Analysing Labour Demand and Supply Using Web Mining and Data Mining

A case study on Romanian Labour Market

Ciprian Panzaru and Claudiu Brandas

West University of Timisoara, Romania
Context

Over 500 Career Sites
Over 5,000,000 CV’s
Over 50,000 jobs
SEMM: 1.6 mil. CV’s and 16,000 Jobs

10.3 mil. Internet Users
59% Used Search Engines
52% Shop Online
Why Data Mining?

Because:
• There is a huge (virtually infinite) amount of data;
• Present technological development allow the automatic extraction of knowledge from large databases.

Allow us to:
• Classification;
• Clustering;
• Discovering the Hidden Patterns;
• Modeling;
• Build forecasting models.
Labor Market Analysis in the Big Data Era

Big Data is an important source for information about the Labour Market.

Labour Market Analysis in the Big Data context using Web Mining and Data Mining:

- Labor demand profile based on the characteristics of the employment ads;
- Labor supply profile based on the characteristics of the users (visitors) and job searchers.
Web Mining

• Web content mining

• Web structure mining

• Web usage mining
Research Framework

• Quantitative analysis;

• Based on employment ads published by the National Agency for Employment (ANOFM) via the Electronic Labour Mediation Service (SEMM);

• Volume of analysed data:
  
  ✓ 16,827 job ads (representing labour demand);
  
  ✓ 18,511 CVs (representing labour supply);

• Web content mining for extraction and structuring of data (using import.io);

• Data mining clustering techniques (Berkhin 2006) using the simple k-means algorithm (using WEKA);

• Data Spatialisation in GIS (using Google Table Fusion).
DATA SOURCES
Romanian Public Employment Service
Electronic Service for Mediation of Labour Demand and Supply

PROCESS 1
Web Content Mining using Import.io
Output 1: structured data

PROCESS 2
Data Mining using WEKA
Output 2: final data and results

PROCESS 3
Data Spatialisation in GIS using Fusion Tables and Google Maps
Output 3: spatialisation of the results

BENEFICIARIES
Government, public administration, Public Employment Services, individuals
Web Content Mining – Process 1

Web Content Mining process using **Import.io**

- **job_demand_data_set** table with the following attributes (**location**, **job_demand_description**): table contain data about job demand.

- **job_supply_data_set** table with the following attributes (**location**, **job_supply_description**): table contain data about job supply.
# Web Content Mining – Process 1

Web Content Mining process using **Import.io**

![Import.io screenshot](image.png)

<table>
<thead>
<tr>
<th>Numar total oferte</th>
<th>5171</th>
<th>Total locuri vacante</th>
<th>16827</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Cod</th>
<th>Jude</th>
<th>Judeanca</th>
<th>Companie</th>
<th>1600,000,000</th>
<th>1600,000,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
</tr>
<tr>
<td>2</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
</tr>
<tr>
<td>3</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
</tr>
<tr>
<td>4</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
</tr>
<tr>
<td>5</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
<td>AEA</td>
</tr>
</tbody>
</table>

**Note:**
- The screenshot shows a webpage with data and a table.
- The table contains columns for code, name, company, and other relevant information.
- The process involves using Import.io to extract data from the webpage.

**Steps:**
1. **Data Scraper:** Import.io is used to scrape data from the webpage.
2. **Data Validation:** The data is validated for accuracy.
3. **Data Analysis:** The cleaned data is analyzed for insights.

**Tools Used:**
- Import.io
- Data scraping software
- Data validation tools

**Outcome:**
- A comprehensive dataset for further analysis.
- Insights derived from the analyzed data.
For `job_demand_data_set`:

- six clusters with one instance for each cluster representing the cluster centroid (mean vectors for each cluster).
- numbers and percentage of instances for each cluster.
For the job_supply_data_set:

- six clusters with one instance for each cluster
- numbers and percentage of instances for each cluster.
Results – Process 2

Job demand density
Results – Process 2

Job supply density
Spatialisation – Process 3

Job demand density

Job supply density
Conclusions

In Cluster #0, we have the most instances (64%), followed by the Cluster #5 (15%). In Cluster #4 we have the fewest instances (2%). For the representative Cluster #0 we have the cluster centroid formed by location "CLUJ" and job_demand_description "confectioner-asamblor articole din textile". For the least representative Cluster #4 we have the cluster centroid formed by location "BRASOV" and job_demand_description "montator subansamble".

<table>
<thead>
<tr>
<th>Job Demand</th>
<th>Job Supply</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Cluster #0, we have the</td>
<td>In Cluster #0, we have the</td>
</tr>
<tr>
<td>most instances (67%)</td>
<td>most instances (67%)</td>
</tr>
<tr>
<td>followed by Cluster #5</td>
<td>followed by Cluster #5</td>
</tr>
<tr>
<td>(11%) and Cluster #3</td>
<td>with 11% and Cluster #3</td>
</tr>
<tr>
<td>with 10%.</td>
<td>with 10%.</td>
</tr>
<tr>
<td>In Cluster #1 and Cluster</td>
<td>In Cluster #1 and Cluster</td>
</tr>
<tr>
<td>#4 we have the fewest</td>
<td>#4 we have the fewest</td>
</tr>
<tr>
<td>instances (3%)</td>
<td>instances (3%)</td>
</tr>
<tr>
<td>For the representative</td>
<td>For the representative</td>
</tr>
<tr>
<td>Cluster #0 we have the</td>
<td>Cluster #0 we have the</td>
</tr>
<tr>
<td>cluster centroid formed by</td>
<td>cluster centroid formed by</td>
</tr>
<tr>
<td>location &quot;GALATI&quot; and</td>
<td>location &quot;GALATI&quot; and</td>
</tr>
<tr>
<td>job_supply_description</td>
<td>job_supply_description</td>
</tr>
<tr>
<td>&quot;muncitori necalificati in</td>
<td>&quot;muncitori necalificati in</td>
</tr>
<tr>
<td>ferme mixte&quot;.</td>
<td>ferme mixte&quot;.</td>
</tr>
<tr>
<td>For the least representative</td>
<td>For the least representative</td>
</tr>
<tr>
<td>Cluster #1 we have the</td>
<td>Cluster #1 we have the</td>
</tr>
<tr>
<td>cluster centroid formed by</td>
<td>cluster centroid formed by</td>
</tr>
<tr>
<td>location &quot;COVASNA&quot; and</td>
<td>location &quot;COVASNA&quot; and</td>
</tr>
<tr>
<td>job_supply_description</td>
<td>job_supply_description</td>
</tr>
<tr>
<td>&quot;dulgheri si tamplari&quot;.</td>
<td>&quot;dulgheri si tamplari&quot;.</td>
</tr>
</tbody>
</table>
Conclusions

The Internet based technologies has generated in recent years huge volumes of diverse data both as source and as content. In the particular case of this research, web mining and data mining have helped to determine clusters for types of careers. These results can be further used to identify decision trees for modeling career paths.

We tried to show that web mining and data mining techniques can be used to obtain extremely valuable information on the labor market.
Advantages of using Data Mining in Analyzing Labor Market

Data mining solves a widespread paradox in classical statistical analysis: the more data there is to analyse, the more difficult it is to analyse and retrieve information.

If the traditional methods only modelling data (looking for a model, function), data mining involves the discovery of intelligible patterns expressed as trees, rules and data associations.

Data Mining is a technique by which huge amounts of data can be interpreted in such a way as to be easily read and understood.
Problems

- Lack of structural and semantic compatibility of data occurs frequently;
- Diversity of sources of information and attribute values.
Recommendations

• Using: Ontologies, Web Semantics and Web 3.0
Thank you!