</td>
</tr>
<tr>
<td>database system</td>
<td>10000</td>
<td>N/A</td>
<td>1000</td>
</tr>
</tbody>
</table>
Mining Quality Phrases from Massive Text Corpora (SegPhrase)

Figure origin: surveyed paper
Key Technical Phrase Identification

- With high quality phrases extracted
- We need to identify which phrases are key so as to be representative concepts
- We can categorize papers using venue information (survey methods that can use this information)
Key Technical Phrase Identification

- Interesting-Phrase Mining for Ad-Hoc Text Analytics
- Mining Quality Phrases from Massive Text Corpora (SegPhrase)
- Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents (KERT)
- Scalable Topical Phrase Mining from Text Corpora (ToPMine)
Interesting-Phrase Mining for Ad-Hoc Text Analytics

• Current subcollection (category) compared to whole corpus

• Interestingness

\[ I_D (p, D^\prime) = \frac{freq(p, D^\prime)}{freq(p, D)} \]
Mining Quality Phrases from Massive Text Corpora (SegPhrase)

- Key technical phrases identification reported as an application of SegPhrase
- Interestingness * frequency

\[ \frac{freq(v, C')^2}{freq(v, C)} \]
Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents (KERT)

- Among four measures coverage, purity, phraseness and completeness

- Purity is about representativeness

\[
\pi_{t}^{pur}(p) = \log \frac{P(e_t(p))}{\max_{t' \neq t} P(e_{t'}(p))} = \log \frac{f_t(p)}{|D_t|} - \log \max_{t' \neq t} \frac{f_t(p) + f_{t'}(p)}{|D_{t'}|}
\]

- Similar to interestingness
Scalable Topical Phrase Mining from Text Corpora (ToPMine)

- If forgo category information
- Frequent phrases in each topic could be representative (PhraseLDA)
- # of topics need to be big to identify key phrase in highly specified sub-category
- LDA running time linear to # of topics
Similar Document Search

- Similarity Measures for Text Document Clustering
- Detecting Similar Documents Using Salient Terms
- Unifying Learning to Rank and Domain Adaptation: Enabling Cross-Task Document Scoring
- Beyond Keyword Search: Discovering Relevant Scientific Literature
- Citation Prediction in Heterogeneous Bibliographic Networks
- ClusCite: Effective Citation Recommendation by Information Network-Based Clustering
Similarity Measures for Text Document Clustering

• Most basic document representation methods (vector space)
  • TF, TF-IDF

• Most basic similarity measures
  • Euclidean, cosine, Jaccard, Pearson correlation, averaged Kullback–Leibler divergence
Similarity Measures for Text Document Clustering

- Evaluate using benchmark datasets with class labels after clustering (k-means)

<table>
<thead>
<tr>
<th>Data</th>
<th>Euclidean</th>
<th>Cosine</th>
<th>Jaccard</th>
<th>Pearson</th>
<th>KLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20news</td>
<td>0.1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.38</td>
</tr>
<tr>
<td>classic</td>
<td>0.56</td>
<td>0.85</td>
<td>0.98</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td>hitech</td>
<td>0.29</td>
<td>0.54</td>
<td>0.51</td>
<td>0.56</td>
<td>0.53</td>
</tr>
<tr>
<td>re0</td>
<td>0.53</td>
<td>0.78</td>
<td>0.75</td>
<td>0.78</td>
<td>0.77</td>
</tr>
<tr>
<td>tr41</td>
<td>0.71</td>
<td>0.71</td>
<td>0.72</td>
<td>0.78</td>
<td>0.64</td>
</tr>
<tr>
<td>wap</td>
<td>0.32</td>
<td>0.62</td>
<td>0.63</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>webkb</td>
<td>0.42</td>
<td>0.68</td>
<td>0.57</td>
<td>0.67</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Detecting Similar Documents Using Salient Terms

- Using Textract (software, NLP-based) to extract salient terms
- (Terms may not be key technical terms)
- Two document dissimilar
  
- # of distinct salient terms > theta
Unifying Learning to Rank and Domain Adaptation: Enabling Cross-Task Document Scoring

• Decouple feature into intra-features and meta-features as TF-IDF to TF and IDF

\[ F(d, t) = \sum_{i,j} \beta_{ij} \sum_{w \in W(d, t)} f_i^{(M)}(w, t) \cdot f_j^{(I)}(w, d) \]

• Recouple to compute similarity

• Learn with Tree-Structured Restricted Boltzmann machine

• Very general
Citation Prediction in Heterogeneous Bibliographic Networks

- Papers with citation relation are in a sense similar
- Meta-path based similarities & logistic regression
- Term bucket (cluster): may or may not boost performance
ClusCite: Effective Citation Recommendation by Information Network-Based Clustering

- Heterogeneous information network

- Soft cluster into $K$ interest groups

- For each interest group $k$, model similarity and authority

$$s(q, p) = \sum_{k=1}^{K} \theta_q^{(k)} \cdot \{r^{(k)}(q, p) + f_p^{(k)}(p)\}.$$
• Phrase Mining from Corpora

• Key Technical Phrase Identification

• Similar Document Search

• With emphasis on unsupervised methods (or little supervision)
Thanks!