Document Clustering via Matrix Representation

Xufei Wang,
Jiliang Tang and Huan Liu
Arizona State University
News Article

- Lead Paragraph
- Explanations
- Additional Information
Research Paper

• Introduction
• Related Work
• Problem Statement
• Solution
• Experiment
• Conclusion
Book

- Chapters
- References
- Appendix
Document Organization

• Not randomly organized

• Put relevant content together

• Logically independent segments
Matrix Space Model

• Represent a document as a matrix
  – Segment
  – Term

• Each segment is a vector of terms
  – Terms + frequency
Vector Space Model

- Oversimplify a document
  - Mixing topics
  - Word order is lost
  - Susceptible to noise

\[ V_d = (w_{1,d}, w_{2,d}, \ldots, w_{N,d})^T \]
The IEEE International Conference on Data Mining series (ICDM) has established itself as the world's premier research conference in data mining. It provides an international forum for presentation of original research results, as well as exchange and dissemination of innovative, practical development experiences. The conference covers all aspects of data mining, including algorithms, software and systems, and applications.

Vancouver is a coastal seaport city on the mainland of British Columbia, Canada. It is the hub of Greater Vancouver, which, with over 2.3 million residents, is the third most populous metropolitan area in the country, and the most populous in Western Canada.
### Segment 1

<table>
<thead>
<tr>
<th>conference</th>
<th>data</th>
<th>mining</th>
<th>research</th>
<th>international</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

### Segment 2

<table>
<thead>
<tr>
<th>vancouver</th>
<th>populous</th>
<th>canada</th>
<th>metropolitan</th>
<th>columbia</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Vector Space Model

<table>
<thead>
<tr>
<th>conference</th>
<th>data</th>
<th>mining</th>
<th>research</th>
<th>international</th>
<th>canada</th>
<th>vancouver</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Pros of a Matrix Representation

• Interpretation:
  – Segments vs. topics

• Finer granularity for data management
  – Segments vs. document

• Multiple or Single class labels
  – Flexibility
Verify the Effectiveness via Clustering

- **Information Retrieval**
  - Indexing based on segments

- **Classification**
  - New approaches based on Matrix inputs

- **Clustering**
  - New approaches based on Matrix inputs
A Graphical Interpretation

Figure 1. Graph Interpretation
Step 1: Obtaining Segments

• Many approaches for segmentation
  – Terms
  – Sentences (Choi et al. 2000)
  – Paragraphs (Tagarelli et al. 2008)

• Determining the number of segments
  – Open research problem
Step 2: Latent Topic Extraction

- Non-negative Matrix Approximation (NMA)
  \[
  \min_{L \in \mathbb{R}^{r \times \ell_1}, \quad M_i \in \mathbb{R}^{\ell_1 \times \ell_2}, \quad R \in \mathbb{R}^{c \times \ell_2}} \quad \sum_{i=0}^{n} ||A_i - LM_i R^T||_F^2
  \]
  \[
  \text{subject to:} \quad L \geq 0, \quad M_i \geq 0, \quad R \geq 0
  \]

- $LM_i$ represents the probability of a term belonging to a latent topic
- $M_i R^T$ represents the probability of a segment belonging to a latent topic
Step 3: Clustering

- Un-overlapping clustering

\[ \min_k \sum_i \| d_i - \text{centroid}(c) \|^2 \]

\[ d_i = \sum_j d_{ij} \]

- Overlapping clustering

\[ \min_k \sum_{ij} \| d_{ij} - \text{centroid}(c) \|^2 \]
Datasets

- 20newsgroup
  - 20 classes
  - 6,038 documents

- Reuters-21578
  - 26 clusters
  - 1,964 documents

- Classic
  - 3 clusters
  - 1,486 documents
Experimental Method

• Generate Datasets (by specifying k)

• Evaluate the accuracy

• Repeat
Number of Latent Topics

Accuracy vs Number of Latent Topics

- Red line: Newsgroup
- Blue line: Reuters
- Black line: Classic
### Number of Segments

#### Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td><strong>Newsgroup</strong></td>
<td>22.85</td>
</tr>
<tr>
<td><strong>Reuters</strong></td>
<td>7.61</td>
</tr>
<tr>
<td><strong>Classic</strong></td>
<td>11.74</td>
</tr>
</tbody>
</table>

#### Number of Segments

<table>
<thead>
<tr>
<th>Number of Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>

---

**Note:** This table and figure illustrate the distribution of number of segments across different datasets and number of clusters.
Comparative Study

<table>
<thead>
<tr>
<th>k</th>
<th>k-means</th>
<th>NMF</th>
<th>PLSI</th>
<th>LDA</th>
<th>NMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>69.77</td>
<td>72.04</td>
<td>84.87</td>
<td>88.66</td>
<td>88.97</td>
</tr>
<tr>
<td>3</td>
<td>49.58</td>
<td>60.98</td>
<td>76.39</td>
<td>84.84</td>
<td>82.14</td>
</tr>
<tr>
<td>4</td>
<td>40.43</td>
<td>49.93</td>
<td>71.64</td>
<td>81.14</td>
<td>77.79</td>
</tr>
<tr>
<td>5</td>
<td>34.52</td>
<td>41.92</td>
<td>69.91</td>
<td>78.73</td>
<td>69.72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>k</th>
<th>k-means</th>
<th>NMF</th>
<th>PLSI</th>
<th>LDA</th>
<th>NMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>77.68</td>
<td>86.68</td>
<td>79.08</td>
<td>83.33</td>
<td>94.47</td>
</tr>
<tr>
<td>3</td>
<td>70.37</td>
<td>81.75</td>
<td>72.49</td>
<td>80.95</td>
<td>91.32</td>
</tr>
<tr>
<td>4</td>
<td>75.96</td>
<td>82.25</td>
<td>77.27</td>
<td>81.57</td>
<td>86.95</td>
</tr>
<tr>
<td>5</td>
<td>73.85</td>
<td>83.09</td>
<td>73.78</td>
<td>79.11</td>
<td>85.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>k</th>
<th>k-means</th>
<th>NMF</th>
<th>PLSI</th>
<th>LDA</th>
<th>NMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>76.71</td>
<td>87.88</td>
<td>97.49</td>
<td>97.78</td>
<td>98.36</td>
</tr>
<tr>
<td>3</td>
<td>77.26</td>
<td>86.29</td>
<td>96.60</td>
<td>97.57</td>
<td>96.45</td>
</tr>
</tbody>
</table>
Conclusion

• Proposing a matrix representation for documents

• Significant improvements with MSM

• Information Retrieval, classification tasks
Questions