Application of Text Mining in Developing Standardized Descriptions of Taxa in Paleontology: A Framework

Francisca E. Oboh-Ikuenobe, Geological Science & Engineering, ikuenobe@umr.edu

Wen-Bin (Vincent) Yu, Information Science & Technology, yuwen@umr.edu

Bih-Ru Lea, Business Administration, leabi@umr.edu

University of Missouri – Rolla
Rolla, MO 65401, USA

Problems In Palynomorph Classifications And Interpretations

Description #1: Proximate, acavate cyst, large (75-95 µm) with irregular perforated paratubal crests, well developed paratabulation (including parasulcal paraplates), and precingular archaeopyle.

Description #2: Large proximate, cyst with irregular perforated crests, well developed paratabulation.

Subjective nature (human judgments, different levels of training)
Text mining or Information Retrieval?

- **Google Search/ online article database**
  - Key word search (existing knowledge)

- **Web Mining/ Semantic Web**
  - Structured ontology for Information Extraction

- **Text mining**
  - A special case of data mining
  - Discover new knowledge by finding interesting and non-trivial patterns in textural documents
  - Usually starts with clustering similar documents and extracting key terms associated with each cluster

Framework

<table>
<thead>
<tr>
<th>Data Sources</th>
<th>Text Mining</th>
<th>Data Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>DinoSys</td>
<td>Text preprocessing &amp; Pattern Identification Module</td>
<td>Model Creation Module</td>
</tr>
<tr>
<td>Sample Database</td>
<td>Variable Reduction Module</td>
<td>Model Evaluation Module</td>
</tr>
<tr>
<td>CHRONOS</td>
<td>Clustering Module</td>
<td>Model Selection Module</td>
</tr>
<tr>
<td>User Input</td>
<td>...</td>
<td>Descriptive Model</td>
</tr>
</tbody>
</table>

Data Mining/ Software Agent

Standardized Descriptions of Taxa in Paleontology

By
Vincent Yu
Francisca Oboh-Ikenobe
Bih-Ru Lea
Text Mining Process

- Text Mining is to find hidden patterns for textual documents by grouping similar descriptions...
- Portable and mobile video services could be ported to other locations, other devices or other media...
- [IPTV, WiMAX get star billing at Supercomm 2005]

Data (Documents) Cleansing

Text processing
- Parsing
- Stop words and start words
- Parts of speech
- Stemming
- Synonyms, Jargons, Abbreviations

Term Frequency Matrix

Weighting Scheme

Term Frequency Matrix II

- Represent each document with selected terms
- A note on Discriminating power: Terms with the most discriminating power are those that occur frequently, but only in a few documents.
- Determine Term weights and Frequency weights

<table>
<thead>
<tr>
<th></th>
<th>Term 1: process</th>
<th>Term 2: web</th>
<th>Term 3: voice</th>
<th>Term 4: id</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Doc 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Text Mining is to find hidden patterns for textual documents by grouping similar descriptions...

...portable and mobile video services could be ported to other locations, other devices or other media... [IPTV, WiMAX get star billing at Supercomm 2005]

- Data (Documents) Cleansing
- Text processing
  - Parsing,
  - Stop words and start words,
  - Parts of speech
  - Stemming
  - Synonyms, Jargons, Abbreviations

- Term Frequency Matrix
- Weighting Scheme

Term Weights: Entropy Method

Term weight (Gij) =

\[ G_j = 1 + \sum_j p_{ij} \log_2 \left( \frac{p_{ij}}{g_i} \right) \]

where
- \( p_{ij} = a_{ij} / g_i \)
- \( a_{ij} \): frequency that term i appears in document j.
- \( g_i \): frequency that term i appears in document collection.
- \( n \): number of documents in the collection

Weighted Frequency

\[ L_{ij} = \log_2(a_{ij}+1) \]
Text Mining is to find hidden patterns for textual documents by grouping similar descriptions. Portable and mobile video services could be ported to other locations, other devices or other media... [IPTV, WiMAX get star billing at Supercomm 2005]

- Data (Documents) Cleansing
- Text processing
  - Parsing
  - Stop words and start words
  - Parts of speech
  - Stemming
  - Synonyms, Jargons, Abbreviations
- Term Frequency Matrix
- Weighting Scheme

Text mining process III

- Variable reduction (Transformations)
  - SVD (Singular Value Decomposition)
    - Principal components decomposition.
  - Roll Up Term
    - Use the n terms with the largest term weights.

- Application in Data analysis
  - Cluster analysis
  - Classification
  - Predictions
Text Mining Process – Example 1

Text Mining is to find hidden patterns for textual documents by grouping similar descriptions. Portable and mobile video services could be ported to other locations, other devices or other media… [IPTV, WiMAX get star billing at Supercomm 2005]

Data (Documents) Cleansing
Text processing
- Parsing
- Stop words and start words
- Parts of speech
- Stemming
- Synonyms, Jargons, Abbreviations

Term Frequency Matrix

Variable Reduction & Transformation

Clustering

Framework- Text Mining Process
Clusters with SVD 1

Four different descriptions for the same dinocyst

Clusters with SVD 2
Text Mining Process – Example 2

Framework - Data Mining Process

- **Input (independent variable):**
  - the SVD variables
- **output (dependent variable):**
  - clusters
- **Descriptive models:**
  - Regression
  - Neural network
  - decision tree
- **Model Evaluation**
- **Model Selection**

---

**Data Mining**

- Model Creation Module
- Model Evaluation Module
- Model Selection Module

**Data Mining/Software Agent**

- Descriptive Model
- Predictive Model

**Standardized Taxon Recommendation Modeling Module**
Modelling: ANN

Modelling: Regression
Modelling: Decision Tree

Dinocyst Description Modeling
Model Evaluation

Framework – Descriptive Model
Scoring Dinocyst Description

- Assume regression model is selected in the model selection process
- A description is entered, analyzed by text mining process, and feed into the selected model (i.e., regression in this example)

User/Human Interaction

- the number of clusters
- the validity of key terms
- possible start/stop words list
- ...

By
Vincent Yu
Francisca
Obih-Ikenobe
Bih-Ru Lea
Dinocyst descriptions collected and mined periodically to update model

Adoptive Model

Data Sources
- DinoSys
- Sample Database
- CHRONOS
- User Input

Text Mining
- Text preprocessing & Pattern Identification Module
- Variable Reduction Module
- Clustering Module

Data Mining
- Model Creation Module
- Model Evaluation Module
- Model Selection Module
- Descriptive Model
- Predictive Model

Data Mining/Software Agent
- Standardized Taxon Recommendation Modeling Module

The End