CLUSTERING IN ONTOLOGY-BASED ANALYSIS OF RESEARCH PROJECT DESCRIPTIONS

P. Lula[1,a], J. Tuchowski[1], U. Cieraszewska[1], M. Talaga [1]

[1] *Cracow University of Economics, Poland*

E-mail: [a]pawel.lula@uek.krakow.pl

Ontology-based approach in exploratory analysis of textual data can significantly improve the quality of the obtained results. On the other hand, the use of domain knowledge defined in the form of ontologies increases the time needed to prepare a model and makes required calculations more complex. The publication will discuss selected aspects of cluster analysis performed on documents automatically annotated using ontologies. It seems that methodological aspects of cluster analysis process, especially the way in which distances are determined, should depend on the structure of a given ontology. Three cases involving the use of ontologies with linear, hierarchical and network structures will be discussed. The methodological aspects of ontology-based cluster analysis of text documents was used for analysis of projects' descriptions related to the area of economics and registered in the period 2019-2021. Only Horizon and Framework Program projects were included.

Keywords: scientific productivity, ontology-based cluster analysis, CORDIS, JEL

Paweł Lula, Janusz Tuchowski, Urszula Cieraszewska, Magdalena Talaga

1. Introduction

Scientific text annotation has become an important task for scientists. There is an increasing need for the development of intelligent systems to support new scientific findings [1]. Currently, ontologies are viewed as a shared and common understanding of a domain that can be communicated between people and heterogeneous and distributed application systems [2]. Public databases available on the Web provide useful data. Text annotation may help as it relies on the use of ontologies to maintain annotations based on a uniform vocabulary.

Clustering text documents into different category groups is an important step in indexing, retrieval, management and mining of abundant text data on the Web or in corporate information systems. Among others, the challenging problems of text clustering are big volume, high dimensionality and complex semantics [3].

Nowadays, there is an increasing need for decision support systems to guide the investments on new scientific research projects. They need to extract useful information from many different resources. One such resource that allows you to see in which directions research ideas are developing is the Community Research and Development Information Service (CORDIS) [4] which is the European Commission's primary source of results from the projects funded by the EU's framework programmes for research and innovation. It has a rich and structured public repository with all project information.

2. Methodological aspects of ontology-based cluster analysis of text documents

The main assumption which was made by the authors is that the cluster analysis of documents is supported by domain knowledge represented by an ontology. Starting from this assumption the following stages in the analysis process can be defined:

- corpus preparation,
- ontology building (or ontology selection),
- documents' annotation,
- distance matrix calculation,
- conducting a clustering process.

In the corpus preparation phase all documents were transformed to pure text format coded in UTF-8 format. Next all words were transformed to their base form (lemmatization process). Also, numeric values and punctuation marks were omitted.

Providing of a proper ontology is the main goal of the next step. It seems that the adoption of a widely accepted ontology is better than building a new ontology designed exclusively for a given research process.

Next, an annotation process should be conducted. During this stage concepts from a given ontology should be assigned to words or phrases in documents. There are many techniques which can be used for implementing annotation task, but rule-based technique is the most popular.

Ontology classification into three classes (with linear, hierarchical or network structure) is very important from the perspective of cluster analysis because ontology's type determines the way of distance calculation.

For ontologies with linear character (gazetteers) it is only possible to check if two concepts are the same or not. For ontologies having hierarchical structure there are two popular approaches used to calculate distances between concepts. First based on the length of a path connecting to concepts. And the second which is based the information theory. It seems that for ontologies with network structure,

the most convenient way of distance calculation is based on the length of path between nodes representing two concepts.

Having a measure between concepts defined, the ontology-based similarity measure between two documents should be specified. Let's assume that the set D_i contains all concepts occurring in a i-th document. Then a similarity between two documents can be defined depending on the type of a given ontology.

For linear ontologies distances between documents can be calculated as:

- Jaccard distance:

$$dist(D_1, D_2) = 1 - \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

- Hamming distance:

$$dist(D_1, D_2) = |D_1 \oplus D_2|$$

While for hierarchical or network-based ontologies the following formulas for document similarity can be used:

- average distance between all concepts:

$$sim(D_1, D_2) = avg(c_i, c_j), c_i \in D_1, c_j \in D_2$$

- average distance between the nearest concepts:

$$sim(D_1, D_2) = \frac{\sum_{i=1}^{N} \min_j \left(sim(c_i, c_j)\right) + \sum_{j=1}^{M} \min_i \left(sim(c_i, c_j)\right)}{N + M}$$

- average distance between concepts chosen as a solution of the optimal alignment problem defined as:

$$sim(D_1, D_2) = \arg\min_{c_i, c_j} \sum sim(c_i, c_j), c_i \in D_1, c_j \in D_2$$

Formulas presented above allow to define similarity matrix between documents. This matrix is a starting data for distance-based cluster analysis. The authors decided to use hierarchical, agglomerative approach, mostly Ward's method.

3. Analysis of Horizon and Framework Program projects related to the area of economics registered in the CORDIS database

The methodology presented in the previous section was used for analysis of projects' descriptions related to the area of economics and registered in the period 2019-2021. Only Horizon and Framework Program projects were included. The total number of projects was 292.

All documents were annotated with the use of the JEL ontology [5]. During this process all concepts defined in the ontology were identified. Next, the significance of main concepts was evaluated by calculating an average number of occurrences for concepts belonging to every main class. The results are presented in Figure 25.

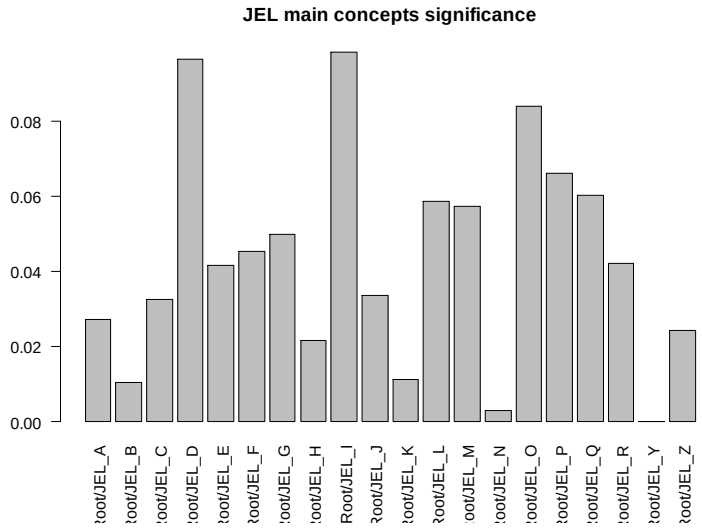

Figure 25. JEL main concepts significance for the whole corpus of project descriptions

For annotated documents cluster analysis may be performed with the use of Hamming distance and Ward's method. The results are presented on Figure 26.

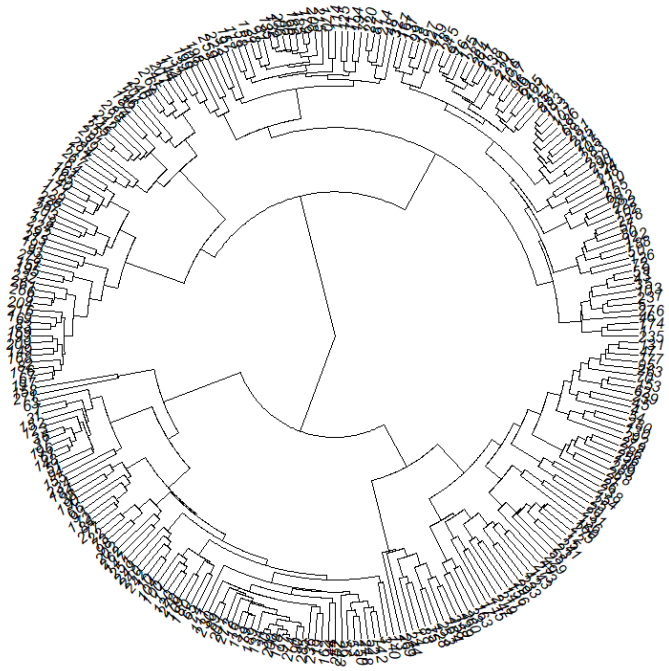

Figure 26. Dendrogram presenting the structure of project descriptions

The shape of the dendrogram suggest that the division of descriptions into two groups. The evaluation of clustering process quality based on silhouette coefficients shows that the structure of clusters is rather weak. It means that clusters are overlapping.

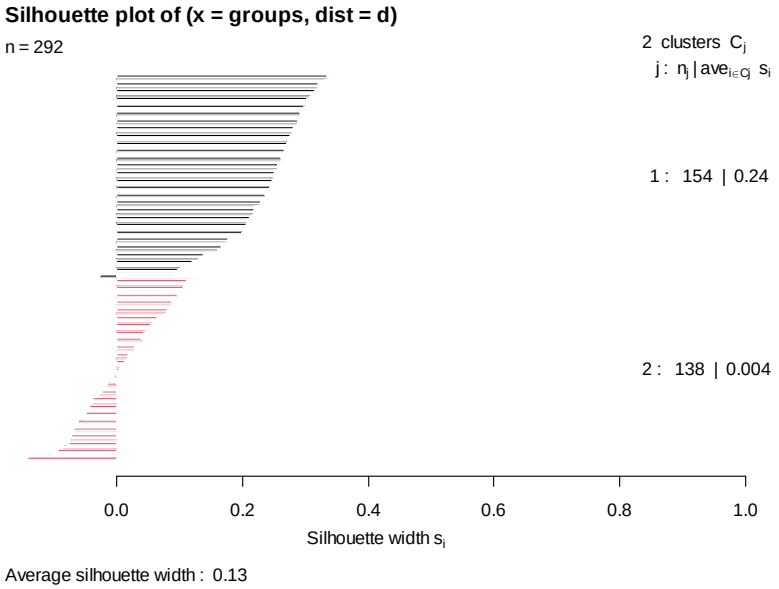

Figure 27. Silhouette plot presenting the quality of project descriptions division into two groups

For more than two clusters an average value of the silhouette index was smaller and therefore further analysis was performed for two groups of projects Figure 27. For every group the significance of JEL main concepts was estimated. The results are presented on Figure 28.

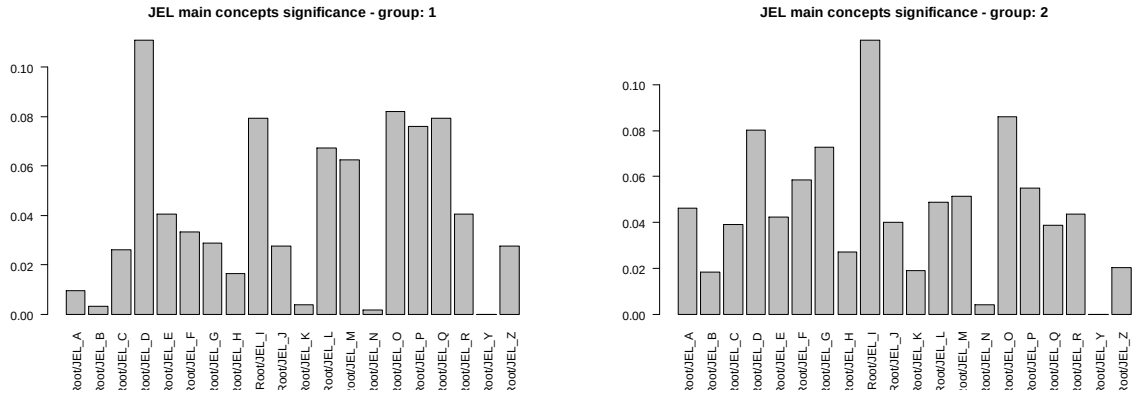

Figure 28. JEL main concepts significance for two clusters of project descriptions

The analysis of descriptions assigned to every group confirms previous observation regarding high similarity between clusters.

4. Conclusions

The results obtained during the analysis show that:

- ontology-based approach allows to perform the analysis of project descriptions to identify concepts related to a given research domain,
- Hamming distance and Ward's method can be used for cluster analysis of documents annotated with automatically identified ontology concepts,
- silhouette coefficients inform about the quality of document clusters identified by cluster analysis methods.

The authors are going to develop the system presented here by adding modules performing concepts' identifications defined in other domain ontologies (MeSH or CSO). Also the analysis of relationships between concepts derived from more than one ontology will be ensured in future solutions.

References

[1] P. C. e C. Gomes, A. M. de C. Moura, and M. C. Cavalcanti, 'A multi-ontology approach to annotate scientific documents based on a modularization technique', J. Biomed. Inform., vol. 58, pp. 208–219, Dec. 2015, doi: 10.1016/j.jbi.2015.09.022.

[2] R. García, 'A Semantic Web Approach to Digital Rights Management', 2006.

[3] L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang, 'Ontology-based Distance Measure for Text Clustering', 2006.

[4] 'European Commission : CORDIS : Search : Results page'. https://cordis.europa.eu/projects/en (accessed Sep. 13, 2021).

[5] 'Journal of Economic Literature'. [Online]. Available: https://www.aeaweb.org/econlit/jelCodes.php?view=jel