A comparison between Latent Semantic Analysis and Correspondence Analysis

Julie Séguéla, Gilbert Saporta

CNAM, Cedric Lab
Multiposting.fr

February 9th 2011 - CARME
Outline

1. Introduction

2. Latent Semantic Analysis
   - Presentation
   - Method

3. Application in a real context
   - Presentation
   - Methodology
   - Results and comparisons

4. Conclusion
Outline

1 Introduction
2 Latent Semantic Analysis
3 Application in a real context
4 Conclusion
Introduction

Context

Text representation for categorization task
Objectives

- Comparison of several text representation techniques through theory and application

- In particular, comparison between a statistical technique: Correspondence Analysis, and an information retrieval (IR) oriented method: Latent Semantic Analysis

- Is there an optimal technique for performing document clustering?
Outline

1. Introduction
2. Latent Semantic Analysis
   - Presentation
   - Method
3. Application in a real context
4. Conclusion
Uses of LSA

- LSA was patented in 1988 (US Patent 4,839,853) by Deerwester, Dumais, Furnas, Harshman, Landauer, Lochbaum and Streeter.
- Find semantic relations between terms
- Helps to overcome synonymy and polysemy problems
- Dimensionality reduction (from several thousands of features to 40-400 dimensions)

Applications

- Document clustering and document classification
- Matching queries to documents of similar topic meaning (information retrieval)
- Text summarization
- ...
LSA theory

How to obtain document coordinates?

1) Document-Term matrix
   \[
   T = \begin{pmatrix}
   \cdots & f_{ij} & \cdots \\
   \vdots & \ddots & \vdots \\
   \vdots & & \ddots
   \end{pmatrix}
   \]

2) Weighting
   \[
   T_W = \begin{pmatrix}
   \cdots & l_{ij}(f_{ij}) \cdot g_j(f_{ij}) & \cdots \\
   \vdots & \ddots & \vdots \\
   \vdots & & \ddots
   \end{pmatrix}
   \]

3) SVD
   \[
   T_W = U\Sigma V' \]

4) Document coordinates in the latent semantic space:
   \[
   C = U_k \Sigma_k
   \]

- We need to find the optimal dimension for final representation
# Common weighting functions

## Local weighting

<table>
<thead>
<tr>
<th>Term frequency</th>
<th>$l_{ij}(f_{ij}) = f_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary</strong></td>
<td>$l_{ij}(f_{ij}) = 1$ if term $j$ occurs in document $i$, else 0</td>
</tr>
<tr>
<td><strong>Logarithm</strong></td>
<td>$l_{ij}(f_{ij}) = \log(f_{ij} + 1)$</td>
</tr>
</tbody>
</table>

## Global weighting

<table>
<thead>
<tr>
<th>Normalisation</th>
<th>$g_j(f_{ij}) = \frac{1}{\sqrt{\sum_i f_{ij}^2}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IDF</strong> (Inverse Document Frequency)</td>
<td>$g_j(f_{ij}) = 1 + \log\left(\frac{n}{n_j}\right)$</td>
</tr>
<tr>
<td>$n$ : number of documents</td>
<td></td>
</tr>
<tr>
<td>$n_j$ : number of documents in which term $j$ occurs</td>
<td></td>
</tr>
<tr>
<td><strong>Entropy</strong></td>
<td>$g_j(f_{ij}) = 1 - \sum_i \frac{f_{ij} \log\left(\frac{f_{ij}}{f_{ij}}\right)}{\log(n)}$</td>
</tr>
</tbody>
</table>
Latent Semantic Analysis

1) \( T = [f_{ij}]_{i,j} \)

2) \( T_W = [l_{ij}(f_{ij}) \cdot g_j(f_{ij})]_{i,j} \)

3) \( T_W = U \Sigma V' \)

4) \( C = U_k \Sigma_k \)

CA: \( l_{ij}(f_{ij}) = \frac{f_{ij}}{\sqrt{f_i}} \) and \( g_j(f_{ij}) = \frac{1}{\sqrt{f_j}} \)

Correspondence Analysis

1) \( T = [f_{ij}]_{i,j} \)

2) \( T_W = \left[ \frac{f_{ij}}{\sqrt{f_i f_j}} \right]_{i,j} \)

3) \( T_W = U \Sigma V' \)

3') \( \tilde{U} = \text{diag}(\sqrt{\frac{f}{f_i}})U \)

4) \( C = \tilde{U}_k \Sigma_k \)
Outline

1. Introduction

2. Latent Semantic Analysis

3. Application in a real context
   - Presentation
   - Methodology
   - Results and comparisons

4. Conclusion
Objectives

- Corpus of job offers
- Find the best representation method to assess "job similarity" between offers in a non-supervised framework
- Comparison of several representation techniques
- Discussion about the optimal number of dimensions to keep
- Comparison between two similarity measures
Data

- Offers have been manually labeled by recruiters into 8 categories during the posting procedure.

- Distribution among job categories:

<table>
<thead>
<tr>
<th>Category</th>
<th>Freq.</th>
<th>%</th>
<th>Category</th>
<th>Freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales/Business Development</td>
<td>360</td>
<td>24</td>
<td>Marketing/Product</td>
<td>141</td>
<td>10</td>
</tr>
<tr>
<td>R&amp;D/Science</td>
<td>69</td>
<td>5</td>
<td>Production/Operations</td>
<td>127</td>
<td>9</td>
</tr>
<tr>
<td>Accounting/Finance</td>
<td>338</td>
<td>23</td>
<td>Human Resources</td>
<td>138</td>
<td>9</td>
</tr>
<tr>
<td>Logistics/Transportation</td>
<td>118</td>
<td>8</td>
<td>Information Systems</td>
<td>192</td>
<td>13</td>
</tr>
</tbody>
</table>

| Total                     | 1483  | 100|

- We keep only the "title"+"mission description" parts ("firm description" and "profile searched" are excluded).
Preprocessing of texts

- Lemmatisation and tagging
- Filtering according to grammatical category (we keep nouns, verbs and adjectives)
- Filtering terms occurring in less than 5 offers
- Vector space model ("bag of words")
Several representations are compared

Representation method
- LSA, weighting : Term Frequency
- LSA, weighting : TF-IDF
- LSA, weighting : Log Entropy
- CA

Dissimilarity measure
- Euclidian distance between documents $i$ and $i'$
- $1 - \cosine similarity between documents i$ and $i'$
Method of clustering

Clustering steps

- Computing of dissimilarity matrix from document coordinates in the latent semantic space
- Hierarchical Agglomerative Clustering until a 8 class partition
- Computation of class centroids
- K-means clustering initialized from previous centroids
Measures of agreement between two partitions

$P_1, P_2 :$ two partitions of $n$ objects with the same number of classes $k$

$N = [n_{ij}]_{i=1,..,k,j=1,..,k} :$ corresponding contingency table

Rand index

$$R = \frac{2 \sum_i \sum_j n_{ij}^2 - \sum_i n_i^2 - \sum_j n_j^2 + n^2}{n^2}, \quad 0 \leq R \leq 1$$

- Rand index is based on the number of pairs of units which belong to the same clusters. It doesn’t depend on cluster labeling.
Measures of agreement between two partitions

- Cohen’s Kappa and F-measure values are depending on clusters’ labels. To overcome label switching, we are looking for their maximum values over all label allocations.

**Cohen’s Kappa**

\[
k^{opt} = \max \left\{ \frac{1}{n} \sum_i n_{ii} - \frac{1}{n^2} \sum_i n_{i.} n_{.i} \right\}, \quad -1 \leq \kappa \leq 1
\]

**F-measure**

\[
F^{opt} = \max \left\{ \frac{2 \cdot \frac{1}{k} \sum_i \frac{n_{ii}}{n_{i.}} \cdot \frac{1}{k} \sum_i \frac{n_{ii}}{n_{.i}}}{\frac{1}{k} \sum_i \frac{n_{ii}}{n_{i.}} + \frac{1}{k} \sum_i \frac{n_{ii}}{n_{.i}}} \right\}, \quad 0 \leq F \leq 1
\]
Correlation between coordinates issued from the different methods

- LSA + TF
- LSA + TF-IDF
- LSA + Log Entropy
- LSA + TF-IDF
- LSA + Log Entropy
- CA
Clustering quality according to the method and the number of dimensions: Rand index
Clustering quality according to the method and the number of dimensions: Cohen’s Kappa
Clustering quality according to the method and the number of dimensions: F-measure

**Euclidian distance**

**Cosine similarity**
Clustering quality according to the dissimilarity function: LSA + Log Entropy
Clustering quality according to the dissimilarity function: CA
Outline

1. Introduction
2. Latent Semantic Analysis
3. Application in a real context
4. Conclusion
Conclusions

- CA seems to be less stable than other methods but with cosine similarity, it provides better results under 100 dimensions.

- As it is said in literature, cosine similarity between vectors seems to be more adapted to textual data than usual dot similarity: slight increase of efficiency and more stability for agreement measures.

- Optimal number of dimensions to keep? It is varying according to the type of text studied and the method used (around 60 dimensions with CA).

- We should prefer a dissimilarity measure which provides stable results with the number of dimensions kept (in the context of automated tasks, it’s problematic if optimal dimension is depending too much on the collection of documents).
Limitations & future work

Limitations of the study
- Clusters obtained are compared with categories chosen by recruiters, which are sometimes subjective and could explain some errors
- We are working on a very particular type of corpus: short texts, variable length, sometimes very similar but not really duplicates

Future work
- Test other clustering methods (the representation to adopt may depend on it)
- Repeat the study with a supervised algorithm for classification (index values are disappointing in unsupervised framework)
- Study the effect of using the different parts of job offers for classification
Some references


Thanks!