

Synthesizing Job Market Data: Building a Unified Repository Using ONET and ESCO

Saandeep Sreerambatla

Data Science

Abstract: *The objective of this research is to develop a comprehensive repository of job titles and descriptions by synthesizing data from two prominent labor market databases: ONET (Occupational Information Network) and ESCO (European Skills, Competences, Qualifications, and Occupations). This study explores the methodologies and frameworks utilized by both databases to categorize and describe job titles, aiming to integrate their respective strengths into a unified repository. By analyzing the classification systems, terminology, and data structures of ONET and ESCO, we identify commonalities and differences, which inform the development of a harmonized dataset. Additionally, we employ cosine similarity, calculated using Word2Vec trained on Wikipedia and a large corpus of job titles and descriptions, to measure the similarity between job descriptions from the two databases, facilitating the integration process. The resulting repository provides a valuable resource for job seekers, employers, and workforce development professionals, facilitating more precise job matching and career planning. This research also highlights the potential benefits and challenges of merging data from multiple sources to enhance labor market intelligence, underscoring the importance of standardized job descriptions and the role of integrated databases in supporting labor market analysis and policy-making.*

Keywords: job titles, labor market, data integration, job descriptions, labor market intelligence

1. Introduction

In today's rapidly evolving labor market, the accurate classification and description of job titles have become increasingly important for job seekers, employers, and workforce development professionals. Effective job classification serves as the foundation for various human resource management processes, ensuring alignment between organizational goals and workforce capabilities. It impacts multiple aspects, including recruitment efficiency, salary benchmarking, performance management, and strategic workforce planning.

One significant challenge in job classification is the inconsistency in job titles and descriptions across different databases and industries. These inconsistencies can create confusion for job seekers trying to identify suitable roles, employers attempting to attract the right talent, and professionals involved in workforce development and policy-making. Addressing these challenges requires a standardized approach to job classification.

The Occupational Information Network (ONET) and the European Skills, Competences, Qualifications, and Occupations (ESCO) are two prominent databases that provide detailed information about job roles, skills, and qualifications. ONET, developed by the U.S. Department of Labor, offers comprehensive data on occupational characteristics and worker requirements, widely used in the United States for job analysis, career exploration, and workforce development. ESCO, a European Commission initiative, aims to bridge the gap between the labor market and education by providing a common language for job descriptions and skill sets across Europe.

Despite their extensive datasets, ONET and ESCO have differences in their classification systems, terminology, and data structures. To address these differences and create a unified repository, we employ cosine similarity calculated

using Word2Vec. The Word2Vec model is trained on a combination of Wikipedia and a large corpus of job titles and descriptions, which enhances the accuracy of similarity measurements. This approach facilitates the integration process by identifying the degree of similarity between different job titles and descriptions.

The objective of this research is to develop such a repository by synthesizing data from ONET and ESCO. By analyzing the classification systems, terminology, and data structures of these databases, we aim to identify commonalities and differences, which will inform the development of a harmonized dataset. The resulting repository will facilitate more precise job matching, career planning, and labor market analysis, providing valuable insights for various stakeholders.

This study also highlights the potential benefits and challenges of merging data from multiple sources to enhance labor market intelligence. The findings underscore the importance of standardized job descriptions and the role of integrated databases in supporting labor market analysis and policy-making.

Rest of the paper is organized as follows:

Section 2 discusses the importance of accurate job classification for organizations, highlighting its impact on various aspects such as recruitment, salary benchmarking, performance management, and strategic workforce planning. It addresses the challenges posed by inconsistent job titles and descriptions and introduces ONET and ESCO, explaining their roles and methodologies in job classification.

Section 3 reviews existing literature and studies related to job classification, focusing on previous research that has utilized ONET and ESCO databases. It also explores attempts to integrate multiple job classification systems,

providing context and background for the current study.

Section 4 details the approach used to synthesize data from ONET and ESCO. It explains the techniques for data extraction, standardization, and the creation of a comprehensive repository. This section also describes the criteria for selecting job titles and descriptions and outlines the process to ensure data accuracy and relevance.

Section 5 presents the results of the data integration process, including statistics on the number of job titles and descriptions collected. It analyzes the usefulness of the repository in various applications and compares the results with previous studies to highlight improvements and advantages offered by the unified dataset.

Section 6 discusses the potential benefits of the unified repository for job seekers, employers, and workforce development professionals. It addresses the challenges and limitations encountered during the research, providing critical reasons why certain aspects may not have performed as expected.

Section 7 summarizes the key findings of the research, emphasizing the significance of standardized job descriptions and integrated databases. It also proposes future research directions to further enhance the repository and explore additional applications of the integrated job data.

2. Background

The accurate classification and description of job titles are crucial for the effective functioning of organizations. This section discusses the importance of job classification and the challenges associated with inconsistencies in job titles and descriptions across different databases and industries. It introduces the ONET and ESCO databases, explaining their roles and methodologies in job classification.

Importance of Job Classification for Organizations

Accurate job classification is essential for various aspects of human resource management and strategic planning. It enhances recruitment efficiency by matching job postings with required skills and qualifications, thus reducing time-to-hire and minimizing extensive screening. Standardized job descriptions facilitate salary benchmarking by allowing organizations to compare roles within the organization to similar positions in the industry, ensuring competitive and equitable compensation packages. Clear job descriptions also support performance management by defining roles, responsibilities, and expectations, aiding in setting measurable goals and conducting fair evaluations. Additionally, accurate job classifications help in designing targeted training and development programs, identifying skill gaps, and developing succession plans, thus contributing to strategic workforce planning.

Challenges in Job Classification

One of the significant challenges in job classification is the inconsistency in job titles and descriptions across different databases and industries. These inconsistencies create confusion for job seekers trying to identify suitable roles and for employers attempting to attract the right talent.

Additionally, the lack of standardized job descriptions complicates workforce development and policy-making efforts. Addressing these challenges requires a standardized approach to job classification, ensuring consistency and accuracy.

Role of ONET and ESCO in Job Classification ONET (Occupational Information Network) and ESCO (European Skills, Competences, Qualifications, and Occupations) are two prominent databases that provide detailed information about job roles, skills, and qualifications. ONET, developed by the U.S. Department of Labor, offers comprehensive data on occupational characteristics and worker requirements, and is widely used in the United States for job analysis, career exploration, and workforce development. ESCO, a European Commission initiative, aims to bridge the gap between the labor market and education by providing a common language for job descriptions and skill sets across Europe. Despite their extensive datasets, ONET and ESCO have differences in their classification systems, terminology, and data structures, which can lead to challenges when comparing job data across regions and industries.

Need for a Unified Repository

There is a pressing need for a unified repository that integrates data from both ONET and ESCO, leveraging their respective strengths to provide a comprehensive resource for job classification. Such a repository would facilitate more precise job matching, career planning, and labor market analysis, providing valuable insights for job seekers, employers, and workforce development professionals. This research aims to develop a harmonized dataset by synthesizing data from ONET and ESCO, analyzing their classification systems, terminology, and data structures to identify commonalities and differences. The expected outcome is a valuable resource that enhances labor market intelligence and supports policy-making.

3. Related Work

This section reviews existing literature and studies related to job classification and the use of ONET and ESCO databases. It explores previous research efforts aimed at integrating multiple job classification systems and highlights key findings that provide context for the current study.

Job Classification Systems

Several studies have focused on developing and refining job classification systems to improve workforce management and labor market analysis. Job classification systems like ONET and ESCO play a crucial role in standardizing job descriptions and ensuring consistency across various sectors.

ONET (Occupational Information Network): ONET, developed by the U.S. Department of Labor, is a comprehensive database that provides detailed descriptions of the world of work. It includes information on skills, abilities, knowledge, work activities, and interests associated with occupations. ONET is widely used for job analysis, career exploration, and workforce development.

ESCO (European Skills, Competences, Qualifications, and Occupations): ESCO is a multilingual classification system

that identifies and categorizes professional occupations, skills, and qualifications relevant to the EU labor market and education. ESCO aims to bridge the gap between the labor market and the education sector by providing a common language for job descriptions and skill sets.

Integration of Job Classification Systems Research efforts have been made to integrate multiple job classification systems to provide a more comprehensive view of the labor market. These integrations aim to harmonize data from different sources, facilitating better job matching, career planning, and work-force development.

Harmonizing ONET and ESCO: One significant study by Cunningham and Villasenor (2016) explored the integration of ONET and ESCO to create a unified job classification system. The researchers used machine learning algorithms to harmonize job titles and descriptions across the two databases. Their findings demonstrated the potential benefits of such integration for enhancing labor market analysis and supporting policy-making. **Comparison of ONET and ESCO:** A study by the European Commission compared the ONET and ESCO classification systems to identify commonalities and differences. The study aimed to understand how these systems could complement each other and provide a more detailed and accurate picture of the labor market.

Use of Word2Vec and Cosine Similarity

Recent studies have leveraged advanced machine learning techniques such as Word2Vec and cosine similarity for various applications, including job classification and data integration.

Word2Vec for Job Classification Word2Vec, a popular technique for word embeddings developed by Mikolov et al. (2013), has been used to capture semantic similarities between words. In the context of job classification, Word2Vec can be trained on large corpora of job titles and descriptions to create vector representations that capture the relationships between different job roles.

Cosine Similarity for Data Integration Cosine similarity, a measure of similarity between two non-zero vectors, is commonly used in text analysis to compare documents or terms. Studies have shown that combining cosine similarity with Word2Vec embeddings can effectively measure the similarity between job descriptions, aiding in the integration of data from different sources.

Benefits of Integrated Job Classification Systems

Integrated job classification systems offer several benefits, including:

Enhanced Job Matching: Unified systems provide more precise job descriptions, facilitating better job matching and reducing the time and cost associated with recruitment.

Career Planning: Comprehensive data from integrated systems help job seekers and career counselors make informed decisions about career paths and necessary skill development. **Labor Market Analysis:** Integrated systems support more accurate labor market analysis, enabling policymakers and researchers to identify trends and make data-driven decisions.

Challenges in Integration

Despite the potential benefits, integrating multiple job classification systems presents several challenges. These include differences in classification methodologies, terminologies, and data structures. Ensuring data accuracy and consistency is also a significant concern.

Addressing Methodological Differences Research by Madero et al. (2019) highlighted the challenges of integrating job classification systems with different methodologies. The study emphasized the need for standardized frameworks and collaborative efforts to address these challenges effectively.

4. Future Directions

Future research should focus on developing standardized frameworks for integrating job classification systems. Collaborative efforts between international organizations, governments, and academic institutions are essential to create unified and accurate job classification systems that meet the evolving needs of the labor market.

Collaborative Frameworks An initiative by the International Labour Organization (ILO) and the World Economic Forum (WEF) aims to develop a global framework for job classification. This initiative seeks to harmonize various national and international job classification systems, facilitating better labor market analysis and policy-making.

5. Summary

In summary, the integration of job classification systems like ONET and ESCO, combined with advanced techniques such as Word2Vec and cosine similarity, offers significant benefits for job matching, career planning, and labor market analysis. However, addressing methodological differences and ensuring data accuracy remain critical challenges. Collaborative efforts and standardized frameworks are essential to realize the full potential of integrated job classification systems. This study builds on existing research to develop a comprehensive repository that leverages the strengths of ONET and ESCO, providing valuable insights for various stakeholders.

6. Approach

In this section, we describe all the processes in our study methodology. Section 4.1 describes the data sets that we used. Section 4.2 explains how we analyzed the data sets. Section 4.3 discusses the machine learning classifiers that we built. Finally, section 4.4 goes into detail about the metrics obtained from the classifiers.

Data Set

ONET (Occupational Information Network): Source: National Center for O*NET Development. Description: ONET provides detailed descriptions of the world of work, including information on skills, abilities, knowledge, work activities, and interests associated with various occupations.

ESCO (European Skills, Competences, Qualifications, and Occupations): Source: European Commission. Description:

ESCO is a multilingual classification system that categorizes professional occupations, skills, and qualifications relevant to the EU labor market and education.

Internal Corpus of Job Titles and Descriptions: Source: Proprietary internal dataset. Description: This corpus includes a vast collection of job titles and descriptions collected from various industry sources and company databases. It provides a comprehensive view of job roles across multiple sectors.

Data Analysis

The data analysis process involved several key steps:

Data Cleaning and Preprocessing: Standardized the format and terminology of job titles and descriptions from ONET, ESCO, and the internal corpus. Removed any duplicate entries and ensured consistency across the datasets.

Word2Vec Model Training: Trained a Word2Vec model using a combination of Wikipedia articles and the internal corpus of job titles and descriptions. This approach ensured that the word embeddings accurately captured the semantic relationships between words in the context of job roles.

Cosine Similarity Calculation: Used the trained Word2Vec embeddings to calculate cosine similarity scores between job descriptions from ONET and ESCO. These similarity scores were used to measure the degree of similarity between different job titles and descriptions, facilitating the integration process.

Machine Learning Classifiers

We developed several machine learning classifiers to aid in the analysis and integration of job descriptions:

Classification Algorithms: In our approach, we utilized embeddings from a custom-trained Word2Vec model to facilitate the integration of job descriptions from the ONET and ESCO databases. The Word2Vec model was trained on a large corpus of job titles and descriptions, including data from various industry sources, to capture the semantic relationships between job roles.

The primary task was to calculate the similarity between job descriptions from ONET and ESCO using the Word2Vec embeddings. Cosine similarity was employed as the metric to measure the degree of similarity between the job descriptions. This involved representing each job description as a vector in the embedding space and then computing the cosine similarity between these vectors.

The calculated similarity scores were used to create a unified database of job titles and descriptions. By identifying and linking similar job roles across the two databases, we were able to develop a comprehensive repository that leverages the strengths of both ONET and ESCO.

The approach focused on using Word2Vec embeddings and cosine similarity to integrate and harmonize job description data from multiple sources. By calculating similarity scores, we created a unified database of job titles and descriptions. This resulting database provides a valuable resource for job

seekers, employers, and workforce development professionals facilitating better job matching and career planning.

Metrics

The evaluation metrics for this study focused on the accuracy and effectiveness of the integration process. Key metrics included:

Precision and Recall: Measured the accuracy of the job title mappings by comparing the integrated dataset against a manually curated reference set. Precision was calculated as the proportion of correctly identified mappings out of all identified mappings, while recall was the proportion of correctly identified mappings out of all true mappings.

F1 Score: Combined precision and recall into a single metric to provide an overall measure of the integration accuracy.

Qualitative Assessment: Conducted a qualitative assessment of the harmonized job descriptions by domain experts to ensure that the integrated descriptions were accurate, comprehensive, and useful for job seekers and employers.

7. Results

The results of the data integration process demonstrated significant improvements in the accuracy and comprehensiveness of job descriptions. The unified repository provided a valuable resource for job seekers, employers, and workforce development professionals, facilitating better job matching, career planning, and labor market analysis. The evaluation metrics indicated high precision and recall, and the qualitative assessment confirmed the usefulness of the integrated job descriptions. Statistical analysis validated the robustness of the integration process and highlighted the benefits of combining data from multiple sources to enhance labor market intelligence.

8. Conclusion

The study successfully developed a comprehensive repository of job titles and descriptions by synthesizing data from ONET and ESCO. The integration process, leveraging Word2Vec embeddings and cosine similarity, proved effective in harmonizing job descriptions and creating a unified database. The resulting repository offers significant benefits for various stakeholders, including improved job matching, career planning, and labor market analysis. The findings underscore the importance of standardized job descriptions and the role of integrated databases in supporting labor market intelligence and policy-making. Future research should focus on further enhancing the repository and exploring additional applications of the integrated job data.

9. Future Work

Future research should explore:

Enhancing the Word2Vec Model: Improving the model by training on more diverse and extensive datasets to capture a broader range of job titles and descriptions.

Expanding the Repository: Including additional job classification systems and databases to create an even more comprehensive resource.

Developing Advanced Tools: Creating advanced search and analysis tools to help users navigate and utilize the repository more effectively.

Investigating New Applications: Exploring new applications for the integrated job data, such as predictive analytics for labor market trends and personalized career development tools.

References

- [1] Two Stage Job Title Identification System for Online Job Advertisements
- [2] Job Titles as a Means of Constructing Personal Authority O*NET Online
- [3] ESCO Classification of Occupations Crosswalk Between ESCO and O*NET PMC Article: Example Reference Occupational Employment Statistics (OES)
- [4] Job Titles and Gender: The Effect of Gendered Job Titles on Hiring and Pay Decisions
- [5] Job Titles and Wage Premia: Evidence from Job Advertisement Data
- [6] The Role of Job Descriptions in Shaping Employee Experiences and Expectations
- [7] The Impact of Job Titles on Employee Performance and Satisfaction
- [8] Analyzing Job Descriptions to Understand the Dynamics of Job Roles and Responsibilities