Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

On semantic mapping to visualize Knowledge Graphs

P. CAMARILLO-RAMIREZ[1], L. F. GUTIÉRREZ-PRECIADO[1], AND F. CERVANTES-ALVAREZ.[1]
[1]Western Institute of Technology and Higher Education, Tlaquepaque, Jalisco 45601 Mexico
(pablo.camarillo@iteso.mx;lgutierrez@iteso.mx;fcervantes@iteso.mx)

Corresponding author: P. Camarillo-Ramirez (e-mail: pablo.camarillo@iteso.mx)

This work was supported in part by the National Council of Science and Technology of Mexico through grant 498322

ABSTRACT Knowledge Graphs (KGs) are one of the most novel technologies used to improve search engines and support decision making in the life sciences since they structure information in graph form by encoding concepts as nodes, and the semantics of the relationship among concepts as edges. The analysis of KGs calls for an effective strategy to visualize them. However, the increasing size of KGs makes the exploring process a big challenge. A semantic map is a visual representation of related concepts that helps humans in the learning process. In this work, we propose to generate a simplified visual representation of a KGs by generating semantic maps. We apply several clustering algorithms to group the *related* concepts in KGs. We used different semantic similarity metrics to compute the matrix consumed by the clustering algorithms.

INDEX TERMS Knowledge graphs; Knowledge graphs visualization; Semantic similarity; Semantic mapping; Big Data;

I. INTRODUCTION

Knowledge Graphs are considered one of the emerging technologies associated with Big Data by providing semantic structured information that can be interpreted by machines, and such attribute is used to speed up the production of more intelligent systems [1].The core idea behind a Knowledge Graph (KG) is to represent knowledge from real world in a graph structure, where nodes represent entities of interest and edges represent relations between these entities [2]. Recently, academic and private organizations have constructed KGs, such as YAGO [3], DBPedia [4], Freebase [5], NELL [6], Google Knowledge Graph [7], Microsoft Satori [8], Facebook Entity Graph [9], and Wikidata [10], which contain millions of entities and billions of relationships. The main applications of KGs include the enhancement of search engines like Google [7] or Bing [8], question answering [11], information retrieval, recommender systems [12], [13], domain-specific KG building [14]–[16], and decision support in the life sciences [1], [17]–[19].

Some definitions describe a KG as a graph-structured knowledge base [20], [21]. In this work, we consider a knowledge base a set of sentences/facts expressed in some formal language such as description logic. According to description logic terminology, knowledge bases have two types of axioms: a terminology box (TBox) and an assertion box (ABox) [22], hence a KG should contain these two sets of axioms to be considered as a knowledge base. To exemplify the above idea, Figure 1 shows sets TBox and ABox of a group of entities and relationships extracted from DBPedia [4] [1]. In KGs, the ontology classes (e.g.,`dbo:Book` [2] or `dbo:Movie`) correspond with the TBox and describe concepts hierarchies, while the ontology instances correspond with the ABox and describe entity instances (e.g., `dbr:Lucasfilm` or `dbr:George_RR_Martin`) and their relationships. Hierarchical relationships like *is a* defines the connection between each pair of concepts in TBox. For example, axioms (`dbo:Book`, *is a*, `dbo:Work`) and (`dbo:film`, *is a*, `dbo:Work`) describe the fact that both `dbo:Book` and `dbo:film` concepts are descendants of the class `dbo:Work`. Alternatively, in ABox, axioms also indicate the list of types that one entity instance may have. For instance, in Figure 1, the axiom (`dbr:A_dance_of_dragons`, *rdf:type*, `dbo:Book`) indicates that resource `dbr:A_dance_of_dragons` is an instance of class `dbo:Book`. Another type

[1]https://dbpedia.org
[2]URIs mentioned in this document use the common prefixes described in https://prefix.cc

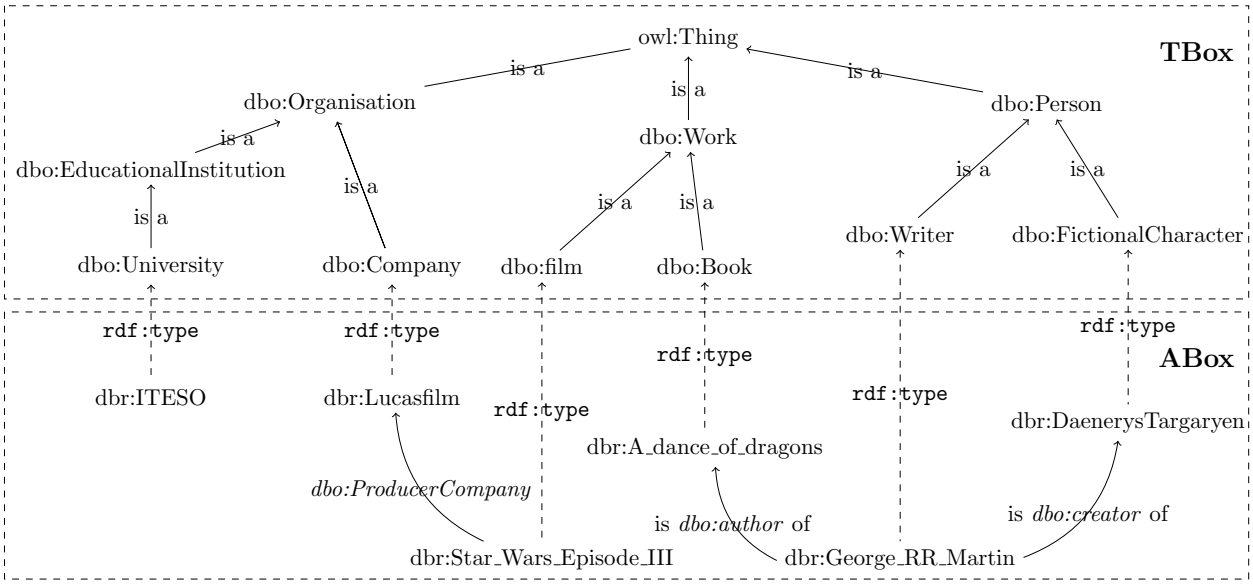

FIGURE 1: Small group of concepts and instances extracted from DBPedia.

of axioms in ABox like (`dbr:George_RR_Martin`, is *dbo:creator_of* of, `dbr:DaenerysTargaryen`) and (`dbr:George_RR_Martin`, is *dbo:author* of, `dbr:A_dance_of_dragons`) indicate that the instance `dbr:George_RR_Martin` has two semantic connections with `dbr:DaenerysTargaryen` and `dbr:A_dance_of_dragons` entity instances.

Considering the continuous increase in size of the KGs, the goal of visual graph analysis become uncertain due to a lacking ability to observe details from the presented information. The visual data exploration is considered as a hypothesis-generator process by allowing users to gain a deep understanding of the data [23], hence producing an effective visual representation of a KG is crucial. Existing approaches to visualize KGs are focussed on drawing the whole structure [24] preventing data analysts to explore the KG beyond its structural information. Consalvi et. al [25] propose an strategy to render large graphs in

Moreover, semantic mapping is a technique widely used to understand complex topics and consists of a categorical structuring of information in a graphic form [26]. Semantic maps have a central word indicating the main topic of the map and it is connected with a set of keywords that groups the rest of the vocabulary. Figure 2 shows an example of a semantic map describing the topic *Water*. It contains three node categories: (1) the central words (root), (2) the set of keywords (e.g., Usages, Living things, etc.), and (3) the vocabulary associated to each keyword, for instance, words `Cooking` and `Bathing` are associated with keyword *Usages*.

In this paper, we hypothesize that semantics maps are useful to visualize the high level of abstraction of a KG based on the semantic closeness of its entity instances. To generate a semantic map it is necessary to find the groups

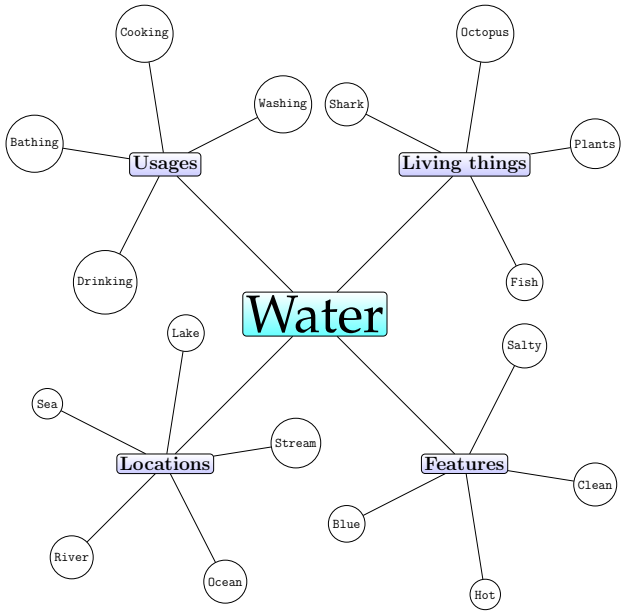

FIGURE 2: Example of a semantic map of concepts and vocabulary associated with topic *Water*.

of related instances. In unsupervised learning, clustering algorithms classify data into one or more classes depending on a similarity or distance measure [27]. Theoretically, if a clustering strategy is applied over the set of entity instances of a KG, it will group those entity resulting groups can be used to build the semantic map. Section IV presents a set of experiments validating the above notion.

The main contribution of this work is the formal definition of the semantic map of a KG as well as a set of experiments showing that semantic maps can be useful to provide a high-

level view of a KG.

The rest of this paper is structured as follows. In the Section II, we review the most relevant works associated with knowledge graph visualization, graph clustering, semantic similarly, and semantic mapping topics. Section III, describes the proposed method to generate the semantic maps of a KG. Section IV presents a set of experiments evaluating the semantic maps generated from a selected group of datasets extracted from DBPedia. Finally, Section V presents a discussion on the obtained results as well as some final remarks for this work.

II. RELATED WORK

Let us provide a formal definition of KG before describing the most relevant topics associated with the semantic mapping process.

Definition 1 (Knowledge Graph): Given a set of entities V, a set of property labels L, and a set of edges E a Knowledge Graph K is defined as $K = (V, L, E)$ where E is a subset of the cross product of entities and property labels defined as $V \times L \times V$. Each member of E is referred to as a triple $(subject - property - value)$.

A. VISUAL DATA EXPLORATION OF KNOWLEDGE GRAPHS

The idea behind the visual data exploration process is to present the data in some visual form, allowing users to draw conclusions of the analyzed phenomena [23]. This process, also known as the *information seeking mantra*, follows three steps: overview, zoom and filter, and details-on-demand [28]. In this context, ontologies are considered one of the most relevant data visualization techniques. In the field of computer science, an ontology is a model for describing the world that consists of a set of types, properties, and relationship types [29] and by providing an initial attempt to visualize linked data.

In regard to visual exploration of KGs, challenges include context adaptation, users input [30], data heterogeneity [31]–[33], supporting diverse analysis tasks (query, combination, filtering, etc.), and performance [24]. In this study, semantic mapping proposal is to combine and reduce the number of edges in the KG using the semantic similarity among its entities.

Recent applications have proven useful for large graph visualizations to understand different phenomena, such as Bitcoin transactions [34] and online discussions [35]. For big knowledge graphs, it is necessary a distributed implementation of the layout algorithms to improve the time needed to generate the visual representation [24].

In addition of recent efforts on KGs visualization, there are some commercial products enabling analysts to visualize RDF graphs like Data Graphs[3] or the family of tools developed by Cambridge Intelligence company: Keylines [4], ReGraph [5], and KronoGraph [6] that offer the capability to render KGs to support tasks in areas like pharmacy and bioscience research or financial analysis. In the area of free tools, there are two online tools that consumes RDF data and produce a visual representation: RDF visualizer [7] and RDF grapher [8]. The main limitation with these tools is the small amount of data they can process.

B. KNOWLEDGE GRAPH SUMMARIZATION

Recent works [32], [33], [36] have shown that visualizing a simplified version of a large graph is an adequate alternative. In the context of data mining, summarization is the process of facilitating the identification of meaningful data. The applications of graph summarization include reduction of data volume and storage, speedup of graph algorithms and queries, interactive analysis support, and noise elimination [37]. Recently, it has been proposed to summarize large graphs in order to enable an efficient visualization of their content. For example, in [30], the authors focus on summarizing KGs by taking advantage of individual interests to generate personalized knowledge graph summaries. In [31], Shen et al. propose a visual analytics tool called OntoVis, which performs both structural and semantic abstractions to offer a summarized version of a large graph and thus being able to visualize a simplified version of the graph. Another related work is presented in [38], which describes the VoG (Vocabulary-based summarization of Graphs) algorithm to summarize and understand large graphs by constructing and visualizing subgraph-types, such as starts, cliques, and chains. The visual abstraction presented in [36], transforms geo-tagged social media data into high-dimensional vectors by utilizing a doc2vec model.

Regardless of the application, one of the main challenges of graph summarization is defining which data is of interest. Every summarization strategy depends on selecting an interest criteria to extract meaningful information [37]. However, to achieve a concise definition of *interesting* is not an easy task. For example, the FUSE algorithm [39] proposes a profit maximization model that seeks to find a summary by maximizing information profit under a budget constraint. On the other hand, VoG [38] exploits the Minimum Description Length (MDL) principle aimed at identifying the best subgraphs by choosing those which save most bits. In the case of semantic abstraction proposed in [36], a dual-objective blue noise sampling model is utilized to select a subset of social media data items supporting the spatial distribution and semantic correlation for the resulting simplified geographical visualization. The personalized summaries of KGs described in [30], the criteria to decide which information is *interesting* for each user is determined by reviewing the users' query history. The work of M. Tasnmin et al. propose a strategy to find equivalent entities in a KG using the context of each

[3]https://datagraphs.com
[4]https://cambridge-intelligence.com/keylines/
[5]https://cambridge-intelligence.com/regraph/
[6]https://cambridge-intelligence.com/kronograph/
[7]https://issemantic.net/rdf-visualizer
[8]https://www.ldf.fi/service/rdf-grapher

RDF Molecule [40]. Semantic mapping process described in this document uses the semantic similarity between each pair of entity instances in the KG to infer the groups of related instances.

C. SEMANTIC SIMILARITY

The semantic similarity is a metric used in Natural Processing Language (NPL) and Information Retrieval (IR) areas [41] that represents how related are two concepts based on their hierarchical relations [42], [43]. In a KG, the semantic similarity between two entities $e_1, e_2 \in V$ is denoted as $sim(e_1, e_2)$. Intuitively, semantic distance between two words is the most easy way to calculate semantic similarity and it is usually determined by the path connecting two entities in KG. Existing semantic similarities metrics are classified in two main groups: corpus-based and knowledge-based approaches [44]. Corpus-based similarity metrics are focused on learning how similar are two concepts based on the information from large corpora. Two examples of corpus-based similarity metrics are pointwise mutual information [45], and latent semantic analysis [46]. In contrast, knowledge-based similarity metrics quantify the degree to which two words are semantically related [47]. In KG, knowledge-based approaches, semantic similarity is determined using the information provided by the TBox. Knowledge-based approaches include path-based metrics such as those proposed by Hulpus et at. [48], Wu & Palmer [49], and Leacock & Chodorow [50]. Other knowledge-based measures utilize the Information Content (IC) metric like Lin [51], Jiang and Conrath [52], and Resnik [42] metrics. IC of concepts is a statistical measure that computes the specificity of a concept over a corpus. Higher values of IC indicate more specific concepts (e.g., `dbo:Book`) and lower values of IC are associated with more general concepts (e.g., `owl:Thing`). Hybrid knowledge-based approaches like IC-graph [53] or Zhou [54] combine IC and some other metrics to compute how related two words are. For instance, graph-based IC [53] uses the counting services provided by DBPedia and it is calculated using the following expression:

$$IC_{graph}(c_i) = -logProb(c_i) \tag{1}$$

Where $Prob(c_i) = \frac{freq_{graph}(c_i)}{N}$ and N is the number of entities in the KG. Let $\mathcal{E}(c_i)$ the set of entities having type of c_i, the frequency of concept c_i in the KG is defined as $freq_{graph}(c_i) = |\mathcal{E}(c_i)|$.

D. CENTROID-BASED CLUSTERING

There exist several techniques to clustering data and recent surveys summarize these clustering approaches based on the application or the type of data to group [55], [56], [57]. Types of clustering include Centroid-based, Density-based, Distribution-based, and Hierarchical clustering [58].

One of the phases of the semantic mapping process is to collocate each entity into the most appropiate cluster based on its semantic similarity. Each resulting cluster needs a node that represents all entities contained on it. We denote the set of this representing nodes as the keywords of the semantic map. Considering these keywords as centroids of clusters, the usage of a centroid-based clustering is crucial.

The main idea of centroid-based clustering is to find k centroids (or centers) followed by computing k sets of data points that minimize the proximity with each center. For instance, K-means algorithm tries to minimize the sim of the squared distance between the data points and the cluster's centroid [59]. A variation of K-means is the PAM (Partitioning Around Medoids) algorithm that minimizes dissimilarities between points in a cluster and the centroids [60]. The CLARA (Clustering Large Applications) algorithm is an extension of PAM for large datasets [61]. On the other hand, CLARANS (Clustering Large Applications based on RANdomized Search) is a partitioning algorithm focused on spatial data mining because it recognizes patters and relationships existing in spatial data such as topological data [62]. One last centroid-based clustering algorithm is the Affinity Propagation (AP) algorithm which consists on a message-passing procedure that looks for broadcasting messages of attractiveness and availability among data points [63].

III. PROPOSED METHOD

The notion behind the semantic map of a KG is to produce a reduced version of the KG by exploiting the semantic similarity between each pair of entities. To illustrate this idea, let us generate a small KG from DBPedia containing the list of some fictional characters from series of fantasy novels by the novelist George R. R. Martin. Figure 3a) presents a visual representation produced by Gephi[9], which is one of the most used tools to visualize networked data. However, this kind of visualization is not visually super informative or visually appealing which may lead in an ineffective exploratory visual analysis. On the contrary Figure 3b) shows a semantic map of the KG by grouping the entity instances and showing the central concept of the map that connects the centroids of the cluster.

A. CLUSTERING ENTITIES OF KNOWLEDGE GRAPHS

The first phase of the semantic mapping process is to group the entities of the KG based on the semantic closeness between each pair of entities in the KG. The main challenge of this phase is to extract numeric data from the KG and generate a set of groups of entities.

Our proposal consist of using a centroid-based clustering algorithm and generate a non-overlapping set of clusters by using the semantic distance matrix as input of the selected clustering algorithm. The semantic relatedness among entities in the KG is encoded in a **semantic distance matrix**.

Definition 2 (Semantic distance matrix): Given a Knowledge Graph $K = (V, L, E)$, and $sim(e_1, e_2)$ the semantic similarity between entities e_1 and e_2, the semantic similarity matrix

[9]https://gephi.org/

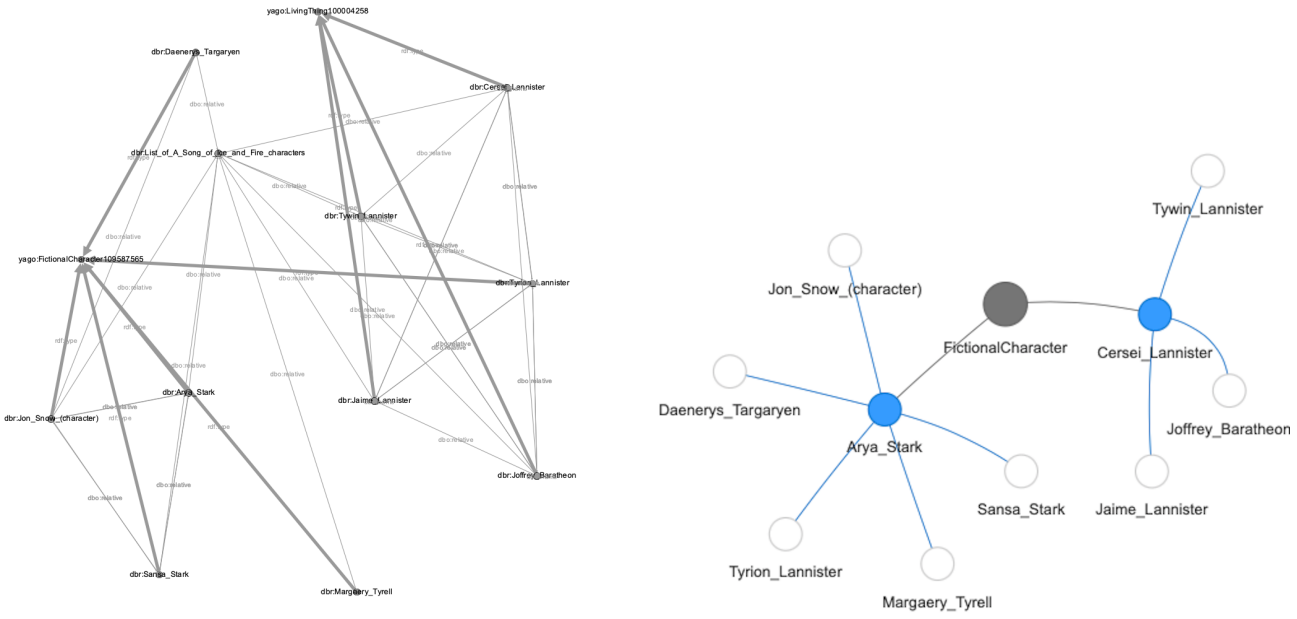

FIGURE 3: Visual representation from a small KG containing some fictional characters by George R.R. Martin. a) Contains the visual representation produced by Gephi. b) Inferred semantic map of the original KG.

$D(K)$ represents the semantic distance between each pair of entities in K i.e., the value for cell $d_{i,j} = 1 - sim(e_i, e_j)$.

The relation between similarity and distance follows the notion that the higher is the similarity between two entities the lower is the distance between these entities.

The $i-th$ row of $D(K)$ is the vector containing semantic distance values between the $i-th$ entity and the rest of entities in the KG. Let $C = \bigcup C_i$ the set of clusters resulting after applying a centroid-based clustering algorithm. Each cluster C_i has a centroid element denoted by $centroid(C_i)$ and the set of centroid elements is denoted by $\mathcal{C} = \bigcup_{C_i \in C} centroid(C_i)$.

B. CENTRAL CONCEPT OF THE SEMANTIC MAP

One of the main features of a semantic map is the **central concept** that represents the main topic of this graphical representation. In this work, we denote this central concept as α. In a regular semantic map, α is connected with a set of selected keywords (e.g., structures, characteristics, size, habitat, movie, kinds in Figure 2). These keywords are used to represent every group of words of the semantic map. This work proposes to use the centroids inferred by centroid-based clustering algorithms [55] as the keywords of a KG. Therefore, we denote these keywords as the set of centroids $\mathcal{C}$ of the entities in a KG.

To infer the central term α, we propose to compute the IC_{graph} measure for all types associated with each centroid in $\mathcal{C}$. Let $types(e_i)$ to be the function to retrieves set of types associated with the entity e_i, we define $\mathcal{T}$ as the set of shared types among all centroids in KG. This definition is formally described in equation 2.

$$\mathcal{T} = \bigcap_{c_i \in \mathcal{C}} types(c_i) \tag{2}$$

Definition 3 (Central concept α): Given a set of shared types $\mathcal{T}$, the central concept α of G is the concept $c_i \in \mathcal{T}$ with maximum IC_{graph}.

Thus, the central term α represents the most specific type among all centroids in the semantic map.

C. SEMANTIC MAP OF A KNOWLEDGE GRAPH

Semantic mapping process aggregates the process of clustering entities of KG and inferring the central term α.

Let us define the semantic map of a KG:

Definition 4 (Semantic map of a Knowledge Graph): Given a Knowledge Graph $K = (V, L, E)$ and $D(K)$ the semantic distance matrix associated with K, the semantic map of K is defined as $\mathcal{SM}(K) = (\alpha, \mathcal{C}, \mathcal{NC}, E_c, E_{nc}, \Psi)$. Table 2 describes the symbols associated with semantic maps of KGs.

IV. EVALUATION STUDY

The goal of experiments described in this section is to validate the process of generating a visualization of a reduced version of a KG through a semantic map. We introduce a tool used to automate the process of generating semantic maps from different datasets. Then, we describe the datasets used to test the semantic mapping process and finally we present the cluster quality obtained for each experiment.

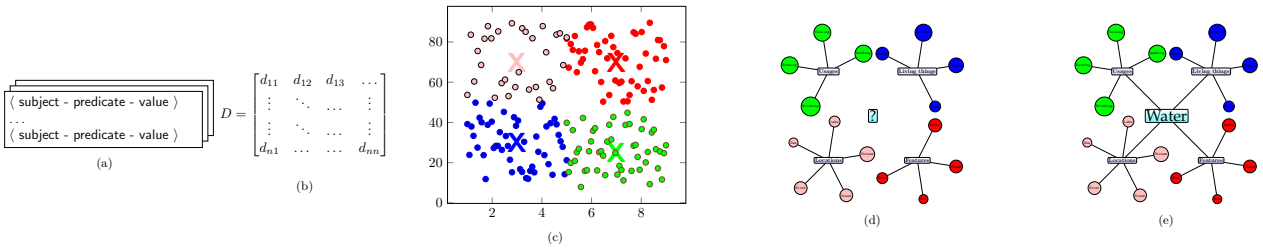

FIGURE 4: Semantic Mapping phases. (a) Consume a KG as a list of n-triples, (b) Generate the semantic distance matrix D, (c) Cluster entities using the matrix D, (d) Infer main term α, and (e) Assemble the semantic map by connecting each centroid with α.

TABLE 1: Symbols associated with semantic maps of Knowledge Graphs.

Symbol	Description
α	Central concept of the semantic map.
$\mathcal{C}$	Centroid entities $\mathcal{C} \subseteq V$ produced by a centroid-based clustering algorithm.
E_c	Set of edges connecting each centroid with the central term α, defined as $E_c = \mathcal{C} \times \alpha$
$\mathcal{NC}$	Non-centroids entities with $\mathcal{NC} \subseteq V$ and $\mathcal{NC} \cap \mathcal{C} = \emptyset$.
Ψ	Function $\mathcal{NC} \times \mathcal{C} \rightarrow E_{nc}$ that connects each non-centroid entity with its unique centroid entity.

A. SEMANTIC MAPPING FRAMEWORK

Experiments are executed using a framework implemented using Python 3 language which depends on the *Sematch* framework [64] to perform SPARQL queries to DBPedia public endpoint and compute the similarity measure used to compute D, i.e., the function $sim(e_i, e_j)$ corresponds with a SPARQL query to DBPedia. Once generated D, the tool produces the set of centroids $\mathcal{C}$ and the set of non-centroid $\mathcal{NC}$ nodes by using two centroid-based clustering approaches: PAM and Affinity Propagation. We infer the main term α by maximizing the IC_{graph} of common types of each centroid in $\mathcal{C}$. These shared types are the result of a series of SPARQL queries to DBPedia (see Figure 5).

```
SELECT DISTINCT ?o WHERE {
<RDF Concept>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
?o .
} LIMIT 5000
```

FIGURE 5: Template of the SPARQL query to get the list of types associated with each centroid.

Finally, our tool, assembles the semantic map by connecting α with each centroid and connecting each centroid with all entities in the corresponding cluster. Figure 4 visually describes the phases of semantic mapping process.

B. QUALITY OF SEMANTIC MAPS

The core of the semantic mapping process is to cluster the entity instances and obtain the set of centroids $\mathcal{C}$. In order to provide a quantitative approach to validate semantic maps, we propose to associate the quality of clusters computed with the quality of semantic maps. With this evaluation strategy, we can learn how reliable are the groups shown in the semantic map.

Literature offers two classes of clustering validation measures: external clustering validation and internal clustering validation [56]. Internal validation metrics evaluate the quality of a clustering algorithm based on its intrinsic properties, while external validation methods evaluate the quality of a clustering solution based on its agreement with a known label of the data. Since there is no known label of the datasets used in the experiments described in this work, our proposal is to use internal validation measures such as Silhouette score, Davies-Bouldin score, and Calinski-Harabasz Index.

Each internal validation metric measure different aspects of the clusters. For example, Silhouette score measures how well each data point fits into its assigned cluster compared to other clusters and the score ranges goes from -1 to 1, where a higher score indicates better cluster quality [60]. Inertia of a cluster, also known as the within-cluster sum of squares (WSS) metric measures how tightly packed the data points are within each cluster. The goal is to minimize inertia, which is equivalent to maximizing the distances between clusters. On the other hand, Dunn index measures the distance between the nearest points in different clusters and the distance between the farthest points in each cluster [65]. A higher Dunn index indicates better cluster quality. Another known quality measure is the Davies-Bouldin index which measures the similarity between each cluster and its closest neighboring cluster, while also considering the cluster's internal similarity [66]. A lower index indicates better cluster quality. Finally, Calinski-Harabasz index measures the ratio of between-cluster variance to to within-cluster variance [67]. A higher value of Calinski-Harabasz index indicates better cluster quality.

C. DATASETS

Datasets used to validate the process are obtained by performing a SPARQL query against DBPedia KG through

its public endpoint [10] and results are saved as N-Triples format[11], i.e., each dataset is a list of subject-predicate-object triples. The intention of each dataset is to represent different knowledge domains accumulated in DBPedia and how they can be reduced and visualized using semantic maps.

1) SCI-FI-MOVIES.NT
A list of triples describing sci-fi movies stashed in DBPedia with a gross greater than eight billion of dollars.

2) FANTASY-NOVELS.NT
This dataset contains all fantasy novels kept in DBPedia KG published after year 2000.

3) CITIES.NT
According to DBPedia, the list of cities with a total population greater than five millions.

4) MAMMALS.NT
This resultset representing the mammals stored in DBPedia.

5) PLANTS.NT
This list contains all plants found in latin america region.

6) SONGS.NT
This collection of triples contains all rock songs reaching Diamond rank worlwide.

7) MOVIES-AND-SONGS.NT
8) MAMMALS-AND-PLANTS.NT

TABLE 2: Dataset summary

Dataset	Number of triples
SCI-FI-MOVIES.NT	188
FANTASY-NOVELS.NT	693
CITIES.NT	127
MAMMALS.NT	
PLANTS.NT	
SONGS.NT	
MOVIES-AND-SONGS.NT	
MAMMALS-AND-PLANTS.NT	

D. EXPERIMENTAL RESULTS

A key hyperparameter in the PAM clustering algorithm is the number of clusters we want to generate. Determining this hyperparameter is a crucial step in clustering and we determine this value by using the elbow method[12]. Table 3 describes the optimal number of clusters suggested by the elbow method for each dataset to use the PAM clustering strategy.

TABLE 3: Optimal number of clusters(k) for PAM clustering suggested by selbow method.

Dataset	Optimal k	WSS
SCI-FI-MOVIES.NT	24	0.27
FANTASY-NOVELS.NT	40	1.73
CITIES.NT	16	1.69

The *preference* parameter is a crucial hyperparameter in the Affinity Propagation clustering algorithm, which determines the number of clusters that will be generated. A higher preference value will result in more clusters, as more data points will be selected as exemplars, while a lower preference value will lead to fewer clusters, as fewer data points will be selected as exemplars. Therefore, it is often necessary to perform sensitivity analysis by trying different values of the preference parameter to find the optimal number of clusters. Our proposal is to maximize the silhouette index of resulting clustering after running Affinity Propagation with preference values that goes from 0.1 to 0.9 since this is the range of possible semantic distance values in the distance matrix.

TABLE 4: Optimal value of *preference* hyperparameter in the Affinity Propagation clustering.

Dataset	Number of clusters k	Preference
SCI-FI-MOVIES.NT	5	0.8
FANTASY-NOVELS.NT	6	0.8
CITIES.NT	5	0.5

V. CONCLUSIONS

REFERENCES

[1] X. Zou, "A survey on application of knowledge graph," *Journal of Physics: Conference Series*, vol. 1487, no. 1, p. 012016, mar 2020. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1487/1/012016

[2] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs," vol. 54, no. 4, 2021.

[3] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 697–706.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The Semantic Web*, K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 722–735.

[5] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 1247–1250.

[6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell, "Toward an architecture for never-ending language learning," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, ser. AAAI'10. Atlanta, Georgia: AAAI Press, 2010, pp. 1306–1313.

[10] https://dbpedia.org/sparql/

[11] N-Triples is a subset of the more complex RDF/XML syntax, and is designed to be human-readable as well as machine-readable. It is a plain text format that represents RDF statements using subject-predicate-object triples, with each element separated by whitespace and terminated by a period.

[12] The elbow method is a heuristic approach to determine the optimal number of clusters in a datase and the idea behind this method is that as the number of clusters increases, the WCSS decreases, as the distance between each data point and its assigned center becomes smaller.

TABLE 5: Semantic mapping quality results

	PAM			Affinity Propagation		
Dataset	Silhouette score	Davies-Bouldin score	Calinski-Harabasz Index	Silhouette score	Davies-Bouldin score	Calinski-Harabasz Index
MOVIES_SCIFI	0.86	0.28	1366.07	0.45	2.33	17.38
FANTASY_NOVELS	0.66	3.73	133.56	0.38	1.53	10.93
CITIES.NT	0.69	0.46	281.70	0.47	1.27	77.45

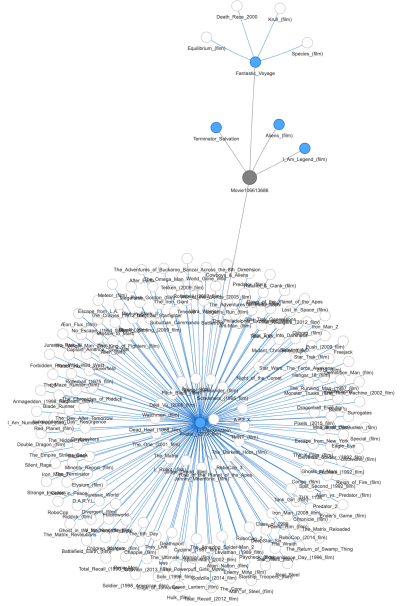

FIGURE 6: Semantic map obtained using AP clustering alforithm for MOVIES_SCIFI dataset.

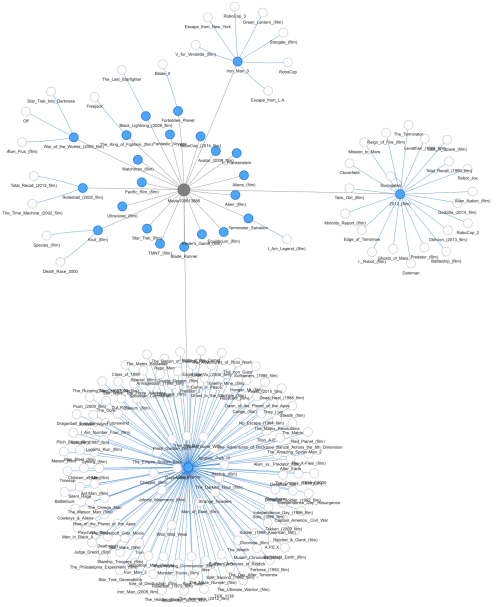

FIGURE 7: Semantic map obtained using PAM algorithm for MOVIES_SCIFI dataset.

[7] A. Singhal. (2012) Introducing the Knowledge Graph: things, not strings. [Online]. Available: https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

[8] R. Qian. (2013) Understand Your World with Bing. [Online]. Available: https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/

[9] F. Engineering. (2013) Under the Hood: The Entities Graph. [Online]. Available: https://www.facebook.com/notes/facebook-engineering/under-the-hood-the-entities-graph/10151490531588920

[10] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić, "Introducing wikidata to the linked data web," in *The Semantic Web – ISWC 2014*, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble, Eds. Cham: Springer International Publishing, 2014, pp. 50–65.

[11] Y.-H. Chen, E. J.-L. Lu, and Y.-Y. Lin, "Efficient sparql queries generator for question answering systems," *IEEE Access*, vol. 10, pp. 99 850–99 860, 2022.

[12] Y. Lin, S. Du, Y. Zhang, K. Duan, Q. Huang, and P. An, "A recommendation strategy integrating higher-order feature interactions with knowledge graphs," *IEEE Access*, vol. 10, pp. 119 290–119 300, 2022.

[13] H. Li, Y. Wang, S. Zhang, Y. Song, and H. Qu, "Kg4vis: A knowledge graph-based approach for visualization recommendation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 195–205, 2022.

[14] K. Zhang, H. Wang, M. Yang, X. Li, X. Xia, and Z. Guo, "A knowledge graph completion method for telecom metadata based on the spherical coordinate system," *IEEE Access*, vol. 10, pp. 122 670–122 678, 2022.

[15] A. Borrego, D. Dessì, I. Hernández, F. Osborne, D. Reforgiato Recupero, D. Ruiz, D. Buscaldi, and E. Motta, "Completing scientific facts in knowledge graphs of research concepts," *IEEE Access*, vol. 10, pp. 125 867–125 880, 2022.

[16] K. Guan, L. Du, and X. Yang, "Relationship extraction and processing for knowledge graph of welding manufacturing," *IEEE Access*, vol. 10, pp. 103 089–103 098, 2022.

[17] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2rdf: towards a mashup to build bioinformatics knowledge systems," *Journal of biomedical informatics*, vol. 41, no. 5, pp. 706–716, 2008.

[18] A. Ruttenberg, J. Rees, M. Samwald, and M. Marshall, "Life sciences on the semantic web: The neurocommons and beyond," *Briefings in bioinformatics*, vol. 10, pp. 193–204, 04 2009.

[19] V. Momtchev, D. Peychev, T. Primov, and G. Georgiev, "Expanding the pathway and interaction knowledge in linked life data," in *Proc. of International Semantic Web Challenge*, 2009.

[20] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 11–33, 2015.

[21] S. Seufert, P. Ernst, S. J. Bedathur, S. K. Kondreddi, K. Berberich, and G. Weikum, "Instant espresso: interactive analysis of relationships in knowledge graphs," in *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016, pp. 251–254.

[22] I. Horrocks, "Ontologies and the semantic web," *Communications of the ACM*, vol. 51, no. 12, pp. 58–67, 2008.

[23] D. A. Keim, "Visual exploration of large data sets," *Communications of the ACM*, vol. 44, no. 8, pp. 38–44, 2001.

[24] J. Gómez-Romero, M. Molina-Solana, A. Oehmichen, and Y. Guo, "Visualizing large knowledge graphs: A performance analysis," *Future Generation Computer Systems*, vol. 89, pp. 224–238, 2018.

[25] L. Consalvi, W. Didimo, G. Liotta, and F. Montecchiani, "Browvis: Visualizing large graphs in the browser," *IEEE Access*, vol. 10, pp. 115 776–115 786, 2022.

[26] D. D. Johnson, S. D. Pittelman, and J. E. Heimlich, "Semantic mapping," *The reading teacher*, vol. 39, no. 8, pp. 778–783, 1986.

[27] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007.

[28] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations," in *Proceedings 1996 IEEE Symposium on Visual Languages*. Boulder, CO, US: IEEE, Sep. 1996, pp. 336–343.

[29] L. M. Garshol, "Metadata? thesauri? taxonomies? topic maps! making sense of it all," *Journal of Information Science*, vol. 30, pp. 378 – 391, 2004.

[30] T. Safavi, C. Belth, L. Faber, D. Mottin, E. Müller, and D. Koutra, "Personalized knowledge graph summarization: From the cloud to your pocket," in *2019 IEEE International Conference on Data Mining (ICDM)*, 2019, pp. 528–537.

[31] Zeqian Shen, Kwan-Liu Ma, and T. Eliassi-Rad, "Visual analysis of large heterogeneous social networks by semantic and structural abstraction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1427–1439, 2006.

[32] L. Shi, Q. Liao, H. Tong, Y. Hu, Y. Zhao, and C. Lin, "Hierarchical focus+context heterogeneous network visualization," in *2014 IEEE Pacific Visualization Symposium*, Yokohama, Japan, 2014, pp. 89–96.

[33] Zeqian Shen, Kwan-Liu Ma, and T. Eliassi-Rad, "Visual analysis of large heterogeneous social networks by semantic and structural abstraction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1427–1439, 2006.

[34] D. McGinn, D. Birch, D. Akroyd, M. Molina-Solana, Y. Guo, and W. J. Knottenbelt, "Visualizing dynamic bitcoin transaction patterns," *Big data*, vol. 4, no. 2, pp. 109–119, 2016.

[35] M. Molina-Solana, D. Birch, and Y.-k. Guo, "Improving data exploration in graphs with fuzzy logic and large-scale visualisation," *Applied Soft Computing*, vol. 53, pp. 227–235, 2017.

[36] Z. Zhou, X. Zhang, X. Zhou, and Y. Liu, "Semantic-aware visual abstraction of large-scale social media data with geo-tags," *IEEE Access*, vol. 7, pp. 114 851–114 861, 2019.

[37] Y. Liu, T. Safavi, A. Dighe, and D. Koutra, "Graph summarization methods and applications: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, pp. 1–34, 2018.

[38] D. Koutra, U. Kang, J. Vreeken, and C. Faloutsos, "Vog: Summarizing and understanding large graphs," in *Proceedings of the 2014 SIAM international conference on data mining*. SIAM, 2014, pp. 91–99.

[39] B.-S. Seah, S. S. Bhowmick, C. F. Dewey, and H. Yu, "Fuse: a profit maximization approach for functional summarization of biological networks," *BMC Bioinformatics*, vol. 13, no. 3, p. S(10), 2012.

[40] M. Tasnim, D. Collarana, D. Graux, and M.-E. Vidal, *Chapter 8: Context-based Entity Matching for Big Data*. Cham: Springer International Publishing, 2020.

[41] E. Hovy, R. Navigli, and S. P. Ponzetto, "Collaboratively built semi-structured content and artificial intelligence: The story so far," *Artificial Intelligence*, vol. 194, pp. 2–27, 2013, artificial Intelligence, Wikipedia and Semi-Structured Resources.

[42] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th Int. Joint Conf. Artif. Intell.* Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995, pp. 448–453.

[43] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, pp. 141–188, 2010.

[44] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Aaai*, vol. 6, no. 2006, 2006, pp. 775–780.

[45] K. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[46] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[47] A. Budanitsky and G. Hirst, "Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures," *Workshop on WordNet and Other Lexical Resources*, vol. 2, 2001.

[48] I. Hulpuş, N. Prangnawarat, and C. Hayes, "Path-based semantic relatedness on linked data and its use to word and entity disambiguation," in *Proc. 14th International Semantic Web Conference 2015*. Bethlehem, PA, USA: Springer, 2015, pp. 442–457.

[49] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *Proc. 32nd Annu. Meeting Assoc. Comput. Linguistics*, Las Cruces, New Mexico, USA, 1994, pp. 133–138.

[50] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," *WordNet: An electronic lexical database*, vol. 49, no. 2, pp. 265–283, 1998.

[51] D. Lin, "An information-theoretic definition of similarity." in *ICML '98: Proc. of the 15th Int. Conf. on Machine Learning*, Madison, Wisconson, USA, 1998, pp. 296–304.

[52] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. 10th Int. Conf. Res. Comput. Linguistics*, Taipei, Taiwan, Aug 1997, pp. 19—33.

[53] G. Zhu and C. A. Iglesias, "Computing semantic similarity of concepts in knowledge graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 72–85, 2017.

[54] Z. Zhou, Y. Wang, and J. Gu, "New model of semantic similarity measuring in wordnet," in *Proc. 3rd Int. Conf. on Intelligent System and Knowledge Engineering*, Xiamen, China, Nov 2008, pp. 256–261.

[55] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015. [Online]. Available: https://doi.org/10.1007/s40745-015-0040-1

[56] C. C. Aggarwal and C. K. Reddy, Eds., *Data Clustering: Algorithms and Applications*. CRC Press, 2014. [Online]. Available: http://www.crcpress.com/product/isbn/9781466558212

[57] S. Firdaus and M. A. Uddin, "A survey on clustering algorithms and complexity analysis," *International Journal of Computer Science Issues (IJCSI)*, vol. 12, no. 2, p. 62, 2015.

[58] "Clustering algorithms," Jul 2022. [Online]. Available: https://developers.google.com/machine-learning/clustering/clustering-algorithms

[59] J. MacQueen, "Classification and analysis of multivariate observations," in *5th Berkeley Symp. Math. Statist. Probability*, 1967, pp. 281–297.

[60] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* Hoboken, New Jersey: John Wiley & Sons, Inc., 1990, ch. Partitioning around medoids (program pam).

[61] ——, *Finding Groups in Data: An Introduction to Cluster Analysis.* Hoboken, New Jersey: John Wiley & Sons, Inc., 1990, ch. Clustering Large Applications (program CLARA).

[62] R. T. Ng and J. Han, "Clarans: A method for clustering objects for spatial data mining," *IEEE transactions on knowledge and data engineering*, vol. 14, no. 5, pp. 1003–1016, 2002.

[63] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[64] G. Zhu and C. A. Iglesias, "Sematch: Semantic similarity framework for knowledge graphs," *Knowledge-Based Systems*, vol. 130, pp. 30–32, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705117302447

[65] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[66] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, vol. 1, no. 2, pp. 224–227, 1979.

[67] T. Calinski and J. Harabasz, "Dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.