Multilingual Document Retrieval through Hub Languages

Jan Rupnik, Andrej Muhič, Primož Škraba

AI Lab
Jozef Stefan Institute
Ljubljana, Slovenia

Si-KDD, October 8, 2012

Data Mining

Data Mining: Describe a research activity of the knowledge discovery in databases (KDD) process, it is the process of discovering new patterns from large data sets involving methods of the analysis of artificial intelligence, machine learning, statistics, and database systems. The goal of data mining is to extract knowledge from a data set in a human-understandable and interpretable form and involve database and data management, data preprocessing, model and hypothesis evaluations, and interestingness metrics. The field of data mining is a relatively young and rapidly developing field of computer science, which is described as "the process of discovering interesting patterns in large data sets". Even the popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally titled "Practical machine learning", and the term "data mining" was only added for marketing reasons.

Often the more general terms "large scale data analysis" or "data mining" or when referring to actual methods, artificial intelligence and machine learning are more appropriate.

Minería de datos

La minería de datos (DM) es el proceso de extraer información de los datos. La minería de datos se puede definir como el proceso de extraer patrones de información de datos grandes. La minería de datos es una técnica de procesamiento de datos que permite analizar grandes cantidades de datos para encontrar patrones interesantes. Estos patrones pueden ser útiles para la toma de decisiones en diferentes campos como la salud, la finanzas, la ciencia de la información, el marketing, la inteligencia artificial, etc.
Two Language Example

- Vector space model document representation
- $X$, $Y$ two vector spaces (two languages)
- Find a basis for $X$: $S_X = \begin{bmatrix} \mathbf{s}_1^X & \mathbf{s}_2^X & \mathbf{s}_3^X & \ldots \end{bmatrix}$
- and a basis for $Y$: $S_Y = \begin{bmatrix} \mathbf{s}_1^Y & \mathbf{s}_2^Y & \mathbf{s}_3^Y & \ldots \end{bmatrix}$

Euclidean space

$s_{X,Y}(x, y) := s_{\text{Euclid}}(S_X(x), S_Y(y))$
The two bases should be aligned: $s_i^X$ should represent the same topic/concept as $s_i^Y$.

**Cross-lingual similarity**: express the documents in the aligned bases and compare the coefficient vectors.
Aligned multilingual collection of documents

\[ D_{\text{Total}} = \left[ D_1^T, D_2^T, \ldots, D_\ell^T \right]^T \]

Compute low rank approximation \( D_{\text{Total}} \approx \sum_{i=1}^{k} s_i \cdot u_i \cdot v_i^T \)

Split each vector \( u_k \) into vectors of dimensions \( n_{x_1}, \ldots, n_{x_\ell} \).

Map to language independent topic space \( V \):

\[ x_1 \mapsto S^{-1} U^{(1)^+} x_1, \quad x_2 \mapsto S^{-1} U^{(2)^+} x_2 \]
Three versions of mapping investigated, aligned sets $X$ and $Y$:

- **Cross-Lingual Latent Semantic Indexing (CL-LSI)**
- **LSQ** Least squares
  
  find $W$ of rank $k$ which minimizes

  $$
  \min ||WX - Y||_F,
  $$

  use truncated SVD of the input $X$

  $$
  W = YX^+ \approx YV_k \Sigma_k^{-1} U_k^*,
  $$

- **RCCA** regression canonical correlation analysis map

  $$
  q \mapsto (XX')^{-1} XY' q \approx U_k \Sigma_k^{-1} V_k^* Y' q
  $$
Multilingual Setting

- Distribution of languages not even
  1. English - 4,069,555
  2. German - 1,475,728
  36. Slovene - 134,447
  40. Hindi - 103,723

- English and German - large intersection

- What about Slovene and Hindi?
The larger the set of correspondences \( \rightarrow \) easier the problem how to use the additional information present in the hub languages?

Combinatorial
Maps between representations
Experiments

- Compare Slovene and Hindi using English as a hub
- Preprocessing: removed stub documents
- Input:
  - 44,426 Slovenian-English correspondences
  - 14,121 Hindi-English correspondences
  - 4,034 Slovenian-Hindi correspondences (4,017 joint English-Hindi-Slovenian correspondences)
Five different ways to do retrieval

1. $sl \leftrightarrow hi$,
2. $hi \leftrightarrow sl$.
3. $sl \leftrightarrow en = hub = en \leftrightarrow hi$,
4. $sl \leftrightarrow en \leftrightarrow hi$
5. $hi \leftrightarrow en \leftrightarrow sl$.
Three different subsets of the data

- *all* – we use all alignment information available,
- *common* – we use only alignment information consistent through all three languages
- *empty* – we remove all common alignment to simulate the case where we are forced to use hubs.
## Results

Slovenian-Hindi MAPR retrieval using different maps

<table>
<thead>
<tr>
<th></th>
<th>sl→hi</th>
<th>hi→sl</th>
<th>sl→en←hi</th>
<th>sl→en→hi</th>
<th>hi→en→sl</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSQ all</td>
<td>0.42</td>
<td>0.49</td>
<td>0.38</td>
<td>0.35</td>
<td>0.43</td>
</tr>
<tr>
<td>RCCA all</td>
<td><strong>0.55</strong></td>
<td>0.45</td>
<td><strong>0.38</strong></td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>LSQ common</td>
<td>0.42</td>
<td>0.48</td>
<td>0.47</td>
<td>0.42</td>
<td><strong>0.49</strong></td>
</tr>
<tr>
<td>RCCA common</td>
<td><strong>0.55</strong></td>
<td>0.46</td>
<td>0.39</td>
<td>0.35</td>
<td><strong>0.38</strong></td>
</tr>
<tr>
<td>LSQ empty</td>
<td>N/A</td>
<td>N/A</td>
<td>0.27</td>
<td>0.28</td>
<td><strong>0.35</strong></td>
</tr>
<tr>
<td>RCCA empty</td>
<td>N/A</td>
<td>N/A</td>
<td><strong>0.32</strong></td>
<td>0.22</td>
<td>0.21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>sl→hi</th>
<th>sl→en←hi</th>
<th>sl→en→hi</th>
<th>hi→en→sl</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL-LSI all</td>
<td><strong>0.585</strong></td>
<td>0.35</td>
<td><strong>0.56</strong></td>
<td>0.54</td>
</tr>
<tr>
<td>CL-LSI common</td>
<td><strong>0.58</strong></td>
<td>0.47</td>
<td><strong>0.61</strong></td>
<td><strong>0.61</strong></td>
</tr>
<tr>
<td>CL-LSI empty</td>
<td>0</td>
<td>0.24</td>
<td><strong>0.48</strong></td>
<td>0.46</td>
</tr>
</tbody>
</table>
Results

Gluing all documents together

Table: Common

<table>
<thead>
<tr>
<th></th>
<th>sl</th>
<th>en</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>sl</td>
<td>0</td>
<td>0.77</td>
<td>0.45</td>
</tr>
<tr>
<td>en</td>
<td>0.73</td>
<td>0</td>
<td>0.64</td>
</tr>
<tr>
<td>hi</td>
<td>0.38</td>
<td>0.67</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: All

<table>
<thead>
<tr>
<th></th>
<th>sl</th>
<th>en</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>sl</td>
<td>0</td>
<td>0.81</td>
<td>0.6</td>
</tr>
<tr>
<td>en</td>
<td>0.77</td>
<td>0</td>
<td>0.71</td>
</tr>
<tr>
<td>hi</td>
<td>0.61</td>
<td>0.76</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: Empty

<table>
<thead>
<tr>
<th></th>
<th>sl</th>
<th>en</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>sl</td>
<td>0</td>
<td>0.37</td>
<td>0.22</td>
</tr>
<tr>
<td>en</td>
<td>0.49</td>
<td>0</td>
<td>0.36</td>
</tr>
<tr>
<td>hi</td>
<td>0.11</td>
<td>0.29</td>
<td>0</td>
</tr>
</tbody>
</table>
Testing the Representation

Ideal retrieval under misalignment

- compare documents between Slovenian and Hindu
- only two aligned sets (en-sl, en-hi) with a third language and no common information
Testing the Representation

Topic drift between two English training sets, (en-sl, en-hi)

- aligned sets, $T_1$ and $T_2$ correspond to $sl \leftrightarrow en_1$ and $hi \leftrightarrow en_2$
- $T_1$ and $T_2$ are topic drifted
- Test the retrieval on ideally projected Slovene and Hindi test documents to English document space.
- MAPR of 0.96
Conclusions

- Take advantage of the distribution of languages
- Representations mean we take advantage of indirect correspondences (can work with no overlap)
  - Linear representation captures relevant information

- Space where we do retrieval matters
- Non-symmetry in maps
- Joint optimization