Visualização Multi-dimensional:

Mineração Visual de Dados multidimensionais e aplicações
Parte II – Variações e Aplicações

Rosane Minghim
2008-2012

Basic Concepts

• Text Preprocessing
• Data and text mining
• Projection techniques
• Point Placement Strategies
Text Preprocessing

1. Stopwords elimination
2. Extraction of words radicals (stemming)
3. Creation of n-grams
4. Frequency count and Luhn's lower cut (n-grams appearing less than x times are ignored)
5. Weighting process (*term-frequency inverse document-frequency* - (tfidf))

Result is a Vector Model

- Attributes: terms (n-grams)
- Value: term weight
- Table Data
Vector Representation – term weighting

- \( tf \) – term frequency
- \( tfidf \) – \( tf \times idf = tf \times \text{inverse document frequency} \)

\[
\omega_{ik} = t_{fk} \times \log \left( \frac{N}{n_k} \right)
\]

Vector Representation

<table>
<thead>
<tr>
<th>Doc (_1)</th>
<th>term(_1)</th>
<th>term(_2)</th>
<th>term(_3)</th>
<th>term(_4)</th>
<th>...</th>
<th>term(_m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.92</td>
<td>0.62</td>
<td>0.92</td>
<td>0.10</td>
<td>...</td>
<td>...</td>
<td>0.67</td>
</tr>
<tr>
<td>Doc(_2)</td>
<td>0.13</td>
<td>0.11</td>
<td>1.00</td>
<td>0.34</td>
<td>...</td>
<td>0.33</td>
</tr>
<tr>
<td>Doc(_3)</td>
<td>0.52</td>
<td>0.00</td>
<td>0.00</td>
<td>0.44</td>
<td>...</td>
<td>0.77</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Doc(_n)</td>
<td>0.02</td>
<td>0.12</td>
<td>0.22</td>
<td>0.92</td>
<td>...</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Vector Representation – Similarity calculation

**EUCLIDEAN**

\[ \text{sim}_{i,j} = \sqrt{(w_{i,1} - w_{j,1})^2 + \ldots + (w_{i,k} - w_{j,k})^2} \]

**MANHATAN**

\[ \text{sim}_{i,j} = |w_{i,1} - w_{j,1}| + \ldots + |w_{i,k} - w_{j,k}| \]

**COSINE**

\[ \text{sim}_{i,j} = \frac{(w_{i,1} \times w_{j,1}) + \ldots + (w_{i,k} \times w_{j,k})}{\sqrt{(w_{i,1}^2 + \ldots + w_{i,k}^2) \times (w_{j,1}^2 + \ldots + w_{j,k}^2)}} \]

Alternative to Vector Representation

- Similarity Calculation text against text
- Ex: NCD Normalized Compression Distance
  - Approximation of Kolmogorov Complexity
Visual representations: graphs, surfaces, volumes, triangulations
IN-SPIRE

• Spatial Paradigm for Information Retrieval - Pacific Northwest National Laboratories

• Two Visualization Metaphors:
  • Galaxies – dimensional reduction
  • Themescape

InfoSky

Granitzer (Granitzer et al., 2004) also employs galaxy metaphor
VxInsight

- Sandia National Laboratories, mountain metaphor (Boyack et al., 2002).
HIVE (Ross and Chalmers 2003)

- Interconnected components:
  - Import
  - Transform
  - Render multi-dim data

Projection Explorer (PEx)

- Projection and Point placement
- Precision
- Graphs and surfaces (Super Spider)
Mapping Text Collections via Projections and Point Placement

• Positioning and labeling
• Detailing topics
Finding Relationships
• Building a mesh
• Coloring by degree of proximity
• Coordinating
• Building a Surface
Exemplos de Mapas
Exemplos de Mapas

RSS News Flash

Bird and Flu
Curvas de Nível

Séries Temporais – Vazão em Hidrelétricas

Figure 2. Power plants of the basin Paraná
Further Example - patents

Images
Exercícios

• Download Pex
• Usar os conjuntos de dados presentes no site e explorá-los.
• Registrar as conclusões

• Download Pex-Image
• Fazer o mesmo

• Infoserver.lcad.icmc.usp.br/infovis2
Referências


Referências