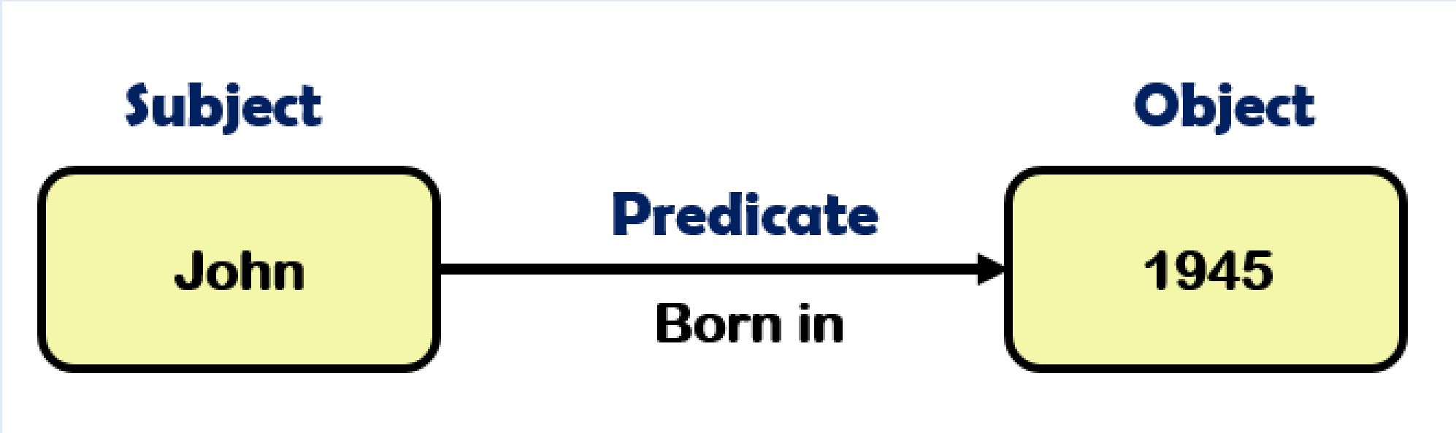
RDF Data Clustering Framework

Presented by

Suraiya Nusrat Tanha, Prinom Maojumder, Nur Nahar Mim, Tasnim Israk Synthia

Background

- Helps in organizing and retrieving knowledge from linked data.
- Combines embedding techniques with clustering to improve RDF data analysis.



Objective

- Develop an RDF data clustering framework.
- Compare clustering on existing and constructed knowledge graphs.
- Improve entity relationship understanding through embedding.

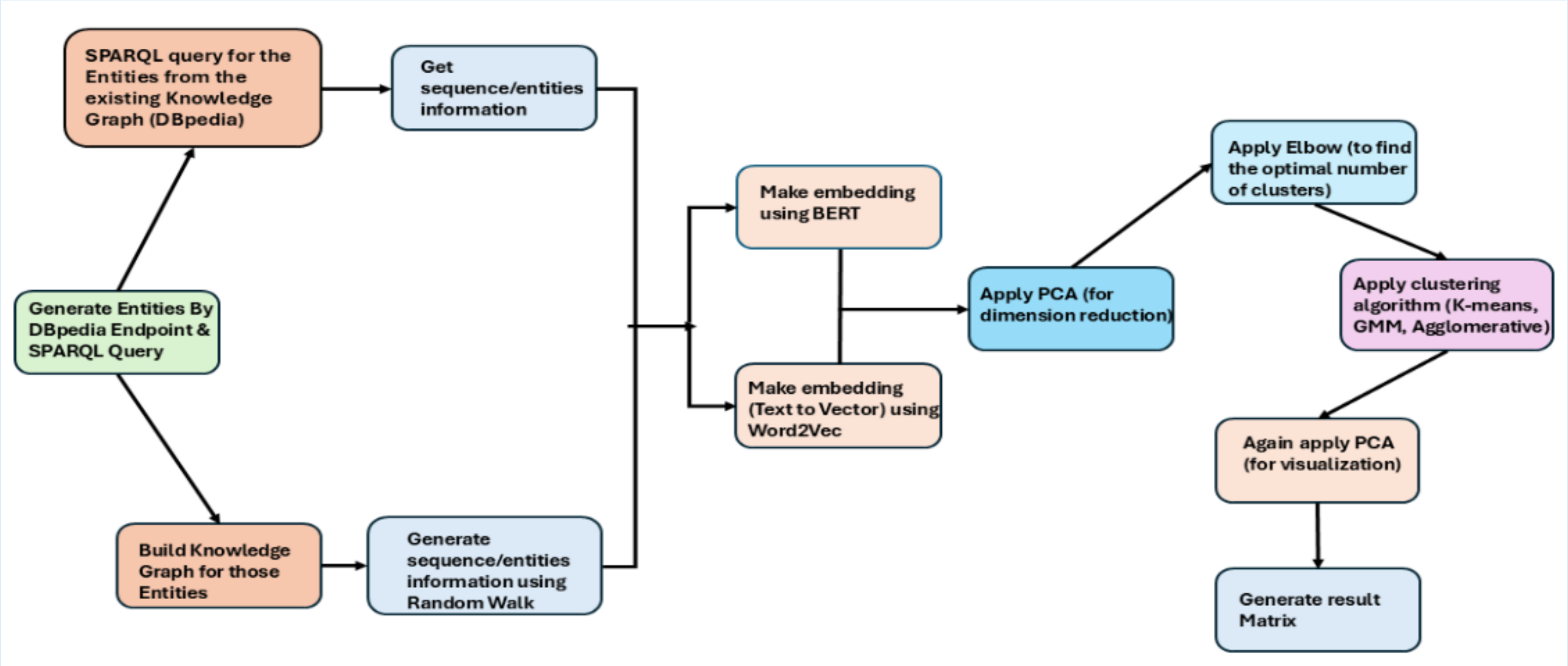
Research Questions

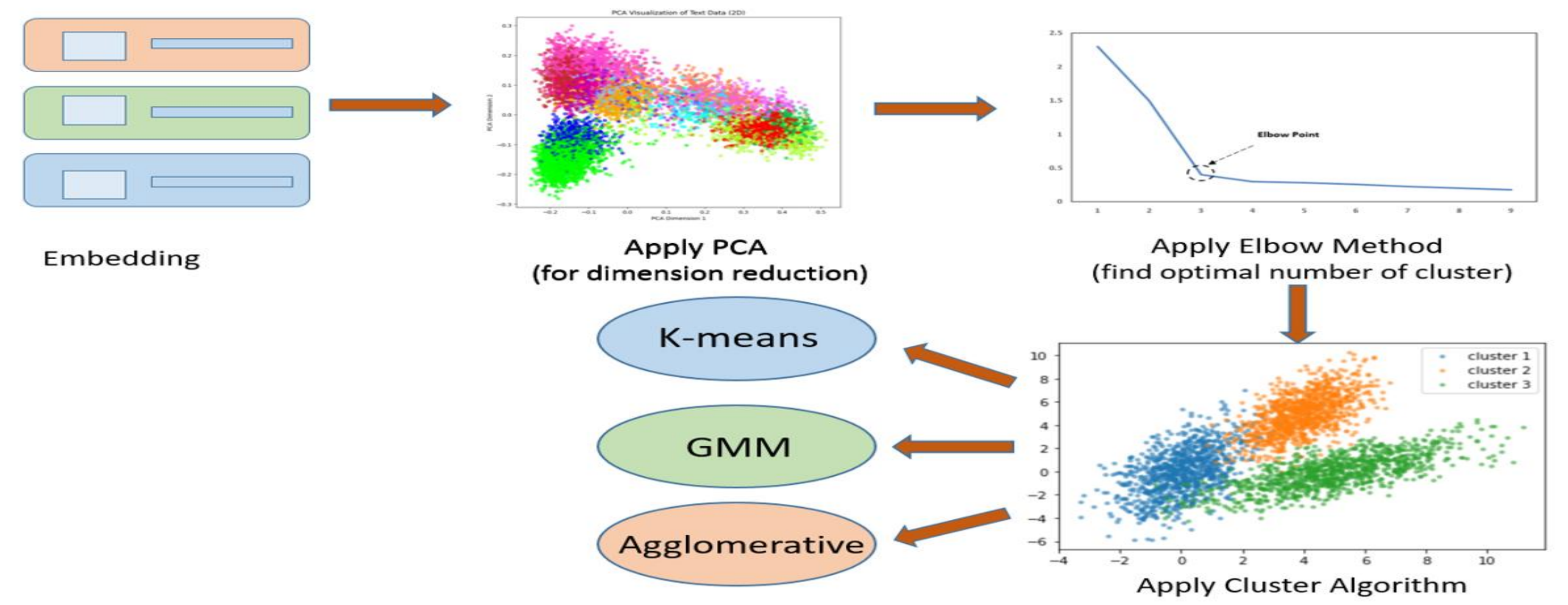
- How do different embedding techniques affect the clustering quality of RDF data?
- Which clustering algorithm performs best for RDF data based on evaluation metrics such as ARI, NMI, and Silhouette Score?
- Can embedding-based clustering improve the interpretability and retrieval efficiency of RDF datasets and find the semantic relationship?

Methodology

Our methodology follows two distinct approaches.

- The first approach extracts entities from DBpedia using SPARQL queries to build a new knowledge graph (KG). It generates entity relationships through Random Walk, applies Word2Vec and BERT embeddings, reduces dimensions with PCA, and clusters using K-Means, GMM, and Agglomerative Clustering.
- The second approach directly uses an existing KG, extracting entities and relationships for further processing. Both approaches involve embedding generation, dimensionality reduction, clustering, and visualization to analyze entity relationships effectively.


SPARQL query for the Entities from the existing Knowledge Graph (DBpedia)
Get sequence/entities information
Generate Entities By DBpedia Endpoint & SPARQL Query
Build Knowledge Graph for those Entities
Generate sequence/entities information using Random Walk
Make embedding using BERT
Make embedding (Text to Vector) using Word2Vec
Apply PCA (for dimension reduction)
Apply Elbow (to find the optimal number of clusters)
Apply clustering algorithm (K-means, GMM, Agglomerative)
Again apply PCA (for visualization)
Generate result Matrix



Result

BERT-based Embeddings:

- Existing KG: K-Means and GMM similar, Agglomerative Clustering best (**ARI: 0.970252**).
- Constructed KG: K-Means outperformed GMM in ARI (0.660588), Agglomerative Clustering slightly better in completeness.

Word2Vec-based Embeddings:

- Existing KG: Moderate performance; K-Means worse than GMM and Agglomerative.
- Constructed KG: All methods underperformed; Agglomerative best (ARI: 0.150356).

2D/3D Visualization:

- Existing KG: **BERT better** separation of clusters.
- Constructed KG: BERT clusters compact, Word2Vec shows overlap.

Comparison between Existing KG & Constructed KG

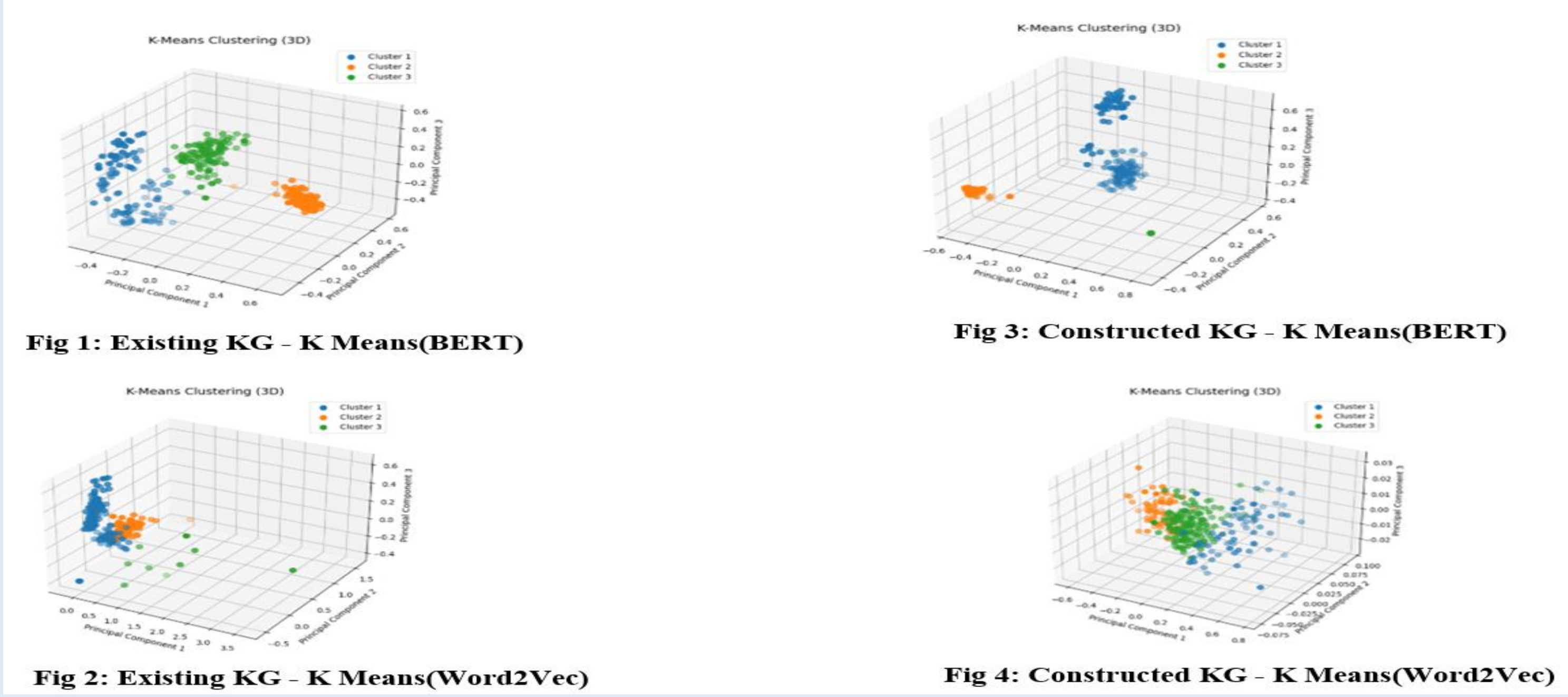


Fig 1: Existing KG - K Means(BERT)

Fig 3: Constructed KG - K Means(BERT)

Fig 2: Existing KG - K Means(Word2Vec)

Fig 4: Constructed KG - K Means(Word2Vec)

Limitation

- Scalability challenges in handling large RDF datasets
- Embedding accuracy may depend on the quality of the knowledge graph
- Clustering performance varies across different algorithms

Conclusion

The RDF Data Clustering Framework effectively organizes knowledge graphs using embedding and clustering. The results indicate that BERT-based embedding perform better than Word2Vec in structuring entity relationships.

Reference

[1] P. Ristoski and H. Paulheim, "RDF2Vec: RDF graph embeddings for data mining," in Proceedings of the International Semantic Web Conference (ISWC), Mannheim, Germany, 2016, pp. 498–514. DOI: 10.1007/978-3-319-46547-0_30.

[2] S. Eddamiri, E. Zemmouri, and A. Benghabrit, "RDF Data Clustering based on Resource and Predicate Embeddings," in Proceedings of the 10th Int. Joint Conf. on Knowledge Discovery, Lisbon, Portugal, 2018, pp. 367–373. DOI: 10.5220/0007228903670373.