IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 November 2023, accepted 4 January 2024, date of publication 8 January 2024,
date of current version 17 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3351170

RESEARCH ARTICLE

Semantic Maps for Knowledge Graphs: A Semantic-Based Summarization Approach

PABLO CAMARILLO-RAMIREZ, FRANCISCO CERVANTES-ALVAREZ, AND LUIS FERNANDO GUTIÉRREZ-PRECIADO
Western Institute of Technology and Higher Education, Guadalajara 45601, Mexico

Corresponding author: Pablo Camarillo-Ramirez (pablo.camarillo@iteso.mx)

This work was supported in part by the National Council of Science and Technology of Mexico under Grant 498322.

ABSTRACT Knowledge Graphs (KGs) have emerged as a powerful tool for representing semantic structured information and enabling the development of intelligent systems. This paper focuses on the generation of semantic maps as summarization method for KGs. We propose a strategy that utilizes centroid-based clustering algorithms, namely Affinity Propagation and Partitioning Around Medoids (PAM), to capture the semantic distance between nodes in the KG and generate meaningful clusters. Our experiments demonstrate divergent results between the two clustering algorithms, with Affinity Propagation showing qualitative coherence and meaningfulness, while PAM performs well in terms of internal validation metrics. We leverage the computed centroids to infer a main term of the semantic map, which contributes to the visually appealing and informative representation of the KG. The combination of semantic distance capture, clustering algorithms, and centroid-based inference facilitates a comprehensive understanding of the KG. Our findings highlight the importance of considering both qualitative and quantitative evaluation measures in assessing clustering results. The effectiveness of semantic maps is showcased in visualizing KGs and advancing the field of knowledge graph visualization. The integration of centroid-based clustering algorithms, qualitative evaluation, and inference methods offers improved clarity and interpretability for complex KG analysis.

INDEX TERMS Knowledge graphs, knowledge graph visualization, semantic distance, semantic mapping, visual data exploration, clustering.

I. INTRODUCTION

Knowledge Graphs are considered as the fundamental building blocks for AI systems, providing the necessary foundation for representation and reasoning capabilities to address the imperative design requirement of involving humanity in the loop [1]. The 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], [15], [16], and decision support in the life sciences [17], [18], [19], [20].

Considering the continuous increase in the use of KGs in decision-making applications, it becomes important to compress and summarize KGs for efficient representation of data [21]. In general, one of the applications of summarizing graphs is to reduce the data volume and storage to facilitate the process of graph visualization [22]. Visual data exploration is considered as a hypothesis-generator process by allowing users to gain a deep understanding of the data [23], hence summarizing a KG is crucial to produce an effective visual representation to understand relationships

The associate editor coordinating the review of this manuscript and approving it for publication was Walter Didimo.

between entities and concepts in a domain. By representing the information in a visual format, users can quickly identify patterns, trends, and clusters or related information that may be difficult to see in a text-based representation. Existing approaches to visualize KGs are focussed on drawing the whole structure [24] preventing data analysts to explore the KG beyond its structural information.

Semantic maps, on the other hand, are widely used technique to understand complex topics [25]. They involved a categorical structuring of information in a graphic form [26]. A semantic map typically has a central word that represents the main topic, connected to a set of keywords that groups the remaining vocabulary. Generating a semantic map requires identifying groups of related words. Unsupervised learning provides clustering algorithms that classify data into one or more classes based on similarity or distance measures [27]. In theory, applying a clustering strategy to the vocabulary in a KG will result in groups that can be used to construct the semantic map. Section V presents a set of experiments validating the above notion.

The main contribution of this work is to provide a novel approach for summarizing KGs by leveraging the power of semantic maps. Some KG applications, such as query answering or KG visualization, require reduced versions of the original graphs [24], [28]. To address this challenge, we propose a method that generates a reduced version of a KG by creating semantic maps from the knowledge graph, enabling users to explore and comprehend the underlying structure and relationships of the graph in a visually intuitive manner. In addition, we present a formal definition of the semantic map of a KG and propose a strategy to measure the quality of these semantic maps. To demonstrate the utility of these semantic maps in providing a high-level view of a KG, we conduct a survey on a group of experts that compared the use of semantic maps with the classical visual representation of KGs. These experiments showcase the value of semantic maps in offering a comprehensive understanding of the KG's structure and relationships.

The paper is structured as follows. Section II provides details on KGs and establishes a definition that will be used throughout the rest of the paper. In Section III, we review the most relevant literature related to knowledge graph visualization, graph clustering, semantic similarly, and semantic mapping topics. Section IV, describes the proposed method and the algorithms necessary for generating semantic maps of a KG. In Section V, we discuss the results obtained from a series of experiments evaluating the method of generating semantic maps generated from a collection of datasets extracted from DBPedia. Finally, Section VI presents concluding remarks and outlines future work.

II. FOUNDATIONS OF KNOWLEDGE GRAPHS

Some definitions describe a KG as a graph-structured knowledge base [29]. In this work, we consider a knowledge base a set of sentences/facts expressed in some formal language such as description logic. In other words, KGs consist in a collection of facts formed by `<subject,predicate,object>`. These collections are typically represented in languages, such as RDF (Resource Description Framework) [31], OWL (Ontology Web Language) [32], or N-Triples which is a subset of the more complex RDF/XML syntax, is designed to be both human-readable and machine-readable. It is a plain text format that represents RDF statements using subject-predicate-object triples, with each element separated by white space and terminated by a period.

In the field of computer science, an ontology serves as a descriptive model of the world, comprising a collection of types, properties, and relationship types [33]. According to description logic terminology, knowledge bases have two types of axioms: a terminology box (TBox) and an assertion box (ABox) [34], 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 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.

Let us begin by proposing a formal definition of a KG before proceeding to describe the remaining relevant topics associated with the semantic mapping process outlined in this paper.

Definition 1 (Knowledge Graph): Let V the set of entities, where each entity $v \in V$ can be uniquely identified. Let L denote the set of property labels or attributes associated with the entities in the knowledge graph. Each label $l \in L$ represents a specific property or characteristic of an entity. Let E the set of edges in a Knowledge Graph K. A knowledge

[1] https://dbpedia.org (Last visited: 2023-11-16)

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

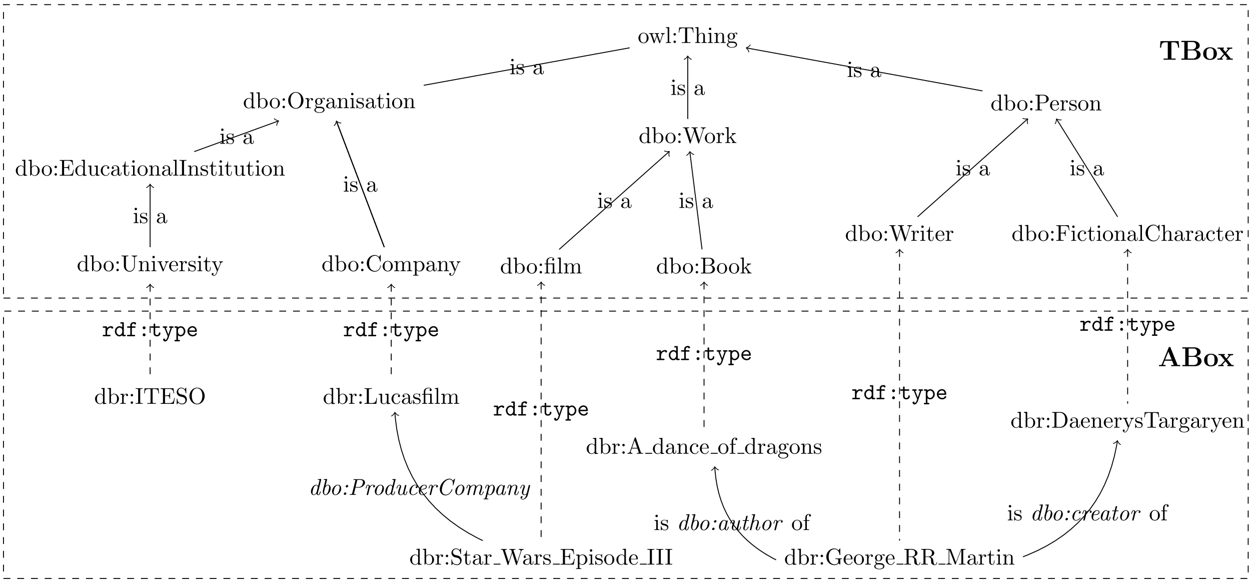

FIGURE 1. Small group of concepts and instances extracted from DBPedia.

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 $(subject - property - value)$.

From the definition above, we can deduce that a KG can be represented as a list of triples that capture axioms from ABox or TBox. In this work, our focus lies on the triples that represent axioms of the ABox set. Specifically, whenever an entity instance is mentioned, it is associated with the subjects of the ABox triples.

III. RELATED WORK

In this section, we present a comprehensive overview of key topics that are fundamental to the understanding and analysis of semantic maps of knowledge graphs. We begin by discussing some knowledge graph summarization approaches, which aim to distill large-scale graphs into concise and meaningful summaries, aiding in knowledge extraction and decision-making processes. Additionally, we explore the concept of semantic maps, which provide a visual depiction of the relationships and interdependencies among entities in a knowledge graph. We further examine semantic similarity measures, which quantify the relatedness between entities based on their semantic attributes or contextual information. We then go through centroid-based clustering algorithms, which leverage the concept of centroids to group similar entities within knowledge graphs. Lastly, we describe the significance of visual data exploration techniques on knowledge graphs and how semantic maps of knowledge graphs can be.

A. KNOWLEDGE GRAPH SUMMARIZATION

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 [22]. Recently, it has been proposed to summarize large graphs in order to enable an efficient visualization of their content. For example, in [35], the authors focus on summarizing KGs by taking advantage of individual interests to generate personalized knowledge graph summaries. In [36], 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 [37], 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 [38], transforms geo-tagged social media data into high-dimensional vectors by utilizing a doc2vec model.

Every summarization strategy depends on selecting an interest criteria to extract meaningful information [22]. 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 [37] 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 [38], 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 [35], the criteria to decide which information is *interesting* for each user is determined by reviewing the users' query history. The work of Tasnmin et al. propose a strategy to find equivalent entities in a KG using the context of each RDF Molecule [40]. In terms of summary quality, Riondato et al. [41] have proposed two quality metrics (ℓ_p-reconstruction error and cut-norm error) primarily focused on determining the error generated in the adjacency matrix from summarized graph. However, these metrics have not been proven with summarization techniques associated with KGs. The semantic mapping process described in this document can be perceived as a summarization strategy that is aimed at minimizing the semantic distance between each pair of entity instances in the KG as the criterion of interest.

B. SEMANTIC MAPS

A semantic map is a type of graphical representation that shows the relationships between different concepts or words within a particular domain or field of study [26]. The purpose of a semantic map is to visually organize and display the meaning and connections between various terms or concepts, highlighting their semantic similarities and differences. In other words, a semantic map provides a visual representation of how different ideas or concepts are related to each other and how they are grouped together based on their shared meanings or semantic properties. However, there are another mathematical representations for semantic maps such as graph and Euclidean space [42]. Figure 2 shows an example of a semantic map describing the topic *Water*. It contains three node categories: (1) the central word (root), (2) the set of keywords (e.g., Usages, Living things, etc.), and (3) the vocabulary associated with each keyword, for instance, words `Cooking` and `Bathing` are associated with keyword *Usages*.

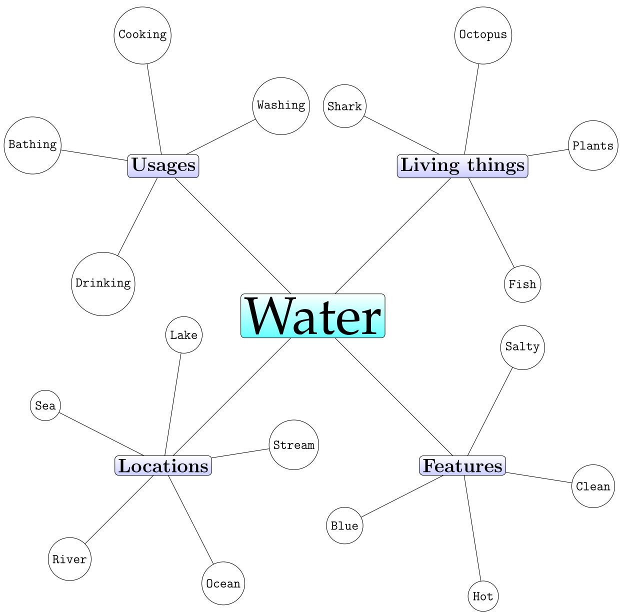

FIGURE 2. **Example of a semantic map of concepts and vocabulary associated with topics *Water*.**

Definition 2 (Semantic Map): Let C the set of groups of related words in the vocabulary, C_i the $i-th$ group of related words, kw the set of keywords representing each group of related words, and α the main subject of the vocabulary. Let us define a semantic map as a tree $T = (\alpha, E_T)$, where α represents the root of this tree, and E_T contains the set of edges that connects all nodes of the semantic map. E_T is defined as follows: $E_T = (E_{kw} \cup E_w)$, where $E_{kw} = (\alpha \times kw)$, and $E_w = (kw_i \times w_i)\forall_{kw_i \in k}, \forall_{w_i \in C_i}$.

C. SEMANTIC SIMILARITY

The semantic similarity is a metric used in Natural Processing Language (NPL) and Information Retrieval (IR) areas [43] that represents how related are two concepts based on their hierarchical relations [44], [45]. 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 [46]. 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 [47], and latent semantic analysis [48]. In contrast, knowledge-based similarity metrics quantify the degree to which two words are semantically related [49]. 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 Hulpuş et al. [50], Wu and Palmer [51], and Leacock and Chodorow [52]. Other knowledge-based measures utilize the Information Content (IC) metric like Lin [53], Jiang and Conrath [54], and Resnik [44] 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 [55] or Zhou et al. [56] combine IC and some other metrics to compute how related two words are. For instance, graph-based IC [55] uses a SPARQL query on DBPedia to compute $freq_{graph}(c_i)$ and N values in 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 as type the concept 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 [57]. Types of clustering include Centroid-based, Density-based, Distribution-based, and Hierarchical clustering [60].

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 [61]. A variation of K-means is the PAM (Partitioning Around Medoids) algorithm that minimizes dissimilarities between points in a cluster and the centroids [62]. The CLARA (Clustering Large Applications) algorithm is an extension of PAM for large datasets [63]. On the other hand, CLARANS (Clustering Large Applications based on RANdomized Search) is a partitioning algorithm focused on spatial data mining because it recognizes patterns and relationships existing in spatial data such as topological data [64]. 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 [65].

E. CLUSTER QUALITY

Literature offers two classes of clustering validation measures: external clustering validation and internal clustering validation [58]. 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 [62], Davies and Bouldin score [66], and Calinski-Harabasz Index [67].

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 [62]. 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 [61]. 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 [68]. Another known quality measure is the Davies and Bouldin index which measures the similarity between each cluster and its closest neighboring cluster, while also considering the cluster's internal similarity [66]. Finally, Calinski and Harabasz index measures the ratio of between-cluster variance to within-cluster variance [67].

F. 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 from the analyzed phenomena [23]. This process, also known as the *information seeking mantra*, follows three steps: overview, zoom and filter, and details-on-demand [69]. In this context, the visual representation of KGs offered by semantic maps proposed in this paper, provides a new way to visualize KGs. It does so by emphasizing the semantic closeness of their entity nodes. Recent works [36], [38], [70] have shown that visualizing a simplified version of a large graph is an adequate alternative.

In regard to visual exploration of KGs, challenges include context adaptation, users input [35], data heterogeneity [36], [70], 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 to compute clusters of related entities.

Recent applications have proven useful for large graph visualizations to understand different phenomena, such as Bitcoin transactions [71] and online discussions [72]. 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]. Actually, Consalvi et al. [73] propose a self-contained system to compute interactive visualizations of thousands elements in a mobile browser.

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, ReGraph, and KronoGraph[4] that offer the capability to render KGs to support tasks in areas like pharmacy and bio-science 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[5] and RDF grapher.[6] The main limitation with these tools is the small amount of data they can process.

IV. 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 the online RDF graph visualizer,[7]

[3] https://datagraphs.com (Last visited: 2023-11-16)

[4] https://cambridge-intelligence.com/ (Last visited: 2023-11-16)

[5] https://issemantic.net/rdf- visualizer (Last visited: 2023-11-16)

[6] https://www.ldf.fi/service/rdf-grapher (Last visited: 2023-11-16)

[7] https://issemantic.net/rdf-visualizer (Last visited: 2023-11-16)

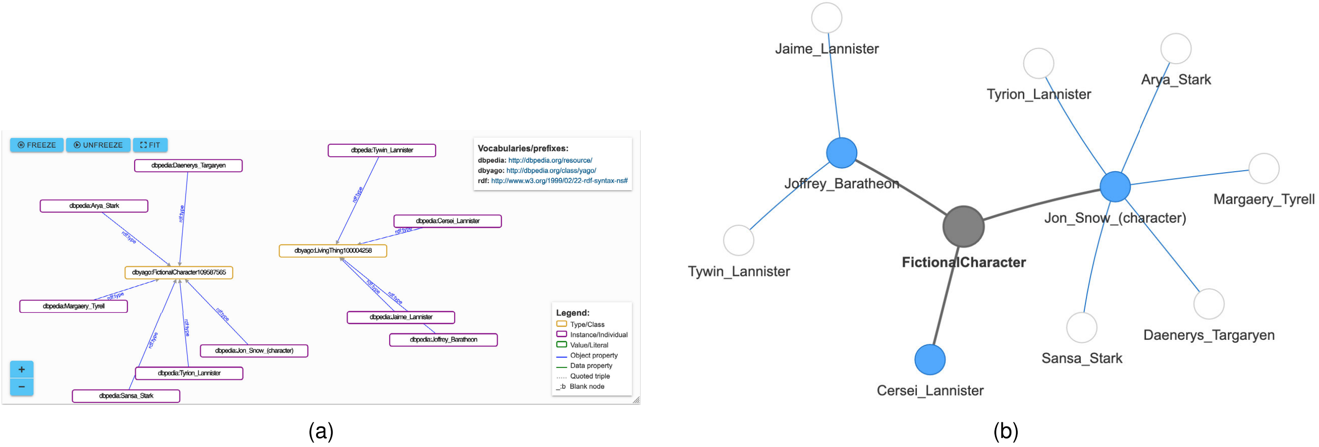

FIGURE 3. Visual representation from a small KG containing some fictional characters by George R.R. Martin. a) Contains the visual representation produced by the online RDF visualizer. b) Inferred semantic map of the original KG.

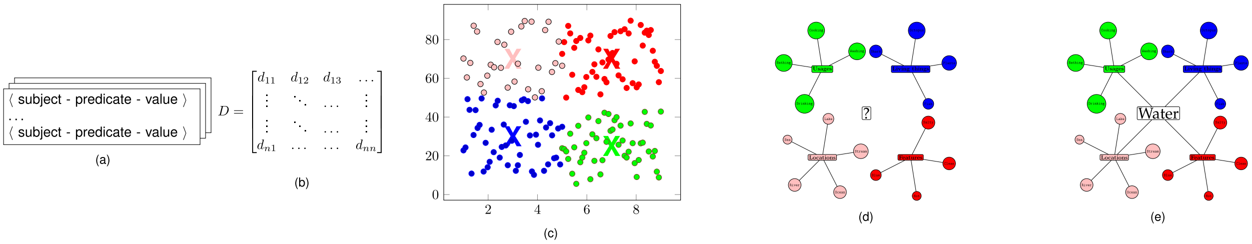
⟨ subject - predicate - value ⟩
...
⟨ subject - predicate - value ⟩
(a)
$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & \dots \\ \vdots & \ddots & \dots & \vdots \\ \vdots & \ddots & \dots & \vdots \\ d_{n1} & \dots & \dots & d_{nn} \end{bmatrix}$
(b)
80
60
40
20
0
2
4
6
8
(c)
Cooking
Washing
Bathing
Usages
Drinking
Shark
Octopus
Living things
Plants
Fish
?
Lake
Sea
Stream
Locations
River
Ocean
Salty
Features
Clean
Blue
Hot
(d)
Water
(e)

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.

which is an online tool that allows easy to visualization and analysis RDF datasets. However, this kind of visualization is not visually informative or visually appealing which may lead in an ineffective exploratory visual analysis. Contrarily Figure 3b) exhibits a semantic map of the KG. This semantic map uses a node-link representation to display the set of entity instances of the KG connected with their corresponding centroid and all centroids connected with the central concept of the semantic map.

A. EXTRACTING SEMANTIC DISTANCE OF ENTITIES IN A 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 in this phase is to extract numeric data from the KG and generate a set of entity group. We propose computing the semantic distance for each pair of entities in the KG by generating a semantic distance matrix.

Definition 3 (Semantic Distance Matrix): Given a Knowledge Graph $K = (V, L, E)$, and $sim(e_1, e_2)$ the semantic similarity between entity instances e_1 and e_2, the semantic similarity matrix $D(K)$ represents the semantic distance between each pair of entity instances in E. Specifically, the value for cell $d_{i,j} = 1 - sim(e_i, e_j)$.

Algorithm 1 describes the process to compute the distance semantic matrix D. The algorithm begins by generating the set of triples that represent the KG from the set of edges E. For each edge $e \in E$, the subject, property label, and target entity instance are extracted using the functions $subject()$, $property()$, and $value()$, respectively. Subsequently, for each pair of triples $t_i, t_j \in T$, the semantic similarity is computed using the $sim()$.

The computational complexity of the function $sim()$ is not detailed by original authors in [74], however, we can provide a brief discussion on the computational complexity based on the provided description of the algorithm and the available code of $sim()$.[8] The function $sim()$ receives as input parameters the reference of two entities in the YAGO KG, let's call them e_1 and e_2. It consists of five phases:

1) Extracting concepts from the YAGO KG,
2) Mapping concepts to synsets,
3) Calculating the IC metric,
4) Calculating scores for synsets, and
5) Obtaining the final score for the given entities.

The filtering step takes linear time, specifically $O(N_1 + N_2)$, where N_1 and N_2 represent the number of concepts associated with entities e_1 and e_2, respectively. Similarly, mapping concepts to synsets for each entity takes linear time, $O(N_1 + N_2)$, for each entity. Calculating the IC for each synset and finding the most common synsets also takes linear time, $O(N_1 + N_2)$. The comparison and score calculation involve nested loops. In the worst case, there are N_1 iterations for the

[8] https://github.com/gsi-upm/sematch (Last visited: 2023-11-16)

Algorithm 1 Algorithm to Build the Semantic Distance Matrix

```
Input: Set of edges $E$ of K
Output: Semantic distance matrix $D$
$T = \emptyset$
for all $e \in E$ do
    $T = T \bigcup \{(subject(e), property(e), value(e))\}$
end for
for all $(t_i, t_j) \in T \times T$ do
    if $t_i = t_j$ then
        $D(i, j) = 0$
    else
        $e_a = subject(t_i)$
        $e_b = subject(t_j)$
        $D(i, j) = D(j, i) = 1 - sim(e_a, e_b)$
    end if
end for
return $D$
```

score of e_1 and N_2 iterations for the score of entity e_2. This results in a time complexity of $O(N_1 \cdot N_2)$. The final score calculation is a constant-time operation, $O(1)$. Overall, the most time-consuming part of the code is the nested loop for comparing synsets and calculating scores, which results in a time complexity of $O(N_1 \cdot N_2)$.

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. The semantic distance between each entity and itself is 0. Our proposal consist of using a centroid-based clustering algorithm and generate a non-overlapping set of clusters by using the semantic distance matrix D as input of the selected clustering algorithm.

B. CLUSTERING ENTITIES OF KNOWLEDGE GRAPHS

One of the central needs of semantic maps is to find specific nodes in the KG that represent each group of entity instances. Center-based clustering algorithms seem to be suitable for this purpose. We propose using PAM and Affinity Propagation center-based clustering algorithms due to their compatibility with handling distance matrices as input instead of feature vectors [62]. PAM is a clustering algorithm that works by iteratively selecting a set of k medoids from the data points and assigning each non-medoid point to its closest medoid. The algorithm tries to minimize the sum of distances between each data point and its assigned medoid. On the other hand, Affinity Propagation is a clustering algorithm that works by propagating messages between data points to determine which points should be exemplars (i.e., representatives of their clusters).

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 defined in the Equation 2.

$$\mu = \bigcup_{C_i \in C} centroid(C_i). \tag{2}$$

C. 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 [57] as the keywords of a KG. Therefore, we denote these keywords as the set of centroids μ of the entities in a KG, where each μ_i represents the centroid of i-th cluster C_i, i.e., $\mu_i = centroid(C_i)$.

To infer the central term α, we propose to compute the IC_{graph} measure for all types associated with each centroid in μ. 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 μ This definition is formally described in equation 3.

$$\mathcal{T} = \bigcap_{\mu_i \in \mu} types(\mu_i) \tag{3}$$

Definition 4 (Central Concept α): Given a set of shared types $\mathcal{T}$, the central concept α of K is the concept $c_i \in \mathcal{T}$ with maximum IC_{graph}.

Algorithm 2 Algorithm to Infer Main Term α of K

```
Input: $\mu$: Set of centroids of $C$.
Output: $\alpha$: Main term of $K$
$\mathcal{T} = types(\mu_0)$
for all $\mu_i \in \mu - \mu_0$ do
    $\mathcal{T} = \mathcal{T} \cap types(\mu_i)$
end for
$\alpha = \max_{t \in \mathcal{T}} IC_{graph} t$
return $\alpha$
```

Algorithm 2 formalizes the process of inferring the main term of K by initializing the set of shared types, denoted as $\mathcal{T}$, with the types associated with the centroid of cluster C_0. The computational complexity of $IC_{graph}()$ (See eq. 1) depends on the complexity of building the set of entities having as type the concept c_i, denoted as $\mathcal{E}(c_i)$. To achieve this, we need to traverse the entire KG, which has a time complexity of $O(N)$, where N is the number of nodes in the KG. For each node, we must retrieve the list of types and search for the concept c_i. Assuming that the complexity of obtaining the list of types is $O(t)$, where t is the number of types associated with concept c_i, the overall complexity of $IC_{graph}()$ is $O(N \cdot t)$.

D. SEMANTIC MAP OF A KNOWLEDGE GRAPH

Let us define the semantic map of a KG:

Definition 5 (Semantic Map of a Knowledge Graph): Given a Knowledge Graph $K = (V, L, E)$, a semantic distance matrix $D(K)$, the main term of K: α, the semantic map of K is defined as $\mathcal{SM}(K) = (\alpha, E_K)$.

Table 1 describes the symbols associated with semantic maps of KGs.

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

Symbol	Description
α	Main concept of the semantic map.
μ	Set of centroid entities produced by a centroid-based clustering algorithm, where $\mu \subseteq V$.
C	Set of clusters resulting from running a centroid-based algorithm.
$\mathcal{NC}$	Non-centroids entities in KG, where $\mathcal{NC} \subseteq V$ and $\mathcal{NC} \cap \mu = \emptyset$.
E_{nc}	Set of edges connecting all members of the clusters with their corresponsing centroid, defined as $\mu_i \times x, \forall x \in \mathcal{NC}$ and $\forall \mu_i \in \mu$.
E_μ	Set of edges connecting all centroids with the main term α, definded as $E_\mu = \mu_i \times \alpha, \forall \mu_i \in \mu$.
E_K	Set of edges connecting each all elements in the semantic map, defined as $E_K = E_\mu \bigcup E_{nc}$.

Semantic mapping process aggregates the process of clustering entities of KG and inferring the central term α. Algorithm 3 describes the process to build the semantic map of a KG. Figure 4 visually describes the phases of semantic mapping process. Algorithm 3 consists of four main steps: building the semantic distance matrix, computing the clusters, inferring the main term, and creating the semantic map. The complexity of each step depends on the number of triples in the input KG and the number of clusters. The semantic distance matrix is a square matrix of size $n \times n$, where n is the number of triples in the input KG. The complexity of building this matrix involves performing $n \times n$ calls to the function *sim*(), which has a computational complexity of $O(N_1 \cdot N_2)$. However, in the worst case, N_1 and N_2 can be equal to n. Consequently, the complexity of building the semantic distance matrix is $O(n^4)$. The complexity of computing the clusters varies depending on the clustering algorithm used. For example, Affinity Propagation has a complexity of $O(n^2 \cdot T)$, where T is the number of iterations, while PAM has a complexity of $O(k \cdot (n-k)^2)$, where k is the number of clusters. The complexity of inferring the main term α has a time complexity of $O(N \cdot t)$, where N is the number of nodes in the entire KG ($N \gg n$). The complexity of creating the semantic map is $O(n+k)$. This involves creating edges between each node and its centroid, and between each centroid and the main term. Therefore, the overall complexity of the generating semantic maps of a KG is dominated by the process of inferring the main term α, which is $O(N \cdot t)$.

Algorithm 3 Process of Building a Semantic Map of a KG

```
Input: $E$: Edges associated with the KG
Output: $\mathcal{SM}(\mathcal{K})$
$D = buildSemanticDistanceMatrix(E)$
$C, \mu = computeClusters(D)$
$\alpha = inferMainTerm(\mu)$
Initialize set $\mathcal{SM} = \emptyset$
for all $C_i \in C$ do
    $\mu_i = centroid(C_i)$
    for all $x \in C_i$ do
        $E_{nc} = E_{nc} \cup create\_edge(x, \mu_i)$
    end for
end for
for all $\mu_i \in \mu$ do
    $E_\mu = E_\mu \cup create\_edge(\mu_i, \alpha)$
end for
$E_K = E_\mu \bigcup E_{nc}$
$\mathcal{SM}(\mathcal{K}) = (\alpha, E_K)$
return $\mathcal{SM}$
```

Our proposed method offers a novel approach for generating a semantic map of a KG using the results of a clustering algorithm applied to nodes in the KG based on their semantic distance. By leveraging the inherent semantic relationships among the nodes, our method enables the creation of a map that captures and visualizes the underlying semantic structure of the data. This approach provides a valuable tool for gaining insights into complex datasets and facilitating knowledge discovery.

V. EVALUATION STUDY

The experimental results section aims to demonstrate the effectiveness of our proposed method for generating semantic maps and how these maps can be used to visualize KGs. We begin by describing the Python framework developed for testing our approach. Next, we discuss the datasets used, which are the result of a set of SPARQL queries. We then detail the process of selecting hyperparameter values for the PAM and Affinity Propagation algorithms. Afterward, we focus on the quality assessment of semantic maps of KGs, emphasizing the quantitative evaluation of clusters generated by the algorithms. Furthermore, we present the centroid-based inference method for term α. Additionally, we describe the survey we conducted to validate the effectiveness of semantic maps in completing some visual data exploration tasks on a KG. Finally, we provide a comprehensive analysis of the obtained results, discussing the implications and significance of our findings

A. SEMANTIC MAPPING FRAMEWORK

Experiments are executed using a framework[9] implemented using Python 3 language which depends on the *Sematch*

[9] https://github.com/pcamarillor/semantic_mapping (Last visited: 2023-11-16)

TABLE 2. Dataset summary.

Dataset	Description	Number of triples
SCI-FI-MOVIES.NT	List of triples describing sci-fi movies with a gross greater than eight billion of dollars.	188
FANTASY-NOVELS.NT	This dataset contains a set of triples describing fantasy novels published after year 2000.	693
CITIES.NT	Collection of triples describing cities with a total population greater than five millions.	127
DISEASES.NT	List of triples that enumerates infectious diseases.	36
DRUGS.NT	List that contains triples of medicines associated with infectious diseases.	54
ACTORS.NT	This collection of triples contains actors starring american sci-fsi movies.	166
MOVIES-AND-ACTORS.NT	This dataset combines a subset of `SCI-FI-MOVIES.NT` and `ACTORS.NT` datasets.	72
DISEASES-AND-DRUGS.NT	This collections of triples combining selected triples from `DISEASES.NT` and `DRUGS.NT`.	50

```
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.

framework [74] to perform SPARQL queries to DBPedia public endpoint and compute the similarity measure used to compute D. The function $sim(e_i, e_j)$ mentioned in Algorithm 1 is implemented through a SPARQL query to DBPedia. Once generated D, our tool produces the set of centroids μ and the set of non-centroid $\mathcal{NC}$ nodes by using centroid-based clustering strategies: PAM and Affinity Propagation. We infer the main term α by implementing the Algorithm 2. These shared types are the result of a SPARQL query to DBPedia that follows the path shown in Figure 5. Finally, our tool, assembles the semantic map by implementing Algorithm 3 and generating maps using `pyvis` library which is a wrapper around the Javascript `visJS` library.

B. DATASETS

Datasets used to validate the semantic mapping building process are the result of performing a SPARQL query to DBPedia KG through its public endpoint[10] and results are saved in N-Triples format, 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. Table 2 contains a summary of datasets used for experiments. Detailed SPARQL queries used to generate these datasets are described in the repository that contains the framework developed to run these experiments.

C. HYPERPARAMETERS SELECTION

A key hyperparameter in the PAM clustering algorithm is the number of clusters we want to generate (k). Determining this hyperparameter is a crucial step in clustering and we determine this value by using the elbow method [75].[11] The *preference* parameter is a crucial hyperparameter in the Affinity Propagation clustering algorithm, which is a parameter that help to determine 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 3 describes the selection of *preference* and k hyperparameters for Affinity Propagation and PAM algorithms, respectively.

D. QUALITY OF SEMANTIC MAPS

The core of the semantic mapping process is to cluster the entity instances and obtain the set of centroids μ. 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.

Table 4 includes three columns describing the semantic map quality for two centroid-based clustering algorithms

[10] https://dbpedia.org/sparql/ (Last visited: 2023-11-16)

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

TABLE 3. Hyperparameter selection. The PAM algorithm uses the elbow method to determine the optimal number of clusters k. The elbow method uses the WSS metric, which decreases as the distance between each data point and its assigned center becomes smaller. For Affinity Propagation, the preference value is a parameter that helps determine the number of clusters that will be generated. A higher preference value results in more clusters, while a lower preference value leads to fewer clusters.

	PAM		Affinity Propagation	
Dataset	Optimal number of cluster k	WSS	Preference	Number of generated clusters
SCI-FI-MOVIES.NT	24	0.27	0.8	5
FANTASY-NOVELS.NT	40	1.73	0.8	6
CITIES.NT	16	1.69	0.5	5
DISEASES.NT	9	0.80	0.6	3
DRUGS.NT	10	8.40	0.8	52
ACTORS.NT	15	19.57	0.1	3
MOVIES-AND-ACTORS.NT	13	4.23	0.7	4
DISEASES-AND-DRUGS.NT	10	6.40	0.7	3

TABLE 4. Quality of clusters in the entity clustering phase of semantic mapping process. The silhouette score measures how similar each entity is to its own cluster compared to other clusters, with scores closer to 1 indicating better cluster quality. The Davies-Bouldin index measures the ratio of the within-cluster scatter to the between-cluster separation, with lower scores indicating better cluster quality. The Calinski-Harabasz index measures the ratio of between-cluster variance to within-cluster variance, with higher scores indicating better cluster quality.

	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
DISEASES.NT	0.44	0.68	46.56	0.43	2.84	5.52
DRUGS.NT	0.33	1.32	11.06	-0.02	0.71	0.63
ACTORS.NT	0.40	1.26	52.87	0.13	2.44	37.63
MOVIES-AND-ACTORS.NT	0.55	0.67	103.26	0.54	1.41	85.44
DISEASES-AND-DRUGS.NT	0.54	0.94	91.39	0.42	0.57	60.25

(PAM and Affinity Propagation) in terms of silhouette score, Davies-Bouldin score, and Calinski-Harabasz index. The silhouette score measures how similar each entity is to its own cluster compared to other clusters, **with scores closer to 1** indicating better cluster quality. The Davies-Bouldin index measures the ratio of the within-cluster scatter to the between-cluster separation, **with lower scores** indicating better cluster quality. The Calinski-Harabasz index measures the ratio of between-cluster variance to within-cluster variance, **with higher scores** indicating better cluster quality.

E. INFERRED MAIN TERMS

For each experiment, semantic mapping process infers the main term α based in the IC_{graph} metric. Table 5 describes the inferred α for each dataset.

F. QUALITATIVE ASSESSMENT OF SEMANTIC MAPS

One of the goals of graph summarization is to facilitate the process of visual data exploration [22]. In this context, we propose to evaluate the effectiveness of the summarization process described in this work by measuring how well semantic maps serve as a visualization strategy for fulfilling visual exploratory tasks. A diagram that represents a knowledge graph visually is called a classical visual representation.

TABLE 5. Inferred main terms.

Dataset	**Inferred main term α**
SCI-FI-MOVIES.NT	`yago:Movie106613686`
FANTASY-NOVELS.NT	`yago:WikicatFantasyNovels`
CITIES.NT	`yago:City108524735`
DISEASES.NT	`yago:AlimentCondition` (Affinity Propagation) `yago:Disease114070360` (PAM)
DRUGS.NT	`dbo:Drug`
ACTORS.NT	`yago:WikicastActors` (Affinity Propagation) `yago:Actor109765278` (PAM)
MOVIES-AND-ACTORS.NT	`yago:Whole100003553`
DISEASES-AND-DRUGS.NT	`yago:Abstraction100002137`

In this diagram, the entities and relationships in the graph are shown as nodes and edges, respectively. The nodes are labeled with the names of the entities they represent, and the edges are labeled with the names of the relationships they represent.

TABLE 6. Effectiveness of semantic maps as strategy to visualize Knowledge Graphs.

		DISEASES-AND-DRUGS.NT		MOVIES-AND-ACTORS.NT		CITIES.NT	
		Classical Visualization	Semantic Maps	Classical Visualization	Semantic Maps	Classical Visualization	Semantic Maps
Search for one specific item	Strongly agree	8%	20%	12%	24%	8%	20%
	Agree	44%	64%	48%	40%	12%	60%
	Neutral	20%	12%	20%	12%	4%	12%
	Disagree	24%	4%	16%	24%	56%	8%
	Strongly disagree	4%	0%	4%	0%	20%	0%
Identification of main term	Strongly agree	44%	36%	32%	48%	24%	44%
	Agree	28%	24%	36%	20%	32%	40%
	Neutral	16%	28%	20%	16%	16%	12%
	Disagree	12%	12%	12%	12%	28%	0%
	Strongly disagree	0%	0%	0%	4%	0%	4%
Explore and comprehend knowledge graphs	Strongly agree	12.5%	20.8%	25%	41.7%	16.7%	41.7%
	Agree	62.5%	45.8%	54.2%	37.5%	25%	29.2%
	Neutral	16.7%	20.8%	12.5%	12.5%	25%	12.5%
	Disagree	8.3%	8.3%	8.3%	4.2%	29.1%	12.5%
	Strongly disagree	0%	4.2%	0%	4.2%	4.2%	4.2%

This diagram can be used to explore the relationships between the entities and visualize the structure of the knowledge graph. We conducted a survey that consists of the following three sections:

- The first section aims to understand the profile of the respondents.
- The second section employs a Likert scale [76] to measure the effectiveness of classical visual representations of knowledge graphs for three datasets: `DISEASES-AND-DRUGS.NT`, `MOVIES-AND-ACTORS.NT`, and `CITIES.NT`.
- The third section inquires about which method is easier to use and more effective in representing knowledge graphs.

This survey was applied to a group of 25 experts and it is fully detailed in Appendix A.

In terms of the profiles of the experts who participated in this study, we offered six different profiles associated with the fields of knowledge discovery, artificial intelligence, and data science. Participants had the option to select more than one profile. From these responses, we found that

- 88.5% mentioned having a background in computer science or a related field.
- 73.1% mentioned being familiar with graph theory.
- 65.4% expressed curiosity about exploring and discovering new insights from data.
- 53.8% were curious about natural language processing, machine learning, and semantic web technologies.
- 38.5% reported having experience in querying and manipulating data using languages such as SPARQL.
- Only 3.8% mentioned having experience in crafting knowledge models.

Regarding the frequency of using knowledge graphs in their daily duties:

- 38.5% of participants mentioned using knowledge graphs sometimes.
- 11.5% always use knowledge graphs to fulfill their daily duties.

Table 6 presents the effectiveness results using a Likert scale to evaluate the effectiveness of two different visual representations of KGs in fulfilling three exploratory tasks related to visual data. When we asked experts about which method they find easy to use, 76.9% of them selected semantic maps, while 23.1% favored classic visual representation. Explicitly inquiring about which method experts believe is more effective in representing KGs, 88.5% chose semantic maps, and 11.5% opted for classic visual representation.

G. DISCUSSION

The quality analysis of semantic maps (Table 4) demonstrates the superior performance of the PAM algorithm over the Affinity Propagation algorithm. The PAM algorithm consistently achieved higher silhouette scores for all datasets, indicating better-defined and well-separated clusters compared to the Affinity Propagation algorithm. Moreover, the Davies-Bouldin scores for the PAM algorithm indicated compact and well- separated clusters in 5 out of 8 datasets, whereas the scores for the Affinity Propagation algorithm indicated significant overlap and poor separation. The Calinski-Harabasz index further reinforced the superiority of the PAM algorithm in generating high-quality semantic maps, as its scores were significantly higher than those of the Affinity Propagation algorithm across all datasets. Consequently, the PAM algorithm emerges as the preferred choice, producing superior semantic maps with better separation.

Regarding to inferred main concepts (α), there are two particular cases to analyze. For the `DISEASES.NT` dataset, semantic map produced using the Affinity Propagation algorithm infers that the main concept is `yago:AlimentCondition`, contrarily main term inferred using the PAM algorithm is the concept `yago:Disease114070360`. Similarly, for the `ACTORS.NT` experiment, the semantic map produced by using the Affinity Propagation algorithm infers the main concept as the concept `yago:WikicastActors` but the main concept inferred when the PAM algorithm is applied to build the semantic map is the concept `yago:Actor109765278`. The reason of this difference between the main concepts inferred is that each clustering algorithm produces different set of centroid elements (μ), which affects directly the inference of the main concept.

Hybrid datasets `MOVIES-AND-ACTORS.NT` and `DISEASES-AND-DRUGS.NT` are used to validate the process to infer the term α when the instances inside datasets comes from different classes, however our study yielded a particularly intriguing finding in the resulting semantic maps of these hybrid datasets. Resulting clusters exhibit a high level of coherence and meaningfulness.

In our effectiveness survey, we delved into three exploratory tasks related to KGs. We examined the effectiveness of semantic maps and classical visual representation across different datasets. Let us explore the key findings in detail. For the first exploratory task (Search for one specific item), semantic maps consistently garnered strong support. Specifically, for the `DISEASES-AND-DRUGS.NT` dataset, 20% of participants strongly agreed that semantic maps effectively located requested item. In the context of the `MOVIES-AND-ACTORS.NT` dataset, 24% of participants found semantic maps useful for search tasks. Meanwhile, classical visual representation received 8% and 12% endorsement for the same tasks in the two datasets, respectively. For the `CITIES.NT` dataset, 20% of participants favored semantic maps, while only 8% found classical visualization helpful. Regarding to second exploratory task aimed to identify the main term of the KG, for the `DISEASES-AND-DRUGS.NT` dataset, 44% of participants found classical visual representation effective in identifying the main topic, while 36% favored semantic maps. In the context of the `MOVIES-AND-ACTORS.NT` dataset, 48% of participants preferred semantic maps, whereas 32% relied on classical visual representation for this task. For the `CITIES.NT` dataset, 44% of participants leaned toward semantic maps, while 24% opted for classical visual representation in identifying main terms. These results underscore the varied utility of semantic mapping and visual representation in identifying the main topic of a certain RDF vocabulary. Lastly, for the third task associated with the process of exploration and comprehension of knowledge graphs we found that for `DISEASES-AND-DRUGS.NT` dataset, 20.8% of participants strongly agreed that semantic maps facilitated understanding. Similarly, in the context of the `MOVIES-AND-ACTORS.NT` and `CITIES.NT` datasets, 41.7% of participants endorsed semantic maps for this task. Conversely, classical visual representation received varying levels of support: 12.5%, 25%, and 16.7% for the respective datasets. These findings emphasize the role of semantic mapping in enhancing knowledge graph exploration across diverse domains.

Although our proposal presents an effective approach to summarize KGs by generating semantic maps, we acknowledge one difference from the description of semantic maps provided in Section III. In the semantic map shown in Figure 2 the centroids are concepts i.e., the centroids are members of TBox set, but the semantic mapping process we are describing in this paper, centroids are entity instances, i.e, $\mu_i \in ABox \; \forall \mu_i \in \mu$. This difference, though significant, falls beyond the scope of the current proposal to generate semantic maps.

VI. CONCLUSION AND FUTURE WORK

In conclusion, our study focused on the generation of semantic maps as a summarization strategy based on semantic similarity. Through the utilization of centroid-based clustering algorithms, specifically Affinity Propagation and PAM, we successfully captured the semantic distance between nodes in the KG and generated meaningful clusters.

Our experiments revealed a notable divergence between the two clustering algorithms. While the Affinity Propagation algorithm produced clusters with qualitative coherence and meaningfulness for hybrid datasets, the PAM algorithm excelled when evaluated using internal validation metrics. This emphasizes the importance of considering both qualitative and quantitative evaluation measures in assessing clustering quality.

Additionally, we leveraged the computed centroids to infer the main term α, resulting in visually appealing and informative representations of the KG. This inference method, outlined in Algorithm 2, facilitated a comprehensive understanding of the encoded information.

Our qualitative study demonstrates the effectiveness of the use of semantic maps as a visualization strategy for fulfilling diverse exploratory analysis tasks. For locating specific items, semantic maps consistently garnered strong support, while classical visual representation played a complementary role. Similarly, in identifying main terms, semantic maps prevailed, but classical visual representation remained relevant. Lastly, for exploring and comprehending knowledge graphs, semantic maps found favor among participants, while classical visual representation received varying levels of endorsement.

As future work, we consider evaluating the summarization technique we propose by reducing the ℓ_p-reconstruction error and the cut-norm error quality metrics proposed by Riondato et al. [41].

In summary, our work successfully integrates centroid-based clustering algorithms, qualitative evaluation, and inference methods to generate semantic maps for visualizing

KGs. This approach offers a comprehensive understanding of the data, combining qualitative and quantitative assessments. We believe that our findings significantly contribute to the field, enabling researchers and practitioners to effectively visualize and analyze complex KGs with improved clarity and interpretability.

APPENDIX A
SURVEY ON EFFECTIVENESS OF KNOWLEDGE GRAPH REPRESENTATIONS

This appendix describes the survey we conducted as part of the qualitative assessment of the visual representation offered by semantic maps. For questions 3 to 8, each item could be answered by selecting one option from the choices: *Strongly agree, Agree, Neutral, Disagree*, and *Strongly disagree* choices. Figures shown in this appendix are scaled versions of the actual pictures that were exposed to participants.

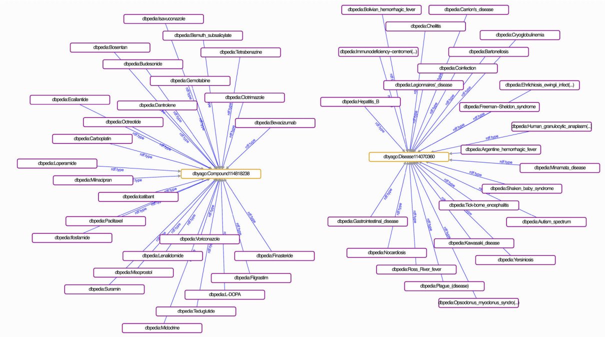

FIGURE 6. Dataset `DISEASES-AND-DRUGS.NT` **represented using RDF Visualizer online tool.**

1) Select the description that matches your professional or academic profile.
   - I have a background in data science, computer science, information science, or a related field.
   - I am familiar with graph theory, graph databases, and graph algorithms.
   - I have experience in querying and manipulating data using languages such as SPARQL, Cypher, or Gremlin.
   - I have expertise in crafting knowledge models using standards like RDF, OWL, or Schema.org.
   - I am curious of natural language processing, machine learning, and semantic web technologies.
   - I have some curiosity to explore and discover new insights from data.
2) How often do you use knowledge graphs in your work or studies?
   - Never
   - Rarely
   - Sometimes
   - Often
   - Always
3) The following picture (see Figure 6) displays a classic visual representation of certain **drugs and diseases** described in Wikipedia. Please evaluate the visual representation displayed based on the following statements:
   - I was able to find easily the item `dbpedia:Ross_River_fever` using this visual representation
   - I was able to identify the central concept that represents all items in the picture
   - This representation helped me understand the relationships between different entities and concepts

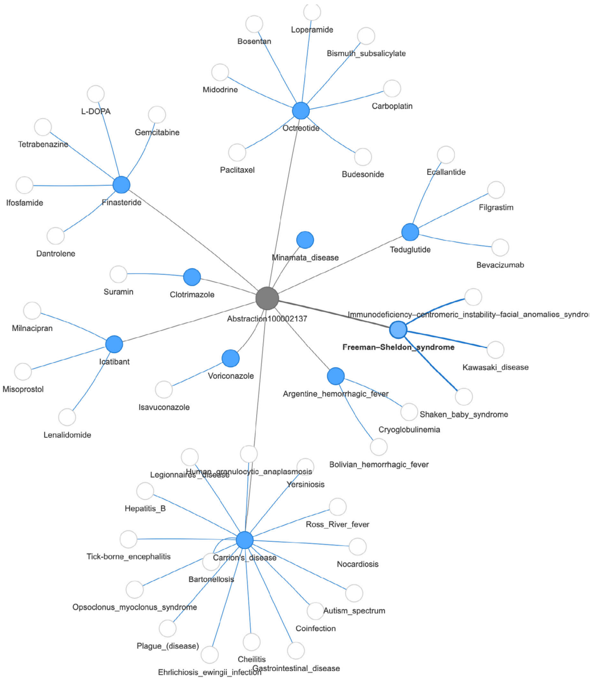

FIGURE 7. Dataset `DISEASES-AND-DRUGS.NT` **represented using the generated semantic map.**

4) The following picture (see Figure 7) displays a semantic map that summarizes certain **drugs and diseases** described in Wikipedia. Please evaluate the visual representation displayed based on the following statements::
   - I was able to find easily the item `Ross_River_fever` using this visual representation
   - I was able to identify the central concept that represents all items in the picture
   - This representation helped me understand the relationships between different entities and concepts
5) The following picture (see Figure 8) displays a classic visual representation of certain **actors and movies** described in Wikipedia. Please evaluate the visual representation displayed based on the following statements:
   - I was able to find easily the item `dbpedia:Ender's_Game_(film)` using this visual representation
   - I was able to identify the central concept that represents all items in the picture

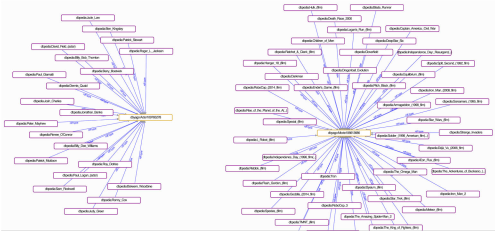

FIGURE 8. **Dataset** ACTORS-AND-MOVIES.NT **represented using RDF Visualizer online tool.**

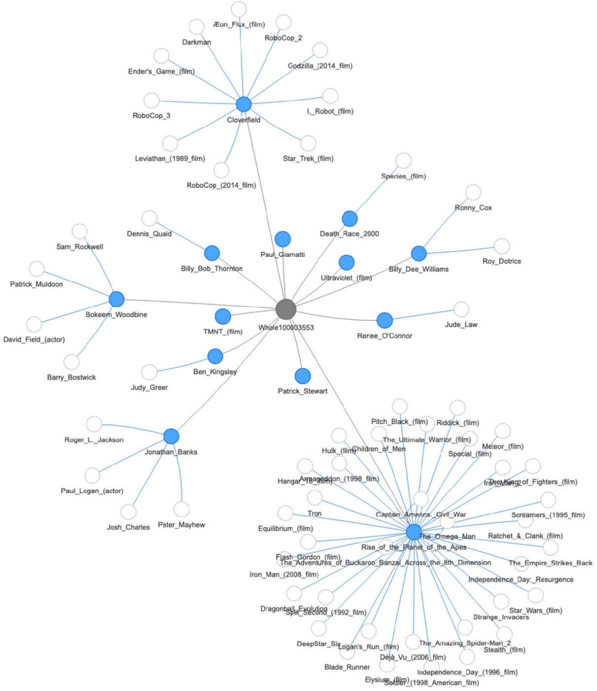

FIGURE 9. **Dataset** ACTORS-AND-MOVIES.NT **represented using the generated semantic map.**

- This representation helped me understand the relationships between different entities and concepts

6) The following picture (see Figure 9) displays a semantic map that summarizes certain **actors and movies** described in Wikipedia. Please evaluate the visual representation displayed based on the following statements::
   - I was able to find easily the item Ender's_Game_(film) using this visual representation
   - I was able to identify the central concept that represents all items in the picture
   - This representation helped me understand the relationships between different entities and concepts
7) The following picture (see Figure 10) displays a classic visual representation of certain **cities** described in Wikipedia. Please evaluate the visual representation displayed based on the following statements:
   - I was able to find easily the item dbpedia: Bogotá using this visual representation

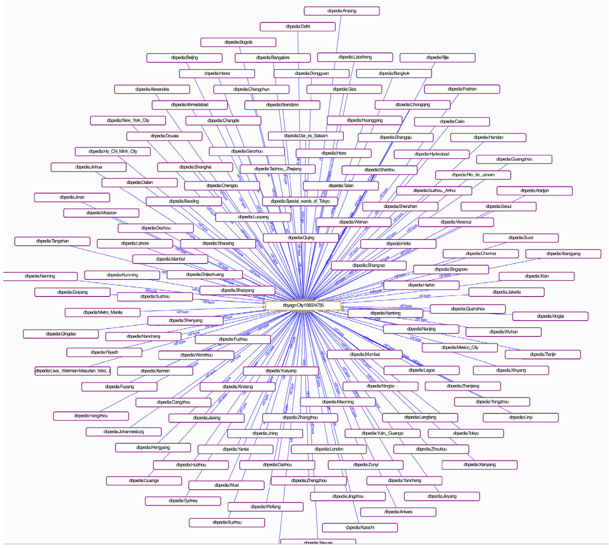

FIGURE 10. **Dataset** CITIES.NT **represented using RDF Visualizer online tool.**

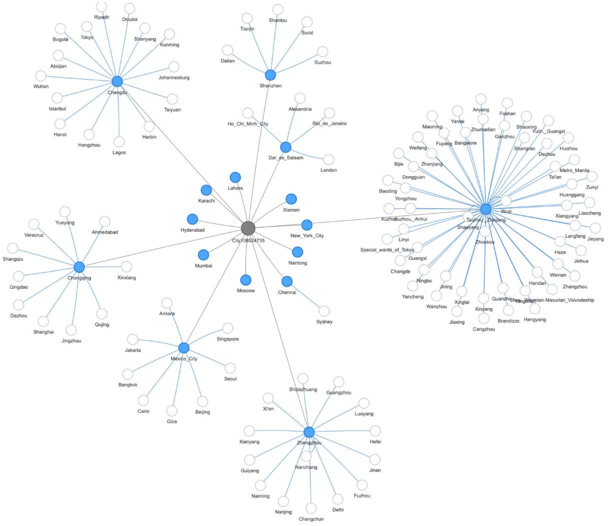

FIGURE 11. **Dataset** CITIES.NT **represented using the generated semantic map.**

   - I was able to identify the central concept that represents all items in the picture
   - This representation helped me understand the relationships between different entities and concepts
8) The following picture (see Figure 11) displays a semantic map that summarizes certain **cities** described in Wikipedia. Please evaluate the visual representation displayed based on the following statements::
   - I was able to find easily the item Bogotá using this visual representation
   - I was able to identify the central concept that represents all items in the picture
   - This representation helped me understand the relationships between different entities and concepts
9) Which method do you find easier to use?
   - Semantic Maps

   - Classic Visual Representations

10) Which method do you think is more effective in representing knowledge graphs?
    - Semantic Maps
    - Classic Visual Representations

REFERENCES

[1] H. Purohit, V. L. Shalin, and A. P. Sheth, "Knowledge graphs to empower humanity-inspired AI systems," *IEEE Internet Comput.*, vol. 24, no. 4, pp. 48–54, Jul. 2020.

[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," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, 2021.

[3] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge," in *Proc. 16th Int. Conf. World Wide Web*, Alberta, AB, Canada, May 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, Germany: Springer, 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 *Proc, 2008 ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*. 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 *Proc. 24th AAAI Conf. Artif. Intell.* Atlanta, GA, USA: AAAI Press, 2010, pp. 1306–1313.

[7] A. Singhal. (May 2012). *Introducing the Knowledge Graph: Things, Not Strings*. Accessed: Nov. 16, 2023. [Online]. Available: https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

[8] R. Qian. (Mar. 2013). *Understand Your World with Bing*. Accessed: Nov. 16, 2023. [Online]. Available: https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/

[9] E. Sun. (2013). *Under the Hood: The Entities Graph*. Meta. Accessed: Nov. 16, 2023. [Online]. Available: https://m.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*, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble, Eds. Cham, Switzerland: Springer, 2014, pp. 50–65.

[11] Y.-H. Chen, E. J. Lu, and Y.-Y. Lin, "Efficient SPARQL queries generator for question answering systems," *IEEE Access*, vol. 10, pp. 99850–99860, 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. 119290–119300, 2022.

[13] H. Li, Y. Wang, S. Zhang, Y. Song, and H. Qu, "KG4Vis: A knowledge graph-based approach for visualization recommendation," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 195–205, Jan. 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. 122670–122678, 2022.

[15] A. Borrego, D. Dessì, I. Hernández, F. Osborne, D. R. Recupero, D. Ruiz, D. Buscaldi, and E. Motta, "Completing scientific facts in knowledge graphs of research concepts," *IEEE Access*, vol. 10, pp. 125867–125880, 2022.

[16] K. Guan, L. Du, and X. Yang, "Relationship extraction and processing for knowledge graph of welding manufacturing," *IEEE Access*, vol. 10, pp. 103089–103098, 2022.

[17] X. Zou, "A survey on application of knowledge graph," *J. Phys., Conf. Ser.*, vol. 1487, no. 1, Mar. 2020, Art. no. 012016.

[18] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems," *J. Biomed. Informat.*, vol. 41, no. 5, pp. 706–716, Oct. 2008.

[19] A. Ruttenberg, J. A. Rees, M. Samwald, and M. S. Marshall, "Life sciences on the semantic Web: The neurocommons and beyond," *Briefings Bioinf.*, vol. 10, no. 2, pp. 193–204, Mar. 2009.

[20] V. Momtchev, D. Peychev, T. Primov, and G. Georgiev, "Expanding the pathway and interaction knowledge in Linked Life Data," in *Proc. Int. Semantic Web Challenge*, 2009.

[21] K. Gunaratna, "Semantics-based summarization of entities in knowledge graphs," Ph.D. dissertation, Wright State Univ., Dayton, OH, USA, 2017.

[22] Y. Liu, T. Safavi, A. Dighe, and D. Koutra, "Graph summarization methods and applications: A survey," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–34, May 2019.

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

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

[25] T. Georgakopoulos. *Semantic Maps*. Accessed: Nov. 16, 2023. [Online]. Available: https://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0229.xml

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

[27] S. E. Schaeffer, "Graph clustering," *Comp. Sci. Rev.*, vol. 1, no. 1, pp. 27–64, 2007.

[28] R. Jalota, D. Vollmers, D. Moussallem, and A. N. Ngomo, "LAUREN—Knowledge graph summarization for question answering," in *Proc. IEEE 15th Int. Conf. Semantic Comput. (ICSC)*, Laguna Hills, CA, USA, Jan. 2021, pp. 221–226.

[29] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proc. IEEE*, vol. 104, no. 1, pp. 11–33, Jan. 2016.

[30] 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 *Proc. 25th Int. Conf. Companion World Wide Web*, Montreal, CA, USA, Apr. 2016, pp. 251–254.

[31] G. Schreiber and Y. Raimond. (Jun. 2014). *RDF 1.1 Primer*. Accessed: Nov. 16, 2023. [Online]. Available: https://www.w3.org/TR/rdf11-primer/

[32] D. L. McGuinness and F. van Harmelen. (Feb. 2004). *5 Trends Appear on the Gartner Hype Cycle for Emerging Technologies, 2019*. Accessed: Nov. 16, 2023. [Online]. Available: https://www.w3.org/TR/owl-features/

[33] L. M. Garshol, "Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all," *J. Inf. Sci.*, vol. 30, no. 4, pp. 378–391, Aug. 2004.

[34] I. Horrocks, "Ontologies and the semantic Web," *Commun. ACM*, vol. 51, no. 12, pp. 58–67, 2008.

[35] T. Safavi, C. Belth, L. Faber, D. Mottin, E. Müller, and D. Koutra, "Personalized knowledge graph summarization: From the cloud to your pocket," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Beijing, China, Nov. 2019, pp. 528–537.

[36] Z. Shen, K.-L. Ma, and T. Eliassi-Rad, "Visual analysis of large heterogeneous social networks by semantic and structural abstraction," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, no. 6, pp. 1427–1439, Nov. 2006.

[37] D. Koutra, U. Kang, J. Vreeken, and C. Faloutsos, "VOG: Summarizing and understanding large graphs," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, Apr. 2014, pp. 91–99.

[38] 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. 114851–114861, 2019.

[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 Bioinf.*, vol. 13, no. S3, p. S10, Dec. 2012.

[40] M. Tasnim, D. Collarana, D. Graux, and M.-E. Vidal, "Context-based entity matching for big data," in *Knowledge Graphs and Big Data Processing*. Cham, Switzerland: Springer, 2020, ch. 8.

[41] M. Riondato, D. García-Soriano, and F. Bonchi, "Graph summarization with quality guarantees," in *Proc. IEEE Int. Conf. Data Mining*, Shenzhen, China, Dec. 2014, pp. 947–952.

[42] W. Croft, "On two mathematical representations for 'semantic maps,'" *Zeitschrift Für Sprachwissenschaft*, vol. 41, no. 1, pp. 67–87, Jun. 2022.

[43] E. Hovy, R. Navigli, and S. P. Ponzetto, "Collaboratively built semi-structured content and artificial intelligence: The story so far," *Artif. Intell.*, vol. 194, pp. 2–27, Jan. 2013.

[44] Y. Bin, L. Xiao-Ran, L. Ning, and Y. Yue-Song, "Using information content to evaluate semantic similarity on HowNet," in *Proc. 8th Int. Conf. Comput. Intell. Secur.* San Mateo, CA, USA: Morgan Kaufmann, Nov. 2012, pp. 448–453.

[45] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artif. Intell. Res.*, vol. 37, pp. 141–188, Feb. 2010.

[46] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proc. AAAI*, vol. 6, 2006, pp. 775–780.

[47] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguistics*, vol. 16, no. 1, pp. 22–29, Mar. 1990.

[48] 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," *Psychol. Rev.*, vol. 104, no. 2, pp. 211–240, Apr. 1997.

[49] A. Budanitsky and G. Hirst, "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures," in *Proc. Workshop WordNet Other Lexical Resour.*, vol. 2, 2001, p. 2.

[50] 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 Int. Semantic Web Conf.* Bethlehem, PA, USA: Springer, 2015, pp. 442–457.

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

[52] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," *WordNet, Electron. Lexical Database*, vol. 49, no. 2, pp. 265–283, 1998.

[53] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*, Madison, WI, USA, 1998, pp. 296–304.

[54] 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.

[55] G. Zhu and C. A. Iglesias, "Computing semantic similarity of concepts in knowledge graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 72–85, Jan. 2017.

[56] Z. Zhou, Y. Wang, and J. Gu, "New model of semantic similarity measuring in wordnet," in *Proc. 3rd Int. Conf. Intell. Syst. Knowl. Eng.*, Xiamen, China, Nov. 2008, pp. 256–261.

[57] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, Jun. 2015.

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

[59] S. Firdaus and M. A. Uddin, "A survey on clustering algorithms and complexity analysis," *Int. J. Comput. Sci. Issues (IJCSI)*, vol. 12, no. 2, p. 62, 2015.

[60] (Jul. 2022). *Clustering Algorithms*. Accessed: Nov. 16, 2023. [Online]. Available: https://developers.google.com/machine-learning/clustering/clustering-algorithms

[61] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1. Berkeley, CA, USA: Regents of the University of California, Jan. 1967, pp. 281–297.

[62] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program PAM)," in *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 1990, ch. 2.

[63] L. Kaufman and P. J. Rousseeuw, "Clustering large applications (program CLARA)," in *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 1990, ch. 3.

[64] R. T. Ng and J. Han, "CLARANS: A method for clustering objects for spatial data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 5, pp. 1003–1016, Sep. 2002.

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

[66] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[67] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist. Simul. Comput.*, vol. 3, no. 1, pp. 1–27, 1974.

[68] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, Jan. 1974.

[69] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proc. IEEE Symp. Vis. Lang.*, Boulder, CO, USA, Sep. 1996, pp. 336–343.

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

[71] 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, Jun. 2016.

[72] M. Molina-Solana, D. Birch, and Y.-K. Guo, "Improving data exploration in graphs with fuzzy logic and large-scale visualisation," *Appl. Soft Comput.*, vol. 53, pp. 227–235, Apr. 2017.

[73] L. Consalvi, W. Didimo, G. Liotta, and F. Montecchiani, "BrowVis: Visualizing large graphs in the browser," *IEEE Access*, vol. 10, pp. 115776–115786, 2022.

[74] G. Zhu and C. A. Iglesias, "Sematch: Semantic similarity framework for knowledge graphs," *Knowl.-Based Syst.*, vol. 130, pp. 30–32, Aug. 2017.

[75] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, Dec. 1953.

[76] R. Likert, *Arch. Psychol.*, vol. 22, no. 1932, pp. 5–55, 1932.

PABLO CAMARILLO-RAMIREZ received the M.S. degree in computer science from the Meritorious Autonomous University of Puebla, in 2014. He is currently pursuing the Ph.D. degree in engineering sciences with the Western Institute of Technology and Higher Education, Guadalajara, Mexico. He is working for Fair Isaac Corporation, as a Software Engineer. His research interests include semantic web, distributed systems, and big data.

FRANCISCO CERVANTES-ALVAREZ received the Ph.D. degree in computer science from the University of Grenoble Alpes, in 2016. Since 2017, he has been a member of the National System of Researchers, CONACYT. He is currently a full-time Professor and a Researcher with the Department of Electronics, Systems and Computer Science, Western Institute of Technology and Higher Education, Guadalajara, Mexico. He is the Coordinator of the Ph.D. Program in engineering sciences, from 2021 to 2025. His research interests include machine learning, multi-agent systems, and graphs. He has several publications in journal and international conferences on these topics.

LUIS FERNANDO GUTIÉRREZ-PRECIADO received the B.E. degree in computer systems from the Autonomous University of Aguascalientes, Mexico, in 2007, and the M.Sc. degree in computer science and the Ph.D. degree from CINVESTAV Guadalajara, in 2009 and 2013, respectively. He is currently a Professor with the Department of Electronics, Systems, and Informatics, Western Institute of Technology and Higher Education. His research interests include graph analytics and self-organizing networks.