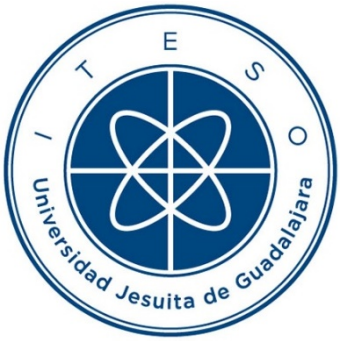
INSTITUTO TECNOLÓGICO
Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación el 29 de noviembre de 1976.

Departamento de Electrónica, Sistemas e Informática

Doctorado en Ciencias de la Ingeniería



EXPLORACIÓN DE LOS MAPAS SEMÁNTICOS COMO MEDIO PARA RESUMIR Y VISUALIZAR GRAFOS DE CONOCIMIENTO

Tesis que para obtener el grado de
Doctor en Ciencias de la Ingeniería
presenta: Pablo Camarillo-Ramirez

Director de tesis: Dr. José Francisco Cervantes-Alvarez
Co-director de tesis: Dr. Luis Fernando Gutierrez-Preciado

Tlaquepaque, Jalisco. Junio de 2024

TITULO: **Exploración de los mapas semánticos como medio para resumir y visualizar grafos de conocimiento**

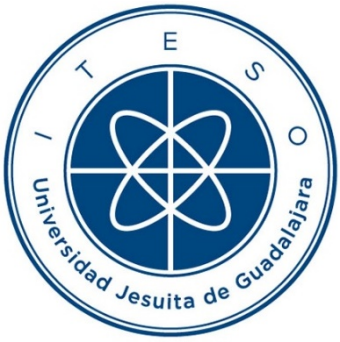
AUTOR: Pablo Camarillo-Ramirez
Ingeniero en Ciencias de la Computación (BUAP, México)
Maestro en Ciencias de la Computación (BUAP, México)

DIRECTOR DE TESIS: José Francisco Cervantes-Alvarez
Departamento de Electrónica, Sistemas e Informática, ITESO
Ingenierio en Sistemas Computacionales (Instituto Tecnológico de Jiquilpan, México)
Maestro en Ciencias de la Computación (CENIDET, México)
Doctor en Ciencias de la Computación (Universidad de Grenoble-Alpes, Francia)

NÚMERO DE PÁGINAS: XXI, 89

ITESO – The Jesuit University of Guadalajara

Department of Electronics, Systems and Informatics
Doctoral Program in Engineering Sciences



EXPLORING SEMANTIC MAPS AS A MEANS OF SUMMARIZING AND VISUALIZING KNOWLEDGE GRAPHS

Thesis to obtain the degree of
Doctor in Engineering Sciences
Presents: Pablo Camarillo-Ramirez

Thesis Director: Dr. José Francisco Cervantes-Alvarez
Thesis Co-director: Dr. Luis Fernando Gutierrez-Preciado

Tlaquepaque, Jalisco, Mexico

June 2024

TITLE: **Exploring Semantic Maps as a Means of Summarizing and Visualizing Knowledge Graphs**

AUTHOR: Pablo Camarillo-Ramirez
Bachelor's degree in Computer Science Engineering (BUAP, Mexico)
Master's degree in Computer Science (BUAP, Mexico)

THESIS DIRECTOR: José Francisco Cervantes-Alvarez
Departament of Electronics, Systems, and Informatics, ITESO
Bachelor's degree in Computer System Engineering (Instituto Tecnológico de Jiquilpan, Mexico)
Master's degree in Computer Science (CENIDET, Mexico)
PhD in Computer Science (University of Grenoble-Alpes, France)

NUMBER OF PAGES: XXI, 89

Resumen

Los grafos de conocimiento (*KGs* por sus siglas en inglés *Knowledge Graphs*) han surgido como una herramienta poderosa para representar información estructurada semánticamente y asistir el desarrollo de sistemas inteligentes. Este trabajo se centra en la generación de mapas semánticos como método de resumen para los KGs. En este trabajo, proponemos una estrategía que utiliza un algoritmo de agrupación basado en centroides para agrupar terminos en el grafo de conocimiento con cierta cercania semántica. Como forma de experimentación se emplean dos algoritmos de agrupación basados en centroides: Propagación de Afinidad y Particionamiento Alrededor de Medoides (*PAM* por sus siglas en inglés *Partitioning Around Medoids*), para capturar la distancia semántica entre los nodos en el KG y generar grupos significativos. Nuestros experimentos muestran resultados divergentes entre los dos algoritmos de agrupación, con la Propagación de Afinidad se observa cierta coherencia cualitativa y significatividad, mientras que PAM se desempeña bien en términos de métricas cuantitativas de validación interna. Aprovechamos los centroides calculados para inferir un término principal del mapa semántico, lo que contribuye a la representación visualmente informativa del grafo de conocimiento. La combinación del proceso de capturar la distancia semántica entre los términos del grafo de conocimiento, el uso de algoritmos de agrupación e la inferencia basada en centroides facilita una comprensión integral del grafo de conocimiento. Nuestros hallazgos destacan la importancia de considerar tanto medidas de evaluación cualitativas como cuantitativas al evaluar los resultados de la agrupación. La efectividad de los mapas semánticos se muestra en la visualización de los KGs y en el avance del campo de la visualización de grafos de conocimiento. La integración de algoritmos de agrupación basados en centroides, evaluación cualitativa y métodos de inferencia ofrece una mayor claridad e interpretabilidad para el análisis de un grafo de conocimiento a partir de un grupo de tareas de exploración claramente definidias en el capitulo 7.

Summary

Knowledge Graphs (KGs) have emerged as a powerfultool for representing semantic structured information and supporting the development of intelligent systems. This document focuses on the generation of semantic maps as summarization method for KGs. In this work, we propose a strategy that uses a centroid-based clustering algorithm to group terms in the knowledge graph with a certain degree of semantic similarity. For experimentation, two centroid-based clustering algorithms are employed: Affinity Propagation and Partitioning Around Medoids (PAM), to capture the semantic distance between nodes in the KG and generate meaningful clusters. Our experiments show divergent results between the two clustering algorithms. Affinity Propagation demonstrates a certain qualitative coherence and significance, while PAM performs well in terms of quantitative internal validation metrics. We leverage the computed centroids to infer a main term of the semantic map, which contributes to the visually informative representation of the KG. The combination of semantic distance capture, the use of 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 interpretability for the analysis of a KG based on a set of exploration tasks clearly defined in Chapter 7.

Acknowledgements

Firstly, the author wishes to convey his appreciation to the National Council of Science and Technology of Mexico for supporting this work under Grant 498322.
The author would like to express his gratitude to Dr. José Francisco Cervantes Álvarez and the co-director of this thesis, Dr. Luis Fernando Gutiérrez Preciado, who have provided guidance and assistance throughout the duration of this project. Specifically, the author would like to thank them for their trust throughout this period and the freedom they granted him to explore various areas and problems until he found the contribution described in this document.
Secondly, the author wishes to extend his thanks to the companies he worked with during this journey: Oracle and FICO. He is grateful for the trust and opportunities provided to him to complete this work. In particular, he would like to express his appreciation to his manager, Timothy Misner, who chose to hire him in the midst of his Ph.D. studies and supported him in obtaining tuition benefits from FICO.
The author would also like to express his gratitude to his family members. He appreciates the support of his parents, Pablo and Angeles, his dear brother Cesar, his beloved sister Lizeth, his incredible nephew Baastian, and his gorgeous niece Amy.
Furthermore, the author wants to thank his best friends Francisco Eduardo Balart and Luis Josue Calva, who were always available for feedback and advice during challenging times.
Finally, but no less important, he wants to express his heartfelt thanks to the *lohl*.

Contenido

Contents

List of Figures

List of Tables

List of Algorithms

Introduction

Knowledge Graphs (KGs) find extensive use across diverse Artificial Intelligence (AI) systems and application domains [72], including question answering [98] and machine translation [101]. The rise of large language models (LLMs) has further underscored the significance of KGs, as they can be leveraged for training and enhancing LLMs [70]. KGs serve as a method for structuring and presenting information, utilizing the concepts and tools of the Semantic Web. The Semantic Web envisions a version of the internet where data is both structured and interconnected, making it comprehensible to both humans and machines. KGs are seen as essential components of AI systems, offering the required basis for representation and reasoning abilities. This addresses the crucial design need of incorporating human involvement in the process [73]. The concept of a Knowledge Graph (KG) involves depicting real-world knowledge in a graph format. In this structure, nodes symbolize the entities of interest, while the edges illustrate the relationships between these entities [35]. Lately, both academic institutions and private companies have developed Knowledge Graphs, for instance, YAGO [91], DBPedia [3], Freebase [10], NELL[16], Google Knowledge Graph [90], Microsoft Satori [74], Facebook Entity Graph [23], and Wikidata [24], which contain millions of entities and billions of relationships. Primary uses of Knowledge Graphs involve improving search engines such as Google [90] or Bing [74], question answering [17], information retrieval, recommender systems [51, 55], domain specific KG building [11, 32, 100], and decision support in the life sciences [8, 64, 80, 106].

Certain KG applications, such as query answering or KG visualization, necessitate condensed versions of the original graphs [31, 40]. Existing strategies to visualize KGs are based on the strategies to visualize regular graphs and few approaches take advantage of the semantics encoded in the KGs [58, 66, 69]. However, These proposals do not fully utilize the knowledge represented in the edges for visualizing the graph and none of these visualization proposals describe a specific strategy to validate how useful the visualizations are for end users in fulfilling visual exploration tasks.

General Objective

Propose a new strategy for visualizing Knowledge Graphs by leveraging the semantic closeness contained in the edges of the graph.

Secondary Objectives

- Describe an evaluation methodology that can be used to validate the effectiveness of a visual representation in fulfilling visual data exploration tasks.
- Evaluate the effectiviness of semantic maps in fulfilling visual data exploration tasks.

Fulfilling these objectives allows users to explore and understand the inherent structure and relationships of the graph in a visually intuitive way.

This document is structured as follows. The first four chapters provide the required information to understand the contributions of this thesis. In Chapter 1, we delve into the features of the Semantic Web as well as its relationship with KGs. To achieve this, we provide a brief description of the most relevant knowledge representation strategies useful for KG representation. In Chapter 2, we explore the main works associated with visualization strategies for KGs. On the other hand, in Chapter 3, we describe the most recent works associated with KG summarization. Semantic maps are discussed in Chapter 4. Semantic Maps are commonly employed methods to comprehend intricate subjects [28]. They involve a categorical organization of information in visual form [42]. A semantic map usually features a central term that signifies the primary topic, linked to a collection of keywords that categorize the remaining vocabulary. These structures are formally described in Chapter 4. Next, the primary contribution of this thesis is fully detailed in Chapter 5. The creation of a semantic map necessitates the identification of groups of related words. Unsupervised learning offers clustering algorithms that categorize data into one or more classes based on measures of similarity or distance [82]. Theoretically, implementing a clustering approach to the vocabulary in a KG should yield groups that can be utilized to build the semantic map. A series of experiments validating this concept is presented in Chapter 6 where in addition we propose a method to evaluate the quality of these semantic maps. Finally, to illustrate the usefulness of these semantic maps in providing a high-level perspective of a KG, we carry out a survey involving a group of experts who compared

the use of semantic maps with the traditional visual representation of KGs. These experiments are described in Chapter 7 and highlight the effectiveness of semantic maps in providing a thorough understanding of the structure and relationships within a KG.

1. Semantic Web and Knowledge Graphs

Tim Berners-Lee introduced the idea of the Semantic Web, envisioning a web comprised of data accessible to machines for processing, rather than solely a web of documents intended for human consumption [9]. The Semantic Web works by enhancing the structure and meaning of information on the World Wide Web, making it easier for computers to understand and process data. The contributions outlined in this document depend on a thorough understanding of Web semantic and knowledge representation strategies, such as KGs. In this chapter, we delve into these topics to enable a comprehensive explanation of the contributions presented in this thesis document.

1.1. Semantic Web

The objective of the Semantic Web is to provide information with clear and precise meanings, paving the way for machine-to-machine interaction and automated services through semantic descriptions [50]. Understanding the Semantic Web requires an examination of the Semantic Web Technology Layer Cake (shown in Figure 1.1), which outlines the technical infrastructure needed to achieve these goals. In the lower layers, we have the XML (eXtended Markup Language), which is used to structure and format data on the web in a tree structure, while the URIs (Universal Resource Identifiers) provide unique identifiers for resources on the web. However, tree structures lack the capability to integrate data from multiple sources. To tackle this limitation, RDF [84] (Resource Description Framework) serves as a framework for providing a standardized way to describe resources on the web and the relationships between them. RDF Schema (RDF-S) builds on RDF by providing a vocabulary for defining ontologies and describing the semantics of RDF data. In computer science, an ontology serves as a descriptive representation of the world, encompassing a variety of types, properties, and relationship types [27].

The Web Ontology Language (OWL) [61] is a description logic-based formal language which is a more expressive language for defining ontologies and describing complex relationships between resources on the web. OWL allows for more precise and detailed modeling of domain knowledge. In addition to ontologies, Semantic Web standards also support languages for describing ontology-

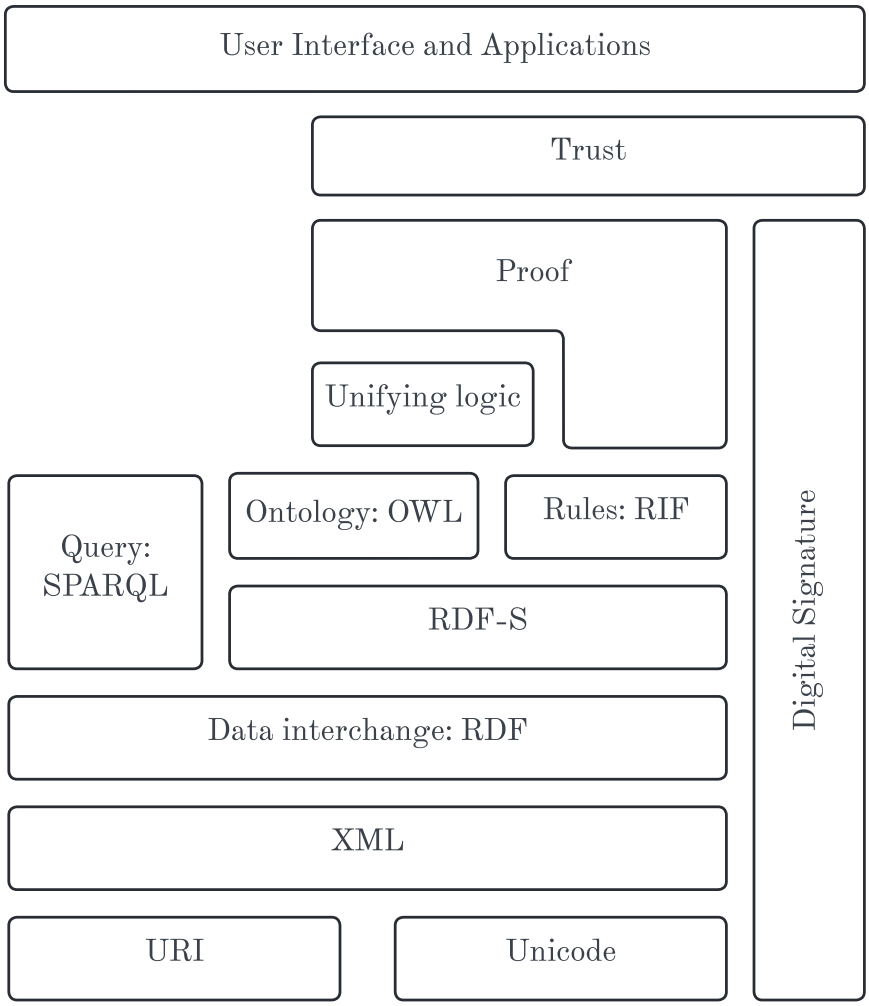


Figure 1.1: Semantic Web Technology Layer Cake

driven rules (i.e., mechanisms to infer data) and queries (i.e., mechanisms to access data). For example, SPARQL is a query language for ontologies, providing powerful capabilities for searching and retrieving information from RDF datasets.

On top of the Semantic Web Technology Layer Cake (also known as the Semantic Web Stack), we have Unifying Logic, which involves the use of formal logic to represent and reason about knowledge on the Semantic Web. The proof layer deals with mechanisms for generating and validating proofs in the Semantic Web. Trust is a critical aspect of the Semantic Web, as it involves assessing the reliability, credibility, and authenticity of information and sources. In the context of the Semantic Web, digital signatures can be applied to semantic data, ontologies, queries, and other artifacts to provide assurances regarding their origin and integrity. By digitally signing data and

documents, individuals and organizations can establish trust in the information exchanged over the web, thereby enabling secure communication, collaboration, and interaction within the Semantic Web environment.

1.2. Knowledge representation

We have discussed that the main goal of the Semantic Web is to provide a mechanism for capturing information in a machine-friendly way. We have also covered the technologies to achieve this. Now, the issue at hand is how to represent the information in a manner that facilitates interoperability among computers. Knowledge representation refers to the process of formalizing expert knowledge extracted from documents or interviews in a structured, computer-readable form. The goal is to store knowledge in a way that can be easily shared and integrated between enterprises. Knowledge representation models, such as ontology, provide a unique understanding of knowledge through formal and semantic definitions, making them suitable for mapping knowledge obtained from natural language documents [34]. This section describes the most relevant strategies to represent knowledge.

The research on Knowledge Representation (KR) can be traced back to the 1980s, when J. Mylopoulos described the terminology and issues associated with KR [65]. KR can be understood as a way to capture expert knowledge in a structured format that computers can process. In this context, it is natural to observe several knowledge fields proposing different methods to capture expert knowledge for developing intelligent systems. For instance, in the field of engineering design, the most relevant KR strategies include Gero's functional behavior structure (FBS) model, which combines the main components involved in engineering design activities [29], the Functional Representation (FR) model, and the Structural Behavior Function (SBF) model, which are similar to the FBS model and have attracted the attention of many scholars [7, 30]. Additionally, utilizing fuzzy comprehensive evaluation to judge the optimal representation method for knowledge, which includes KGs, IF-THEN production rules, and two-dimensional data linked lists [34].

Another field of study that has proposed different ways to represent knowledge is robotics. The survey presented in [71] introduces the concept of ontology as a method for organizing knowledge. It highlights that the primary purpose of knowledge representation (KR) in robotics is to provide

meaning to a robot's actions and its understanding of the environment. It is crucial for building an effective KR as it helps the robot understand the consequences of its actions, make decisions, and plan its tasks effectively. KR involves representing motions or skills for task planning, integrating perception systems to localize objects and communicate with others, grounding perception or control to logical statements for problem-solving, retaining experiences as beliefs to represent uncertainty, committing acquired knowledge for re-use, and defining the environment in which the robot operates. In [71], the concept of *ontology* is introduced as a means of organizing a defined set of terms or language. The objective of an ontology is to establish a framework of concepts and terms that characterize and describe the operational domain of the robot in a format that is comprehensible to humans as well.

1.2.1 Propositional Logic

On the other hand, a knowledge base is a repository or database that stores knowledge in a structured format, typically using a specific knowledge representation language or formalism. A knowledge base may contain various types of knowledge, such as factual information, domain-specific rules, ontologies, and logical relationships, organized in a way that facilitates efficient retrieval and manipulation by computational systems. The primary language of classical logic used in knowledge representation is first-order (predicate) logic, as advocated by John McCarthy [59] and Alan Robinson [79].

Example 1 *Knowledge Representation with propositional logic. Let's say we want to represent the statement:* ***If it is raining, then I will take my umbrella*** *In propositional logic, we can represent this statement using propositions:*

- *P: It is raining.*
- *Q: I will take my umbrella.*

The statement ***If it is raining, then I will take my umbrella*** *can be represented as the implication* $P \rightarrow Q$.

1.2.2 First-order logic

First-order formulas are essential for representing declarative knowledge and for automated proof using resolution. Propositional logic is a subset of first-order logic and has seen increased interest due to the development of fast satisfiability solvers [52].

Example 2 *Knowledge Representation with first-order logic. Now, let's consider a slightly more complex scenario where we want to represent the relationships between people and their ages:*

We have a set of predicates:

- *P(x): x is a person.*
- *A(x): x is an adult.*
- *C(x, y): x is the child of y.*

And constants:

- *Alice*
- *Bob*
- *Charlie*

We can represent the knowledge that Alice and Bob are adults as:
$A(Alice) \wedge A(Bob)$

And the knowledge that Charlie is the child of Alice and Bob as:
$C(Charlie, Alice) \wedge C(Charlie, Bob)$

1.2.3 Description logics

Description logics (DLs) [4, 6, 14] are a family of knowledge representation languages designed to structure and formalize knowledge in specific domains. They utilize concept descriptions, which are expressions constructed from atomic concepts and roles using constructors provided by the DL. DLs are distinguished from earlier approaches like semantic networks and frames by their formal,

logic-based semantics, offering a well-understood framework for representing domain knowledge [5].

In description logics, concept descriptions are used to construct statements within a knowledge base, which consists of two main parts: a terminological part (TBox) and an assertional part (ABox). The TBox allows for the description of relevant notions in the domain by specifying properties of concepts and roles, as well as relationships between them [5]. Essentially, the TBox functions similarly to a schema in a database setting, providing a structured framework for defining and organizing knowledge within the domain. On the other hand, the ABox is the *assertional* part of the knowledge base and it is used to describe a concrete situation by stating properties of individuals [5].

Before delve into an example of representing knowledge with DL, let's analyze briefly the fundamentals of the DL syntax. This description utilizes boolean constructors such as conjunction ($\sqcap$), disjunction ($\sqcup$), and negation ($\neg$), along with existential ($\exists$r.C) and universal ($\forall$r.C) restriction constructors. The DL syntax also includes relationship operators such as the operator $\sqsubseteq$, which represents a subsumption relationship. It denotes that one concept is a sub-concept of another concept, meaning that all instances of the first concept are also instances of the second concept. These constructors allow for the specification of complex relationships and constraints within a knowledge representation framework.

Example 3 *Knowledge Representation with Description logics. Let's say we want to represent knowledge about animals in a zoo. We'll define some basic concepts and relationships using DL:*

a) Concepts:

- *Animal: A general concept representing all animals.*
- *Mammal: A sub-concept of Animal representing mammals.*
- *Bird: A sub-concept of Animal representing birds.*

b) Roles:

- *hasColor: A binary relation representing the color of an animal.*
- *eats: A binary relation representing the diet of an animal.*

Now, let's represent some knowledge using these concepts and roles:

a) Terminological part (TBox):

- *Animal* $\sqsubseteq \exists$*hasColor.*
- *Mammal* $\sqsubseteq$ *Animal.*
- *Bird* $\sqsubseteq$ *Animal.*
- *Mammal* $\sqcap \exists$*eats.Plants* $\sqsubseteq$ *Herbivore: This states that mammals that eat plants are herbivores.*
- *Mammal* $\sqcap \exists$*eats.Insects* $\sqsubseteq$ *Carnivore: This states that mammals that eat insects are carnivores.*

b) Assertional part (ABox):

- *Tiger: A named individual representing a tiger.*
- *Tiger* $\sqsubseteq$ *Mammal: This asserts that a tiger is a mammal.*
- *Tiger* $\sqsubseteq \exists$*hasColor.Orange: This asserts that a tiger is orange in color.*
- *Tiger* $\sqsubseteq \exists$*eats.Meat: This asserts that a tiger eats meat.*

1.3. Knowledge Graphs: A Fundamental Expression of the Semantic Web

As we discussed previously, the Semantic Web is dependent on standards and technologies facilitating the representation, exchange, and inference of semantic data, including: RDF, OWL, SPARQL. Conversely, a KG serves as a method for encoding and storing knowledge in a format readable by machines, employing a graph structure comprising nodes and edges. Nodes denote entities or concepts, while edges signify relationships or properties among them [35].

It is natural to perceive KGs as part of the Semantic Web stack since they encode expert knowledge in a manner that machines can utilize for reasoning and inference purposes. Certain definitions characterize a KG as a knowledge base structured in a graph format [68, 87]. In this work, we consider a collection of sentences or facts articulated in a formal language such as description logic as the knowledge base for a KG. Essentially, KGs can be perceived as an assemblage of facts structured as `<subject,predicate,object>`. These collections are typically represented in languages such

as RDF, OWL, or N-Triples, the latter being a subset of the more intricate RDF/XML syntax. N-Triples is designed to be both human-readable and machine-readable, employing a plain text format that delineates RDF statements using subject-predicate-object triples, with each element separated by white space and concluded by a period.

The Semantic Web embodies a vision of a digitally interconnected environment where data is organized, linked, and comprehensible to both machines and humans. While KGs serve as a method for representing and structuring knowledge in alignment with its principles and technologies.

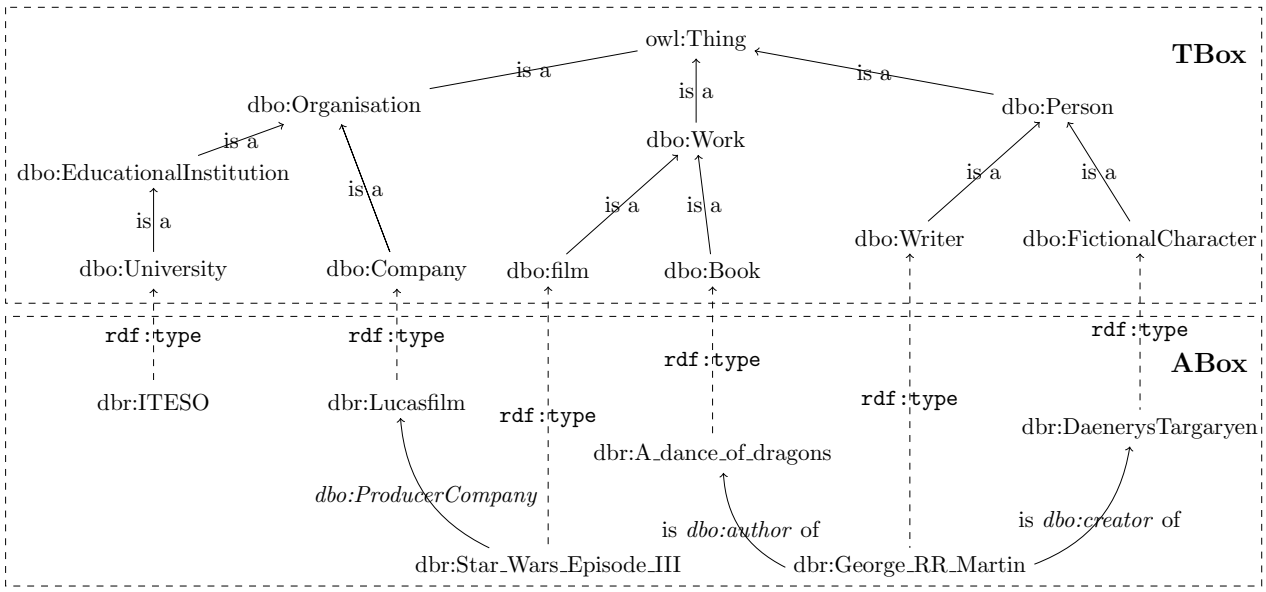


Figure 1.2: Small group of concepts and instances extracted from DBPedia.

Following description logic terminology, knowledge bases consist of two types of axioms: a terminology box (TBox) and an assertion box (ABox) [36]. Therefore, a KG should incorporate both sets of axioms to qualify as a knowledge base. Figure 1.2 illustrates the sets TBox and ABox comprising a group of entities and relationships sourced from DBPedia [3][1]. Within KGs, ontology classes (e.g., `dbo:Book` or `dbo:Movie`) correspond to the TBox, delineating concept hierarchies. On the other hand, ontology instances correspond to the ABox, describing entity instances (e.g., `dbr:Lucasfilm` or `dbr:George_RR_Martin`) and their relationships. Hierarchical relationships like "is a" establish connections between pairs of concepts in the TBox. For example, axioms (`dbo:Book`, *is a*, `dbo:Work`) and (`dbo:film`, *is a*, `dbo:Work`) signify that both `dbo:Book` and

[1]https://dbpedia.org (Last visited: 2024-03-13)

dbo:film concepts are descendants of the class dbo:Work. We can rewrite these assertions from TBox in DL formal language as follows:

- (dbo:Book, *is a*, dbo:Work): dbo:Book $\sqsubseteq$ dbo:Work
- (dbo:film, *is a*, dbo:film): dbo:Book $\sqsubseteq$ dbo:Work

In contrast, axioms in the ABox also indicate the types associated with individual entity instances. As an illustration, in Figure 1.2, the axiom (dbr:A_dance_of_dragons, *rdf:type*, dbo:Book) denotes that the resource dbr:A_dance_of_dragons is an instance of the class dbo:Book. Additional axioms in the ABox, such as (dbr:George_RR_Martin, *dbo:creator_of*, dbr:DaenerysTargaryen) and (dbr:George_RR_Martin, *dbo:author* of, dbr:A_dance_of_dragons), indicate that the instance dbr:George_RR_Martin has semantic connections with dbr:DaenerysTargaryen and dbr:A_dance_of_dragons entity instances. We can rewrite these assertions from ABox in DL formal language as follows:

- (dbr:A_dance_of_dragons, *rdf:type*, dbo:Book):
  dbr:A_dance_of_dragons $\sqsubseteq \exists$*rdf:type*.dbo:Work
- (dbr:George_RR_Martin, *dbo:creator_of*, dbr:DaenerysTargaryen):
  dbr:George_RR_Martin $\sqsubseteq \exists$*dbo:creator_of*. dbr:DaenerysTargaryen
- (dbr:George_RR_Martin, *dbo:author* of, dbr:A_dance_of_dragons):
  dbr:George_RR_Martin $\sqsubseteq \exists$*dbo:author*.dbr:A_dance_of_dragons

We will start by introducing a formal definition of a KG before delving into the other pertinent topics linked to the process to generate semantic maps delineated in this document.

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 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 $(subject - property - value)$.*

Based on Definition 1, it is evident that a KG can be depicted as a compilation of triples, which encapsulate axioms from either the ABox or TBox. Our attention in this study centers on the triples that portray axioms within the ABox set. More precisely, whenever an entity instance is referenced, it corresponds to the subjects of the ABox triples.

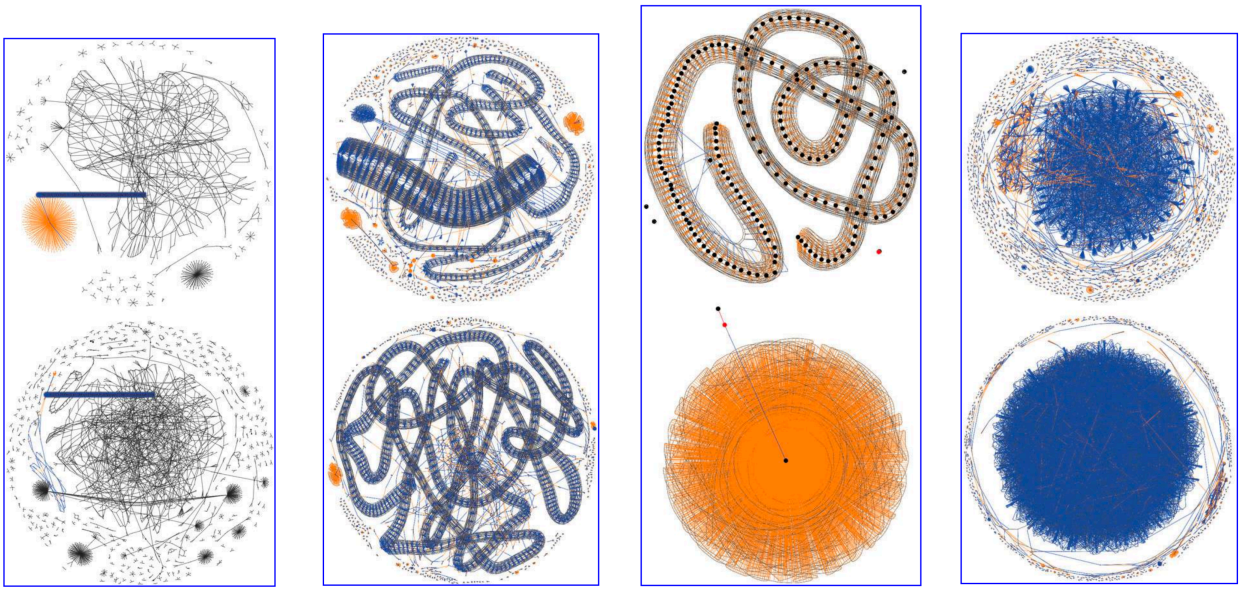
1.4. Conclusion

In conclusion, the Semantic Web and KGs represent pivotal advancements in the realm of data organization and interpretation. While the Semantic Web relies on standardized technologies like RDF, OWL, and SPARQL to facilitate the exchange and inference of semantic data, KGs provide a tangible method for encoding and structuring knowledge in a machine-readable format. Through their structured representation of facts and relationships, KGs enable both humans and machines to navigate and comprehend complex datasets, thereby advancing the vision of a seamlessly interconnected web of knowledge. We have also discussed some works associated with Knowledge Representation and highlighted the relevance of Knowledge Representation in understanding how current technologies support reasoning and inference tasks for KGs.

However, now the concern one may have is whether existing technologies could support the increasing amount of data that the modern web harbors. Are KGs truly useful in providing humans with a simple way to understand the resources allocated on the web?

2. Ongoing challenges in Visual Data Exploration of Knowledge Graphs

Knowledge graphs have become increasingly vital as a data and context source in Data Science. Initial data analysis typically involves exploration, with visualization playing a pivotal role. Visual exploration proves especially valuable when understanding of the source data and analysis objectives is limited. Nowadays, Semantic Web technologies dominate in modeling and querying knowledge graphs. In this chapter, we present some of the most relevant recent works addressing the issue of KG visualization.



(a) (b) (c) (d)

Figure 2.1: Example of visualizing Bitcoin Transactions presented in [60]: (a) Visualization of Bitcoin transactions reported as containing anomalous yet unidentified transactions at the apex of a money laundering operation. (b) Visualization of the initial *parasitic worm* transaction rate attack. (c) Visualization of the initial algorithmic responses to spam. (d) Visual representation of two distinct phases of the second data density-based *tumor* attack

2.1. Knowledge Graph Visualization

The concept underlying the visual data exploration process aims to represent data visually, enabling users to derive insights from the analyzed phenomena [46]. This process, often referred to as the *information seeking mantra*, comprises three stages: overview, zoom and filter, and details-on-demand [89].

Recent applications have demonstrated their utility in visualizing large graphs to comprehend various phenomena, such as Bitcoin transactions [60] and online discussions [63]. Some examples of these visualizations are shown in Figure 2.1. However, the performance evaluation outlined in [31] concludes that a distributed implementation of layout algorithms is essential to enhance the time required for generating visual representations of extensive knowledge graphs. Challenges associated with visualizing knowledge graphs include context adaptation, user interactions [81], data heterogeneity [88, 99], support for diverse analysis tasks (querying, combination, filtering, etc.), and performance concerns [31].

Some studies related to the visual representation of networks consider the objectives of data exploration [2, 102] as the goals for visualizing KGs. Conversely, alternative approaches concentrate on assessing the *usability* of software tools rather than evaluating the visual representation for fulfilling visual data exploration tasks [15, 39, 86].

Topics mas are a method for visualizing document collections [66]. Topic maps are created by applying topic modeling, specifically Latent Dirichlet Allocation (LDA), to a set of documents, which identifies themes or topics by grouping frequently co-occurring words. These topics are then projected into two-dimensional space to create a visual representation. Each document is associated with one or more topics, and the goal is to organize documents so that those sharing similar topics appear close together on the map.

On the other hand, CubeViz [58] is a platform for exploring and visualizing statistical data that adheres to the RDF Data Cube Vocabulary, a W3C standard for representing multi-dimensional statistical data in RDF. CubeViz simplifies the complex structure of RDF data cubes and offers a user-friendly interface for interactive filtering and visualization through faceted browsing and chart generation.

The Sextant [69] is a web-based tool designed to visualize and explore linked spatio-temporal

data. It addresses the challenge of representing and visualizing the temporal evolution of linked data, which is crucial for geospatial datasets frequently updated over time, such as those used in Earth Observation (EO) and environmental monitoring. The paper showcases several use cases, including environmental monitoring, fire detection, and the evolution of land cover, demonstrating the practical applications of Sextant in handling and visualizing large-scale geospatial data.

In [12], authors introduce a workflow for visualizing linked data that can be applied to visualize KGs. These phases are outlined in a generic manner and can be customized to suit the specific data domain intended for visualization. In Figure 2.2, we present two KGs visually represented by applying this workflow.

a) **Data retrieval**: This initial step involves acquiring or generating linked data. For instance, in the context of the Semantic Web or Knowledge graphs, this step may entail executing a SELECT SPARQL query to produce an n-triple file.

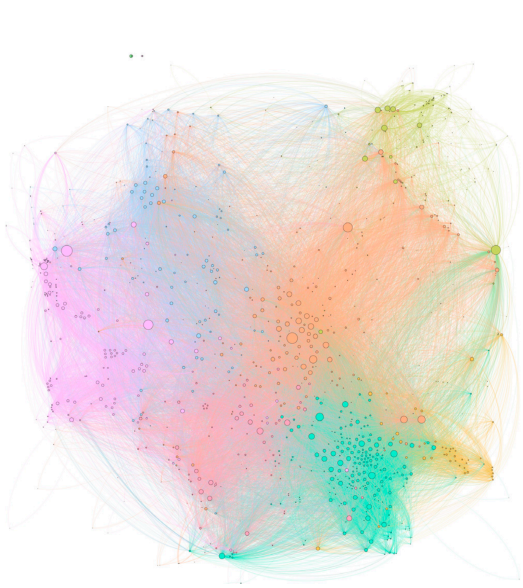
b) **Graph building**: After obtaining the linked data, the next step is to establish connections between entities based on the retrieved information. The objective of this phase is to create a machine-processable representation of the linked data. For example, this could involve constructing an iGraph instance in R.

c) **Graph calculations**: Subsequently, the graph can be manipulated to enhance visualization. For instance, the size and color of vertices and edges can be determined based on metrics such as degree or betweenness centrality.

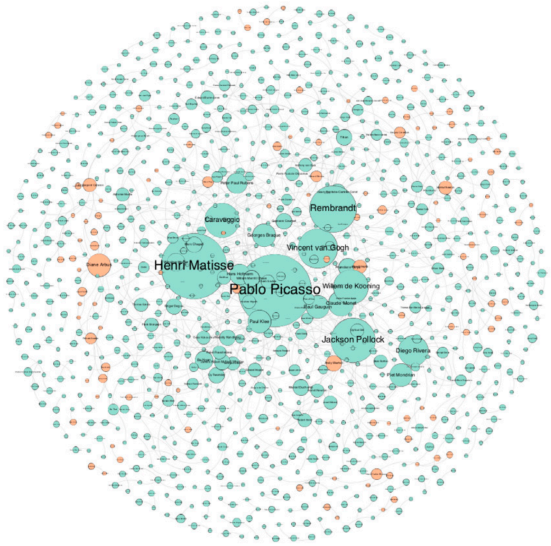
d) **Graph layout**: In this stage, the spatial positions of the vertices are determined to provide an aesthetically pleasing representation to end-users. One commonly used graph layout algorithm is force-based vertex placement, which calculates attractive and repulsive forces between every pair of vertices to determine their optimal proximity.

e) **Rendering**: The final step involves presenting the graph on the desired screen. For example, this could be achieved using libraries such as D3 [1] to create visually appealing graphs within a browser window.

[1] `https://d3js.org` (Last visited: 2024-03-30)



(a) DrugBank drug to drug interactions where colors are assigned to communities.



(b) Artist influences network, filtered by most relevant painters and photographers where vertex size is proportional to degree

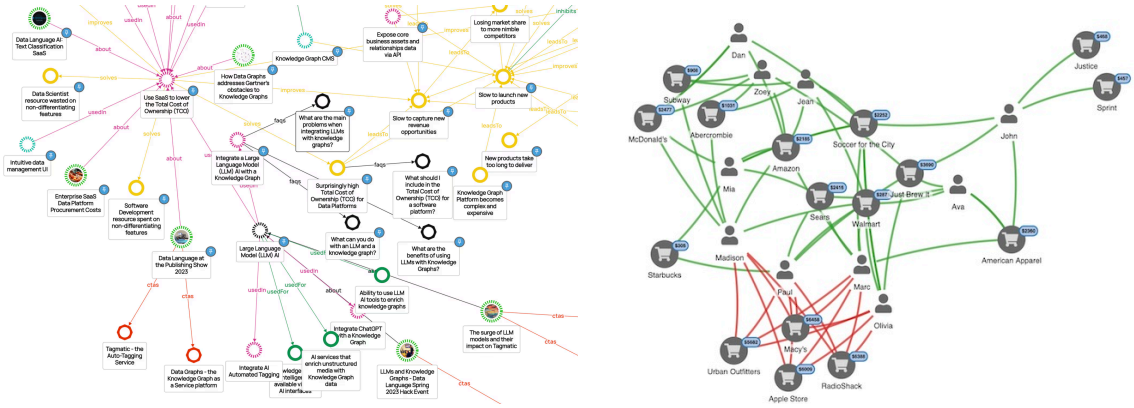
Figure 2.2: Visual representations presented in [31] exemplify the application of the graph visualization pipeline described in [12] to render large KGs.

2.2. Existing tools to visualize Knowledge Graphs

2.2.1 Commercial tools

In addition to recent advancements in KG visualization, several commercial products enable analysts to visualize RDF graphs, including Data Graphs[2] and tools developed by Cambridge Intelligence company: Keylines, ReGraph, and KronoGraph[3]. These tools facilitate rendering KGs to support tasks in domains such as pharmacy and bio-science research or financial analysis. In Figure 2.3 are shown two examples of visual representations of KGs produced by DataGraph and ReGraph.

[2] `https://datagraphs.com` (Last visited: 2024-05-01)
[3] `https://cambridge-intelligence.com/` (Last visited: 2024-05-01)



(a) (b)

Figure 2.3: KGs visualizations produced by some existing tools: (a) Example of visualization produced by DataGraph tool. (b) Practical use case of ReGraph tool to credit card fraud detection [45].

2.2.2 Free tools

In the realm of free tools, two online platforms consume RDF data and produce visual representations: isSemantic visualizer[4] and RDF grapher[5]. However, these tools have limitations in the amount of data they can effectively process. In Figure 2.4 it is shown an example of the visual representation of a KG produced by isSemantic tool.

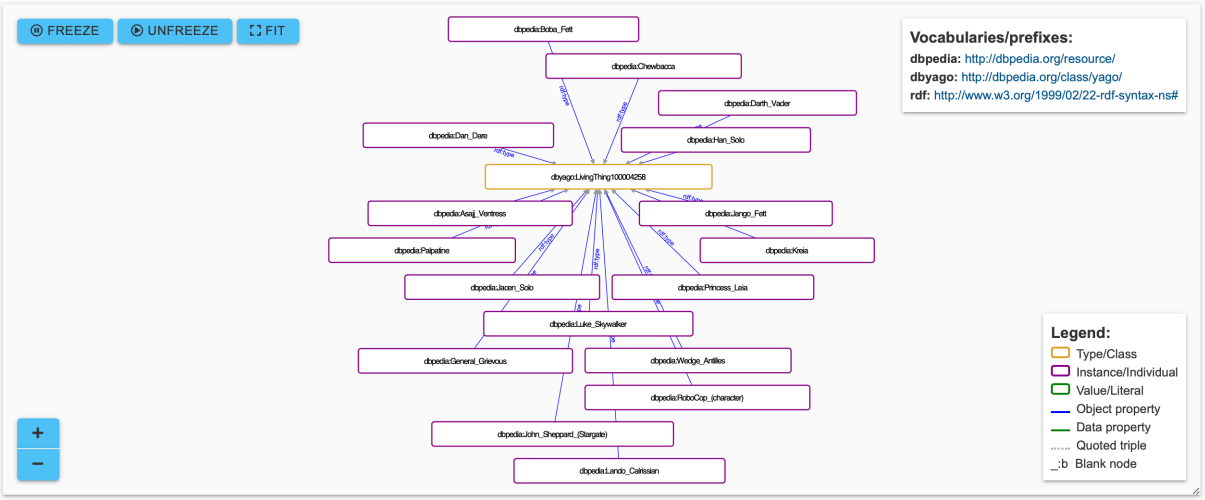


Figure 2.4: KGs visualizations produced by RDF visualizer isSemantic from a KG extracted from DBPedia containing some characters of Sci-Fi movies *Star Wars*

[4] `https://issemantic.net/rdf-visualizer` (Last visited: 2024-05-01)
[5] `https://www.ldf.fi/service/rdf-grapher` (Last visited: 2024-05-01)

2.3. Challenges on Knowledge Graph Visualization

With the increasing adoption of KGs in decision-making applications, there is a pressing need to condense and summarize KGs for effective data representation [33]. Recent research [88, 99, 102] has indicated that visualizing a simplified version of a large graph can serve as a viable alternative. For example, visualizing a summarized graph can prove beneficial when implementing a distributed solution to render a large graph is not feasible. In this context, summarizing graphs is generally beneficial for reducing data volume and storage, thereby facilitating the graph visualization process [56]. Visual data exploration is considered a process that generates hypotheses by enabling users to gain a deep understanding of the data [46]. Hence, summarizing a KG is crucial for creating an efficient visual representation that aids in comprehending the relationships between entities and concepts within a domain. By presenting information visually, users can promptly identify patterns, trends, clusters, or related information that may be challenging to discern in a text-based representation. Current methods for visualizing KGs primarily focus on depicting the entire structure [31], thereby impeding data analysts from exploring the KG beyond its structural information.

2.4. Conclusion

Recent advancements in data visualization techniques have proven instrumental in understanding complex phenomena, as evidenced by their application in visualizing large graphs. However, challenges persist in optimizing the performance of visual representations, as highlighted in [31], where authors advocate for distributed implementations of layout algorithms to improve efficiency, particularly for extensive knowledge graphs. The issue arises when classical visualization strategy to render the whole structure is unavailable; in such cases, summarizing these structures appears to be more feasible. Summarizing graphs not only reduces data volume and storage requirements but also facilitates the visualization process [56]. Thus, addressing the KG summarization challenge and refining visualization techniques are imperative for leveraging the full potential of knowledge graphs in data analysis and decision-making processes.

3. Review of current advances in summarizing Knowledge Graphs

Graph summarization is the field that aims to distill complex networked data into more manageable and interpretable forms. Researchers have explored various techniques to summarize large graphs, addressing challenges related to visualization, storage reduction, and efficient analysis. According to [56], the majority of graph summarization techniques can be categorized into four main groups: grouping or aggregation-based methods, bit compression-based methods, simplification or sparsification-based methods, and influence-based methods. Typically, in knowledge graph summarization, the simplification or sparsification-based approach is favored. KG summarization usually adopts simplification or sparsification because the main goal of KG summarization is to generate a subgraph that accentuates the significant entities and relationships present in the original graph. In this chapter, we explore the most recent strategies proposed for summarizing large graphs and KGs.

3.1. Approaches to Summarize Graphs

Functional summaries, initially introduced by the bioinformatics community to derive a concise and interpretable representation of Protein-Protein Interaction (PPI) networks [85], offer a high-level abstraction of maps derived from knowledge bases. In this chapter, we focus on exploring graph summarization techniques for networked data, also known as networks or graphs. The applications of graph summarization encompass various benefits such as data volume and storage reduction, acceleration of graph algorithms and queries, support for interactive analysis, and noise elimination [56].

Recent research has introduced various approaches to summarize large graphs, aiming to facilitate efficient visualization of their content. Shen et al. propose OntoVis in their work [99], a visual analytics tool that employs both structural and semantic abstractions to provide a condensed version of large graphs for simplified visualization. Another relevant contribution is presented in [47], where authors introduce the VoG (Vocabulary-based summarization of Graphs) algorithm. This

algorithm constructs and visualizes subgraph-types such as stars, cliques, and chains to summarize and comprehend large graphs effectively. Additionally, a visual abstraction method presented in [102] transforms geo-tagged social media data into high-dimensional vectors using a doc2vec model. Furthermore, Koutra et al. [81] focus on summarizing Knowledge Graphs (KGs) by leveraging individual interests to generate personalized knowledge graph summaries.

Koutra et al. [56] present a taxonomy of graph summarization algorithms, categorizing them into two families based on the type of network: static and dynamic. In both static and dynamic networks, summarization techniques encompass grouping-based, bit compression-based, and influence-based methods. In the context of dynamic networks, grouping-based summarization approaches involve recursively aggregating nodes and timesteps to reduce the scale of large-scale dynamic networks. Bit compression-based methods utilize compression to extract significant patterns from temporal data.

3.2. Approaches to Summarize Semantic Graphs

When discussing KG summarization, we are essentially elaborating on the concept of semantic graph summarization, which encompasses knowledge graph summarization as a subset [97]. Regardless of the graph type (semantic or non-semantic), summarization techniques share objectives and methods. Irrespective of the application or network type, one of the primary challenges in graph summarization is determining the data of interest. Each summarization strategy relies on selecting criteria of interest to extract meaningful information [56]. However, defining what qualifies as *interesting* is not straightforward. For instance, the FUSE algorithm [85] introduces a profit maximization model aiming to identify a summary by maximizing information profit within a budget constraint. Conversely, VoG [47] utilizes the Minimum Description Length (MDL) principle to select the best subgraphs, opting for those that conserve the most bits. In the case of semantic abstraction proposed in [102], a dual-objective blue noise sampling model is employed to choose a subset of social media data items, emphasizing spatial distribution and semantic correlation for the resulting simplified geographical visualization. Regarding personalized summaries from KGs [81], the criteria for determining what information is *interesting* for each user are based on their query history. For geographical KGs, [97] proposed to adopt the concept of distributional semantics within

geographical contexts and investigate the similarity and interconnectedness of various types of locations by employing diverse latent representations enhanced with spatial contexts. Lastly, in [83], Scherp et al. delves into the notion of summarization of semantic graphs using quotients, with a focus on creating concise representations of input graphs while maintaining particular structural attributes.

Another approach to address the graph summarization challenge is through *Entity Matching*. This method finds application in fields like data science or data management where matching entities is essential. In the context of knowledge graph summarization, entity matching helps identify entities that should belong to the same summary unit. For example, in [92], the authors outline a summarization framework called COMET (Context-Aware Matching Technique), which relies on semantic data integration and entity matching principles to merge RDF molecules using semantic similarity metrics.

3.3. Knowledge Graph Summarization

Several authors have proposed various definitions for the KG Summarization process. Informally, in [56], it is suggested that the KG Summarization problem can be described as the process of finding a labeled summary graph or a set of labeled structures to concisely describe the given graph. Scherp et al. [83], on the other hand, provide a more formal definition where they define the KG Summarization problem in terms of summary models to preserve selected features and payload functions to associate this graph with a given task. Based on the analyzed works, we define KG summarization as the process of creating a KG that represents the original KG with a reduced number of edges and/or vertices.

3.4. Conclusion

Analyzing KGs poses a significant challenge in the field of data science due to the sheer size and complexity of the encoded information. In recent years, considerable attention has been directed towards two research domains: graph visualization and graph summarization, particularly concerning KGs. In this work, we focused on the visualization challenge for KGs. Many visualization techniques for knowledge graphs primarily focus on presenting the entire structure. However, the

continuous expansion of KGs has begun to surpass the effectiveness of this approach. Conversely, graph summarization techniques for KGs are emerging as a promising avenue for generating informative summaries tailored for analysis purposes. One of the key challenges is to devise a novel approach that integrates cutting-edge strategies for summarizing large graphs with traditional visualization techniques.

4. Semantic Maps as a strategy to Visualize KGs

Semantic maps are graphical representations that show the relationships between different concepts or words within a particular domain or field of study. In this chapter, we consider semantics maps to be useful for visualizing the high level of abstraction of a KG based on the semantic closeness of entities within the KG. We describe some related work associated with KGs and visualization strategies similar to semantic maps. We propose finding a way to extract a semantic map of a KG to obtain a reduced version of the KG. To elaborate on this idea, in this chapter, we describe what a semantic map is and how we can compute the similarity among all entity instances in a KG.

4.1. Semantic Maps

A semantic map serves as a graphical representation that reveals the relationships between different concepts or words within a specific domain or field of study [42]. Its purpose is to visually organize and display the meaning and connections among various terms or concepts, emphasizing their semantic similarities and differences. In essence, a semantic map provides a visual depiction of how distinct ideas or concepts relate to one another and how they group together based on shared meanings or semantic properties. Notably, there exist alternative mathematical representations for semantic maps, including graphs and Euclidean spaces [20]. For an illustrative example, refer to Figure 4.1, which showcases a semantic map centered around the topic of "Water" This map comprises three node categories: (1) the central word (root), (2) a set of keywords (such as Usages and Living things), and (3) the vocabulary associated with each keyword (e.g., words like Cooking and Bathing linked to the keyword Usages).

Let us propose a formal definition of semantic maps as follows:

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*

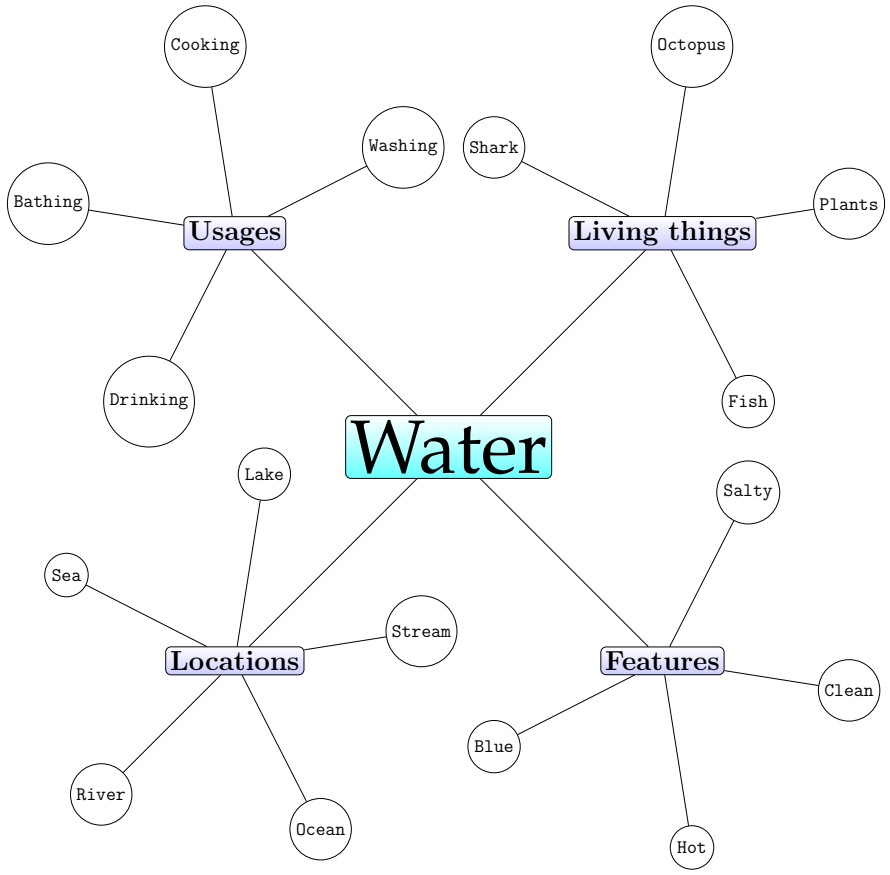


Figure 4.1: Example of a semantic map of concepts and vocabulary associated with topics *Water*.

the semantic map. E_T *is defined as follows:* $E_T = (E_{KW} \cup E_w)$*, where* $E_{KW} = (\alpha \times KW)$*, and* $E_w = (KW_i \times w_i) \forall_{KW_i \in KW}, \forall_{w_i \in C_i}$.

4.2. Semantic Similarity in Knowledge Graphs

Before describing the process to compute the semantic map associated with a KG, we first need to explore the existing approaches to compute the similarity between two entities in a KG. Then, we can use this approach to outline a generalized method to find groups of related entity instances in a KG.

Semantic similarity is a metric widely employed in Natural Language Processing (NLP) and Information Retrieval (IR) domains [37], representing the degree of relatedness between two concepts based on their hierarchical relations [77, 94]. In a KG, the semantic similarity between two entities e_1 and $e_2 \in V$ is denoted as $sim(e_1, e_2)$. Intuitively, the semantic distance between two

words is commonly determined by the path connecting them in a KG. Semantic similarity metrics can be broadly classified into two main groups: corpus-based and knowledge-based approaches [62]. Corpus-based similarity metrics focus on assessing the similarity between two concepts using information extracted from large corpora. Examples of corpus-based similarity metrics include pointwise mutual information [18] and latent semantic analysis [48]. In contrast, knowledge-based similarity metrics quantify the semantic relatedness between two words using information provided by the TBox in a KG. Knowledge-based approaches encloses path-based metrics proposed by Hulpus et al. [38], Wu and Palmer [95], and Leacock and Chodorow [49]. Other knowledge-based measures utilize the Information Content (IC) metric, such as Lin [54], Jiang and Conrath [41], and Resnik [77]. The IC of concepts is a statistical measure that quantifies the specificity of a concept over a corpus. Higher IC values indicate more specific concepts (e.g., `dbo:Book`), while lower values are associated with more general concepts (e.g., `owl:Thing`). Hybrid knowledge-based approaches, such as IC-graph [104] or Zhou [103], combine IC with other metrics to assess the relatedness of two words. For instance, graph-based IC [104] employs a SPARQL query on DBPedia to compute $freq_{graph}(c_i)$ and N values in the following expression:

$$IC_{graph}(c_i) = -logProb(c_i) \tag{4.1}$$

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 $freq_{graph}(c_i) = |\mathcal{E}(c_i)|$.

4.3. Conclusion

In conclusion, semantic maps serve as tools for illustrating the intricate relationships between concepts or terms within a specific domain or field of study. By visually organizing and displaying the semantic connections among various entities, these maps highlight both similarities and distinctions, offering insights into how ideas are interconnected. While there exist diverse mathematical representations for semantic maps, such as graphs and Euclidean spaces, their fundamental purpose remains consistent: to provide a tangible depiction of semantic relationships. Furthermore, in the context of KGs, semantic similarity emerges as a crucial metric, facilitating the assessment

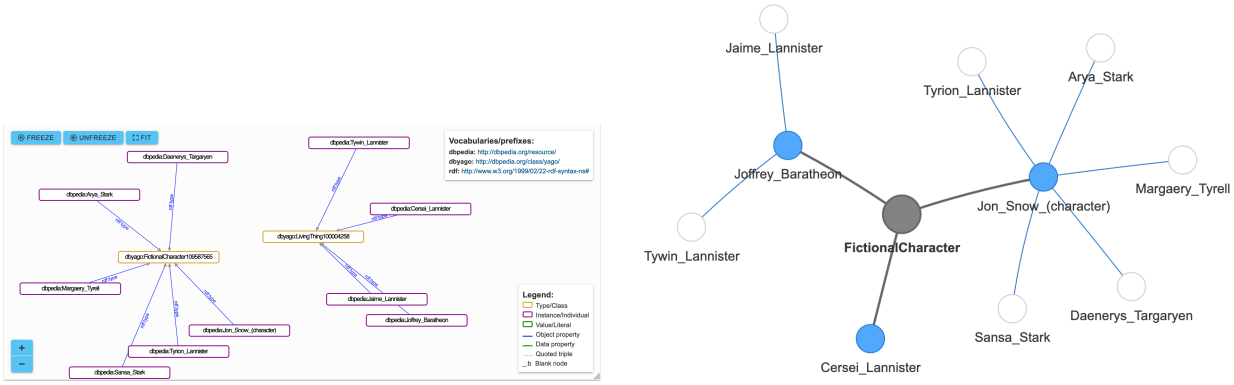
of relatedness between entities. Leveraging semantic maps to visualize the high level of abstraction inherent in KGs, we propose an approach aimed at extracting concise representations of KGs based on semantic proximity. This endeavor implies computing the similarity among all entity instances in a KG, ultimately culminating in the generation of a reduced version of the KG. Through this combination of the process to build semantic maps and similarity computation, we aim to enhance our understanding of KGs while streamlining their representation for practical applications and analysis.

To the best of the author's knowledge, no prior research has directly addressed KG visualization using semantic maps. While several studies have explored related aspects, such as Topic Maps [66], CubeViz [58], Sextant [69], and [20] (explored in Chapter 2), these approaches do not explicitly use semantic maps to visualize KGs. This thesis introduces a methodology for extracting a semantic map from a given KG, offering a unique perspective on the use of semantic maps for performing Visual Data Exploration tasks.

5. Summarize Knowledge Graphs through Semantic Maps

5.1. Semantic Maps as a strategy to reduce the size of a Knowledge Graphs

The idea behind the semantic map of a KG is to condense the content of the KG by exploiting semantic similarities among pairs of entities. To illustrate this concept, let us create a concise KG derived from DBPedia, focusing on fictional characters from fantasy novel series authored by George R. R. Martin. Figure 5.1a) depicts a visual representation generated by an online RDF graph visualizer [1], which facilitates easy visualization and analysis of RDF datasets. However, this visualization may lack clarity and appeal, potentially impeding effective exploratory analysis. In contrast, Figure 5.1b) presents a semantic map of the KG, employing a node-link format to depict entity instances linked to their respective centroids, with all centroids interconnected with the central concept of the semantic map. In this study, we propose to take advantage of the benefits of semantic maps to illustrate the intricate relationships between concepts in a KG by organizing them based on semantic closeness.



(a) (b)

Figure 5.1: 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.

[1] https://issemantic.net/rdf-visualizer (Last visited: 2024-03-29)

Contribution 1 (Semantics maps to reduce size of KGs) *Given a KG, we propose extracting its semantic map to reduce the number of edges and group related entities of the KG by using the semantic similarity among these entities. Formally, we propose a function* $\psi(\lambda) : KG \rightarrow \mathcal{SM}_{\lambda}$*, where* ψ *represents the process of extracting a semantic map, and* λ *signifies the centroid-based clustering algorithm used to group the entities in the KG.*

5.2. Extracting semantic distance of entities in a Knowledge Graph

The initial step in the process to generate semantic maps entails clustering the entities within the KG based on their semantic proximity. The main challenge in this phase is to extract numerical data from the KG and generate a set of entity clusters. Our approach involves computing the semantic distance for each pair of entities in the KG by constructing 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 presents the process for computing the semantic distance matrix D. It begins by obtaining the set of triples that define the KG from the edge set E. For each edge $e \in E$, the functions $subject()$, $property()$, and $value()$ are used to extract the subject, property label, and target entity instance, respectively. Subsequently, for every pair of triples $t_i, t_j \in T$, the algorithm calculates the semantic similarity using the function $sim()$.

The semantic similarity function $sim()$ is defined using a list of state-of-the-art semantic similarity measures such as path [75], Wu and Palmer (wup) [95], Leacock and Chodorow (lch) [49], and information content-based measures: Resnik (res) [76], Jiang and Conrath (jnc) [41], and Lin (lin) [54]. According to the original authors of the function $sim()$ [105], a threshold $\eta \in [0, 1]$ is used to establish the semantic similarity between two synsets [2].

[2] A set of words that share one common sense is called a synset [105]

Algorithm 1: Algorithm to build the semantic distance matrix

Input: Set of edges E of K
Result: Semantic distance matrix D
// Let T be the set of triples associated with the K
$T \leftarrow \emptyset$;
foreach $e \in E$ **do**
  $T \leftarrow T \cup \{(subject(e), property(e), value(e))\}$;
end
foreach $(t_i, t_j) \in T \times T$ **do**
  if $t_i = t_j$ **then**
    $D(i,j) \leftarrow 0$;
  end
  else
    $e_a \leftarrow subject(t_i)$;
    $e_b \leftarrow subject(t_j)$;
    $D(i,j) \leftarrow D(j,i) \leftarrow 1 - sim(e_a, e_b)$;
  end
end
return D;

5.2.1 Complexity analysis of semantic distance extraction

The original authors of [105] do not provide a detailed explanation of the computational complexity of the function $sim()$. However, we can offer a brief analysis of its computational complexity based on the algorithm description provided and the available code of $sim()$ [3]. The function $sim()$ takes as input parameters the references of two entities within the YAGO KG, represented as e_1 and e_2. It involves five phases:

a) Extracting concepts from the YAGO KG,

b) Mapping concepts to synsets,

c) Calculating the IC metric,

d) Calculating scores for synsets, and

e) Obtaining the final score for the given entities.

[3] `https://github.com/gsi-upm/sematch` (Last visited: 2024-03-29)

The filtering stage operates in linear time, precisely $O(N_1 + N_2)$, where N_1 and N_2 denote the number of concepts linked with entities e_1 and e_2, respectively. Similarly, associating concepts with synsets for each entity also requires linear time, specifically $O(N_1 + N_2)$, for each entity. Computing the IC metric for each synset and identifying the most common synsets also consumes linear time, $O(N_1 + N_2)$. The comparison and score computation involve nested loops. In the worst-case scenario, there are N_1 iterations for the score of e_1 and N_2 iterations for the score of entity e_2, resulting in a time complexity of $O(N_1 \cdot N_2)$. The final score computation is a constant-time operation, $O(1)$. Overall, the most time-intensive segment of the code involves the nested loop for comparing synsets and calculating scores, resulting in a time complexity of $O(N_1 \cdot N_2)$.

The relationship between similarity and distance follows the principle that greater similarity between two entities corresponds to a shorter distance between them. Each row in $D(K)$, denoted as the i-th row, comprises a vector containing semantic distance values between the i-th entity and all other entities in the KG. The semantic distance between each entity and itself is defined as 0. Our approach entails employing a centroid-based clustering algorithm to produce a non-overlapping set of clusters, using the semantic distance matrix D as input for the selected clustering algorithm.

5.3. Clustering entities of Knowledge Graphs

An essential aspect of semantic maps involves identifying specific nodes in the KG that represent each group of entity instances. In the context of grouping data points in graphs, there are two main approaches: clustering algorithms and graph community detection algorithms. Since we need to identify a representative item for each group, we will focus on centroid-based clustering algorithms. Centroid-based focuses on partitioning point data in a vector space, while graph community detection focuses on identifying dense subgroups within a network structure.

We propose utilizing PAM and Affinity Propagation center-based clustering algorithms, primarily because they can handle distance matrices as input rather than feature vectors [26, 43]. PAM is an iterative algorithm that selects a set of k medoids from data points and assigns each non-medoid point to its closest medoid, aiming to minimize the sum of distances between each data point and its assigned medoid. Conversely, Affinity Propagation propagates messages between data points to determine which points should serve as exemplars representing their respective clusters.

5.3.1 Clustering-based algorithms

Clustering algorithms based on centroids utilize the notion of centroids to cluster similar entities within KGs. Ultimately, we underscore the significance of visual data exploration techniques in comprehending knowledge graphs and showcase how semantic maps derived from knowledge graphs can aptly visualize their content.

Various methodologies exist for clustering data, with recent surveys categorizing these approaches based on application or data type for grouping [1, 25, 96]. Clustering types encompass Centroid-centric, Density-based, Distribution-based, and Hierarchical clustering [19].

One phase of the process to build semantic maps involves assigning each entity to the most suitable cluster based on semantic similarity. Each resulting cluster requires a node representing all contained entities. The set of these representative nodes is termed the semantic map's keywords. Viewing these keywords as centroids of clusters underscores the importance of employing centroid-based clustering.

The fundamental concept of centroid-centric clustering lies identifying *k* centroids (or centers), followed by computing *k* sets of data points that minimize proximity to each center. For example, the K-means algorithm aims to minimize the sum of squared distances between data points and the cluster centroids [57]. A variant of K-means, the PAM (Partitioning Around Medoids) algorithm, minimizes dissimilarities between points in a cluster and the centroids [43]. CLARA (Clustering Large Applications) extends PAM for large datasets [44]. Conversely, CLARANS (Clustering Large Applications based on RANdomized Search) is a partitioning algorithm focused on spatial data mining, recognizing patterns and relationships in spatial data such as topological data [67]. Another centroid-centric clustering algorithm is the Affinity Propagation (AP) algorithm, employing a message-passing procedure to broadcast messages of attractiveness and availability among data points [26].

Let $C = \bigcup C_i$ denote the set of clusters obtained after applying a centroid-based clustering algorithm. Each cluster C_i contains a centroid element represented by $centroid(C_i)$, and the set of centroid elements is defined by Equation 5.1.

$$\mu = \bigcup_{C_i \in C} centroid(C_i). \tag{5.1}$$

5.4. Central concept of the semantic map

A fundamental aspect of a semantic map is its central concept, which serves as the focal point of the graphical representation. In this investigation, we denote this central concept as α. Within a typical semantic map, α is connected with a selected set of keywords (for instance, Usages, Living things, Locations, and Features of *Water* as illustrated in Figure 4.1). These keywords are employed to represent each group of terms within the semantic map. This study suggests utilizing centroids obtained from centroid-based clustering algorithms [96] as the keywords of a KG. Hence, we denote these keywords as the set of centroids μ of the entities in a KG, where each μ_i represents the centroid of the i-th cluster C_i, i.e., $\mu_i = centroid(C_i)$.

To identify the central term α, we propose computing the IC_{graph} measure for all types associated with each centroid in μ. Information Content (IC) [54] functions as a statistical metric to ascertain the specificity of a concept across a corpus. Higher IC values indicate more specific concepts (e.g., `dbo:Book`), whereas lower IC values are linked to more general concepts (e.g., `owl:Thing`). Hybrid knowledge-based approaches such as IC-graph [104] or Zhou [103] blend IC with other metrics to evaluate the relational closeness between two words. For instance, the graph-based IC [104] utilizes a SPARQL query on DBPedia to compute $freq_{graph}(c_i)$ and N values in the following expression:

$$IC_{graph}(c_i) = -logProb(c_i) \tag{5.2}$$

Defining $types(e_i)$ as the function that retrieves the set of types associated with the entity e_i, we designate $\mathcal{T}$ as the set of shared types among all centroids in μ. This formulation is formally outlined in Equation 5.3.

$$\mathcal{T} = \bigcap_{\mu_i \in \mu} types(\mu_i) \tag{5.3}$$

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: Infer main term α

Input: μ: Set of centroids of C
Result: α: Main term of K
$\mathcal{T} \leftarrow types(\mu_0)$;
foreach $\mu_i \in \mu - \mu_0$ **do**
  $\mathcal{T} \leftarrow \mathcal{T} \cap types(\mu_i)$;
end
$\alpha \leftarrow \max_{t \in \mathcal{T}} IC_{graph}(t)$;
return α;

5.4.1 Complexity analysis of the process to infer the term α

Algorithm 2 formalizes the procedure for deducing the primary term of K by initializing the set of shared types, denoted as $\mathcal{T}$, with the types linked to the centroid of cluster C_0. The computational complexity of $IC_{graph}()$ (refer to Equation 5.2) depends on the complexity of constructing the set of entities categorized under the concept c_i, denoted as $\mathcal{E}(c_i)$. To achieve this, we traverse the entire KG, incurring a time complexity of $O(N)$, where N represents the number of nodes in the KG. For each node, we retrieve the list of types and search for the concept c_i. Assuming that obtaining the list of types involves a complexity of $O(t)$, where t is the number of types associated with concept c_i, the overall complexity of $IC_{graph}()$ is $O(N \cdot t)$.

5.5. Semantic map of a Knowledge Graph

The process to build semantic maps involves clustering KG entities and determining the central term α. Algorithm 3 outlines the steps for constructing the semantic map of a KG, visually depicted in Figure 5.2. The algorithm consists of four main steps: constructing the semantic distance matrix, computing the clusters, inferring the central term, and generating the semantic map. The complexity of each step depends on the number of triples in the input KG and the number of clusters.

Definition 5 formalizes the proposed concept of the semantic map associated with a KG. Symbols mentioned in Definition 5 are described in the Table 5.1.

Definition 5 (Semantic map of a Knowledge Graph) *Given a Knowledge Graph $K = (V, L, E)$, a semantic distance matrix $D(K)$, the centroid-based clustering algorithm λ, and the main term of K resulting of applying the clustering defined by λ: α_λ, the semantic map of K is defined as*

$\mathcal{SM}_\lambda(K) = (\alpha_\lambda, E_K, \mu_\lambda, \mathcal{NC})$.

Algorithm 3: ψ: Process to produce the semantic map of a KG

Input: E: Edges associated with the KG to reduce
Input: λ: Centroid-based Clustering algorithm
Result: $\mathcal{SM}_\lambda(K)$
$D \leftarrow$ buildSemanticDistanceMatrix(E);
$C_\lambda, \mu_\lambda \leftarrow$ computeClusters(D, λ);
$\alpha_\lambda \leftarrow$ inferMainTerm(μ_λ);
Initialize set $\mathcal{SM} \leftarrow \emptyset$;
Initialize set $\mathcal{NC} \leftarrow \emptyset$;
foreach $C_i \in C_\lambda$ **do**
  $\mu_i \leftarrow$ centroid(C_i);
  foreach $x \in C_i$ **do**
    $\mathcal{NC} \leftarrow \mathcal{NC} \cup x$;
    $E_{nc} \leftarrow E_{nc} \cup$ create_edge(x, μ_i);
  end
end
foreach $\mu_i \in \mu_\lambda$ **do**
  $E_\mu \leftarrow E_\mu \cup$ create_edge(μ_i, α_λ);
end
$E_K \leftarrow E_\mu \cup E_{nc}$;
$\mathcal{SM}_\lambda(\mathcal{K}) \leftarrow (\alpha_\lambda, E_K, \mu_\lambda, \mathcal{NC})$;
return $\mathcal{SM}_\lambda$;

5.5.1 Complexity analysis of the process of building a semantic map of a KG

The semantic distance matrix is a square matrix of size $n \times n$, where n represents the number of triples in the input KG. Building this matrix requires $n \times n$ calls to the function $sim()$, which has a

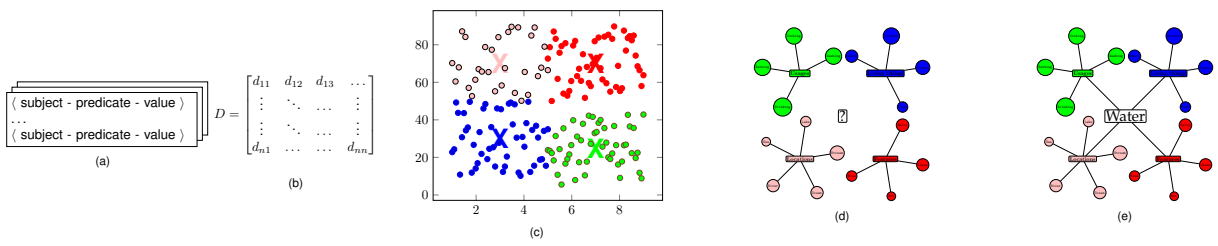


Figure 5.2: Phases of the process to build semantic maps. (a) Consume a KG as a list of n-triples, (b) Generate the semantic distance matrix D, (c) Cluster entities using the matrix D, (d) Infer main term α, and (e) Assemble the semantic map by connecting each centroid with α.

computational complexity of $O(N_1 \cdot N_2)$. In the worst-case scenario, both N_1 and N_2 may equal n, resulting in a complexity of $O(n^4)$ for constructing the semantic distance matrix. The complexity of computing the clusters depends 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 $O(k \cdot (n-k)^2)$, with k denoting the number of clusters. Inferring the central term α has a time complexity of $O(N \cdot t)$, where N represents the number of nodes in the entire KG (which is significantly larger than n), and t is the number of types. Creating the semantic map incurs a complexity of $O(n+k)$, involving the establishment of edges between each node and its centroid, as well as between each centroid and the central term. Thus, inferring the central term α contributes the most to the overall complexity of generating semantic maps for a KG, exhibiting a time complexity of $O(N \cdot t)$.

TABLE 5.1. SYMBOLS ASSOCIATED WITH SEMANTIC MAPS OF KNOWLEDGE GRAPHS.

Symbol	Description
λ	Centroid-based clustering algorithm used to group entities in the KG.
α_k	Main concept of the semantic map associated with the clusters inferred by λ.
μ_λ	Set of centroid entities produced by a centroid-based clustering algorithm λ, where $\mu \subseteq V$.
C_λ	Set of clusters resulting from running a centroid-based algorithm λ.
$\mathcal{NC}$	Non-centroids entities in KG, where $\mathcal{NC} \subseteq V$ and $\mathcal{NC} \cap \mu_\lambda = \emptyset$.
E_{nc}	Set of edges connecting all members of the clusters with their corresponding centroid, defined as $\mu_i \times x, \forall x \in \mathcal{NC}$ and $\forall \mu_i \in \mu_\lambda$.
E_μ	Set of edges connecting all centroids with the main term α_λ, defined as $E_{\mu_\lambda} = \mu_i \times \alpha, \forall \mu_i \in \mu_\lambda$.
E_K	Set of edges connecting each all elements in the semantic map, defined as $E_K = E_{\mu_\lambda} \bigcup E_{nc}$.

5.6. Conclusion

Our approach offers a novel approach to constructing a semantic map of a KG by leveraging the results of a clustering algorithm applied to KG nodes based on their semantic similarity. By exploiting the intrinsic semantic relationships among these nodes, our method enables the creation of a map that succinctly captures and visualizes the underlying semantic structure of the data. This methodology serves as a valuable tool for navigating complex datasets and enhancing the efficiency

of knowledge discovery processes.

In summary, Algorithm 3 represents the function ψ as mentioned in Contribution 1. The process to build semantic maps outlined in this chapter proposes a solution for reducing the size of a given KG by capturing semantic similarity among all pairs of entities within the KG. This approach involves replacing all edges in the original graph with edges connecting the centroid of each inferred group to all items within each cluster. As a result, the semantic map retains all vertices from the original graph but with a reduced number of edges, ensuring that each group represents entities that are semantically close.

6. Knowledge Graphs Visualization through Semantic Maps

This chapter aims to demonstrate the effectiveness of the proposed method in generating semantic maps described in Chapter 5 and showcasing how these maps can utilized to visualize KGs. We commence by outlining the Python framework developed to test our approach. Subsequently, we discuss the datasets utilized, obtained through a series of SPARQL queries. We then elaborate on the process of selecting hyperparameter values for the PAM and Affinity Propagation algorithms. Following this, we delve into the quality assessment of semantic maps of KGs, with a focus on the quantitative evaluation of clusters generated by the algorithms. Moreover, we introduce the centroid-based inference method for the term α. Finally, we offer a comprehensive analysis of the results obtained, discussing the implications and significance of our findings.

6.1. Building Semantic Maps Framework

The experiments were conducted using a Python 3 framework [1]. This framework relies on the Sematch framework [105] for executing SPARQL queries to the DBPedia public endpoint and computing the required similarity measure to generate the semantic distance matrix D. The function $sim(e_i, e_j)$, as described in Algorithm 1, is implemented through a SPARQL query to DBPedia. Once D is generated, the framework utilizes centroid-based clustering strategies, specifically PAM and Affinity Propagation, to produce the set of centroids μ and the set of non-centroid nodes $\mathcal{NC}$. The main term α is inferred by implementing Algorithm 2 to compute the central term. The shared types are obtained from a SPARQL query to DBPedia, following the path illustrated in Figure 6.1. Finally, the tool assembles the semantic map using Algorithm 3. The maps are generated using the `pyvis` library, which is a wrapper around the JavaScript `visJS` library.

[1] `https://github.com/pcamarillor/semantic_mapping` (Last visited: 2024-03-30)

6.1.1 Datasets

The datasets employed for validating the construction of semantic maps are acquired by executing SPARQL queries to the DBPedia via its public endpoint [2]. The results are stored in N-Triples format, where each dataset consists of a list of subject-predicate-object triples. Each dataset is designed to represent various knowledge domains accumulated in DBPedia and demonstrate how they can be condensed and visualized using semantic maps. Table 6.1 offers a summary of the datasets utilized in the experiments. While the datasets used in these experiments are not particularly large, they were carefully selected to align with the objectives of this study. The focus of our work is on demonstrating the effectiveness of the proposed method to extract semantic maps from KGs, which can be achieved with datasets of this size. Additionally, smaller datasets allow for more controlled experimentation and thorough analysis. Nonetheless, we acknowledge that larger datasets may provide further insights and leave this as a consideration for future research. The detailed SPARQL queries utilized to generate these datasets are documented in the repository containing the framework developed for conducting these experiments.

6.1.2 Hyperparameter selection

Determining the optimal number of clusters (k) is a crucial hyperparameter in the PAM clustering algorithm. We identify this parameter using the elbow method [93], a heuristic technique for selecting the ideal number of clusters in a dataset. The elbow method operates under the principle that as the number of clusters increases, the within-cluster sum of squares (WSS) decreases, indicating a reduction in the distance between each data point and its assigned center. The PAM algorithm

[2] `https://dbpedia.org/sparql/` (Last visited: 2024-03-30)

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

Figure 6.1: Template of the SPARQL query to get the list of types associated with each centroid.

TABLE 6.1. 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-fi 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

employs the elbow method to determine the optimal value of k, relying on the WSS metric, which decreases as the distance between each data point and its assigned center diminishes.

In contrast, for the Affinity Propagation algorithm, the preference value plays a crucial role in determining the number of clusters to generate. A higher preference value leads to the formation of more clusters, as more data points are chosen as exemplars. Conversely, a lower preference value results in fewer clusters, as fewer data points are selected as exemplars. Hence, it is essential to conduct sensitivity analysis by testing different values of the preference parameter to identify the optimal number of clusters.

Our methodology involves maximizing the silhouette index of the resulting clustering after running Affinity Propagation with preference values ranging from 0.1 to 0.9. This range covers the possible semantic distance values in the distance matrix. Table 6.2 outlines the process of selecting the $preference$ and k hyperparameters for the Affinity Propagation and PAM algorithms, respectively.

TABLE 6.2. HYPERPARAMETER SELECTION.

	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

6.2. Quality of Semantic Maps a Knowledge Graphs

The core of the process to generate semantic maps involves clustering entity instances and obtaining the set of centroids μ_λ. To provide a quantitative validation method for semantic maps, we propose associating the quality of the computed clusters with that of the semantic maps. This evaluation methodology enables us to determine the dependability of the groups described in the semantic map.

6.2.1 Cluster quality

The literature distinguishes between two categories of clustering validation measures: external clustering validation and internal clustering validation [1]. Internal validation metrics assess the quality of a clustering algorithm based on its intrinsic properties, while external validation methods evaluate the quality of a clustering solution by comparing it with known data labels. Given the absence of known labels for the datasets used in the experiments outlined in this study, we opt to utilize internal validation measures such as the Silhouette score [43], Davies-Bouldin score [21], and Calinski-Harabasz Index [13].

Each internal validation metric evaluates different aspects of the clusters. For instance, the Silhouette score quantifies how well each data point fits into its assigned cluster relative to other clusters [43]. The cluster's inertia, or within-cluster sum of squares (WSS) metric, assesses the den-

sity of the data points within each cluster [57]. The goal is to minimize inertia, which corresponds to maximizing the distances between clusters. Conversely, the Dunn index measures the distance between the nearest points in different clusters and the distance between the farthest points within each cluster [22]. Another well-recognized quality measure is the Davies-Bouldin index, which evaluates the similarity between each cluster and its closest neighboring cluster while also considering internal cluster similarity [21]. Finally, the Calinski-Harabasz index quantifies the ratio of between-cluster variance to within-cluster variance [13].

Contribution 2 (Quality of Semantic Maps a KGs) *Let $\mathcal{SM}_\lambda(K) = (\alpha_\lambda, E_K)$ represent the semantic map associated with the KG K and a clustering algorithm λ. We propose that the quality of $\mathcal{SM}_\lambda(K)$ be denoted by $\phi(E_K)$, where ϕ is the function that evaluates the cluster quality of the clusters captured in the set of edges E_K.*

6.3. Quality of semantic maps

Tables 6.3 and 6.4 presents three columns representing the quality of semantic maps generated by two centroid-based clustering algorithms (PAM and Affinity Propagation), as assessed by the 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 higher scores indicating better cluster quality. Conversely, the Davies-Bouldin index evaluates the ratio of within-cluster scatter to between-cluster separation, with lower scores indicating superior cluster quality. Similarly, the Calinski-Harabasz index quantifies the ratio of between-cluster variance to within-cluster variance, with higher scores suggesting improved cluster quality. Specifically, the silhouette score reflects the similarity of each entity to its own cluster versus other clusters, with scores approaching 1 indicating better cluster quality. Meanwhile, the Davies-Bouldin index quantifies the ratio of within-cluster scatter to between-cluster separation, where lower scores denote higher cluster quality. Lastly, the Calinski-Harabasz index evaluates the ratio of between-cluster variance to within-cluster variance, with higher scores indicating better cluster quality.

TABLE 6.3. QUALITY OF CLUSTERS PRODUCED BY THE PROCESS TO GENERATE SEMANTIC MAP FOR PAM ALGORITHM.

Dataset	Silhouette score	Davies-Bouldin score	Calinski-Harabasz Index
MOVIES_SCIFI	0.86	0.28	1366.07
FANTASY_NOVELS	0.66	3.73	133.56
CITIES	0.69	0.46	281.70
DISEASES	0.44	0.68	46.56
DRUGS	0.33	1.32	11.06
ACTORS	0.40	1.26	52.87
MOVIES-AND-ACTORS	0.55	0.67	103.26
DISEASES-AND-DRUGS	0.54	0.94	91.39

TABLE 6.4. QUALITY OF CLUSTERS PRODUCED BY THE PROCESS TO GENERATE SEMANTIC MAP FOR AFFINITY PROPAGATION ALGORITHM.

Dataset	Silhouette score	Davies-Bouldin score	Calinski-Harabasz Index
MOVIES_SCIFI	0.45	2.33	17.38
FANTASY_NOVELS	0.38	1.53	10.93
CITIES	0.47	1.27	77.45
DISEASES	0.43	2.84	5.52
DRUGS	-0.02	0.71	0.63
ACTORS	0.13	2.44	37.63
MOVIES-AND-ACTORS	0.54	1.41	85.44
DISEASES-AND-DRUGS	0.42	0.57	60.25

6.4. Discussion

The evaluation of the semantic map uses metrics that are already validated and widely used by the community, thereby offering a standardized method to evaluate our proposal. The use of a single metric to evaluate semantic maps is considered part of future work. The assessment of semantic map quality (Table 6.3 and Table 6.4) highlights the PAM algorithm's superior performance compared to the Affinity Propagation algorithm. Across all datasets, the PAM algorithm consis-

tently achieved higher silhouette scores, indicating clearer and more distinct clusters compared to the Affinity Propagation algorithm. Additionally, the Davies-Bouldin scores for the PAM algorithm suggested compact and well-separated clusters in 5 out of 8 datasets, while the Affinity Propagation algorithm exhibited significant overlap and poor separation. The Calinski-Harabasz index further confirmed the PAM algorithm's superiority in generating high-quality semantic maps, with notably higher scores than the Affinity Propagation algorithm across all datasets. Consequently, the PAM algorithm emerges as the preferred choice for producing semantic maps with superior separation and clarity.

7. Assessing the effectiveness of Semantic Maps in visualizing KGs

This chapter describes the evaluation process of the efficacy of semantic maps in summarizing KGs. The evaluation process implies inferring the primary terms of semantic maps and conducting a survey to gauge their effectiveness. The results revealed robust endorsement for semantic maps across various tasks and datasets. Specifically, semantic maps proved efficiency at locating requested items and enhancing comprehension of the core themes within KGs. In contrast, conventional visual representations received comparatively lower levels of support in similar tasks. These findings underscore the utility of semantic maps as a valuable tool for summarizing and visualizing KGs.

7.1. Qualitative assessment of Semantic Maps

One of the goals of graph summarization is to streamline the visual data exploration process [56]. In this study, we propose evaluating the effectiveness of the summarization process by examining how well semantic maps serve as a visualization approach for supporting visual exploratory tasks, as proposed in [15]. Traditional visual representations of KGs, known as *classical visual representations*, represent entities and relationships as nodes and edges, respectively. Nodes are labeled with the names of the entities they represent, while edges are labeled with the names of the relationships they depict. This visualization method can aid in exploring entity relationships and understanding the knowledge graph's structure.

7.2. Main term (α) inference

One of the primary characteristics of a semantic map is the prominent notion known as the **main term**, which embodies the principal theme depicted in the graphical representation. Within this document, we refer to this pivotal concept as α. In each experiment detailed within the chapter 6, the process to generate semantic map deduces the principal term α using the IC_{graph} metric

[104]. Table 7.1 outlines the inferred α for every dataset described in chapter 6.

TABLE 7.1. 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`

There are two specific cases that warrant analysis. In the `DISEASES.NT` dataset, the semantic map generated using the Affinity Propagation algorithm infers the main concept as `yago:AlimentCondition`. In contrast, the main term inferred using the PAM algorithm is the concept `yago:Disease114070360`. Similarly, in the `ACTORS.NT` experiment, the semantic map produced by the Affinity Propagation algorithm identifies the main concept as `yago:WikicastActors`, whereas the main concept inferred by the PAM algorithm is `yago:Actor109765278`. The discrepancy in the inferred main concepts stems from the fact that each clustering algorithm generates a distinct set of centroid elements (μ), directly impacting the inference of the main concept.

The hybrid datasets `MOVIES-AND-ACTORS.NT` and `DISEASES-AND-DRUGS.NT` serve to validate the process of inferring the term α when instances within datasets originate from different classes. However, our study revealed a particularly intriguing finding in the resulting semantic maps of these hybrid datasets. The resulting clusters exhibit a high level of coherence and meaningfulness.

7.3. Survey on Effectiveness of KG Visual Representations

This study evaluates the effectiveness of summarization by examining how well semantic maps support visual exploratory tasks compared to a node-link visual representation of the RDF vocabularies. To assess our approach's efficacy, we conducted a survey comprising three sections:

- The first section aims to understand the profile of the respondents.
- The second section employs a Likert scale [53] to measure the effectiveness of classical visual representations of KGs 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 KGs.

This section describes the survey [1] was administered to a cohort of 25 experts. Questions 3 to 8 prompted respondents to select one option from the following choices: *Strongly agree, Agree, Neutral, Disagree*, and *Strongly disagree*. The images displayed in the Appendix C are scaled-down versions of the original pictures given to participants.

a) 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.

[1] `https://ue8vg07pp4n.typeform.com/to/djZhvStq`

- I have some curiosity to explore and discover new insights from data.

b) How often do you use KGs in your work or studies?

- Never
- Rarely
- Sometimes
- Often
- Always

c) The following picture (see Figure 7.1) 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

d) The following picture (see Figure 7.2) 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

e) The following picture (see Figure 7.3) 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
- This representation helped me understand the relationships between different entities and concepts

f) The following picture (see Figure 7.4) 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

g) The following picture (see Figure 7.5) 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
- 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

h) The following picture (see Figure 7.6) 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

i) Which method do you find easier to use?

   - Semantic Maps
   - Classic Visual Representations

j) Which method do you think is more effective in representing KGs?

   - Semantic Maps
   - Classic Visual Representations

7.4. Results of effectiveness of semantic maps as strategy to visualize KGS

Concerning the profiles of the experts engaged in this study, we introduced six distinct profiles associated with the fields of knowledge discovery, artificial intelligence, and data science. Participants had the option to select more than one profile. Based on the gathered responses, it was 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 KGs in their daily duties:

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

Tables 7.2, 7.3, and 7.4 illustrates the outcomes concerning effectiveness for `CITIES.NT`, `DISEASES-AND-DRUGS.NT, and MOVIES-AND-ACTORS.NT` datasets respectively, employing a Likert scale to evaluate the effectiveness of two different visual representations of KGs across three exploratory tasks related to visual data. When participants were asked about their preference for ease of use, 76.9% favored semantic maps, while 23.1% opted for classic visual representations. Regarding effectiveness in representing KGs, 88.5% of experts chose semantic maps, while 11.5% preferred classic visual representation.

Exploratory tasks definition:

- **01**: Search for one specific item
- **02**: Identification of main term
- **03**: Explore and comprehend KGs

7.5. Conclusions

In our survey on effectiveness, we investigated three exploratory tasks related to KGs to assess the efficacy of semantic maps compared to classical visual representations across various datasets. Let's delve into the key findings:

a) **Search for a Specific Item**: Semantic maps consistently received strong support in locating requested items. In the `DISEASES-AND-DRUGS.NT` dataset, 20% of participants strongly

TABLE 7.2. EFFECTIVENESS ASSESSMENT RESULTS FOR VISUALIZATION ON DATASET `CITIES.NT`

Task	Visualization Strategy	Rating	
Search for one specific item	Node-link visualization	20%	76%
	Semantic Maps	80%	8%
Identification of main term	Node-link visualization	56%	28%
	Semantic Maps	84%	4%
Explore and comprehend KGs	Node-link visualization	42%	33%
	Semantic Maps	72%	16%

Strongly agree Agree Neutral Disagree Strongly Disagree

TABLE 7.3. EFFECTIVENESS ASSESSMENT RESULTS FOR VISUALIZATION ON DATASET `DISEASES-AND-DRUGS.NT`

Task	Visualization Strategy	Rating	
Search for one specific item	Node-link visualization	52%	28%
	Semantic Maps	84%	4%
Identification of main term	Node-link visualization	72%	12%
	Semantic Maps	60%	12%
Explore and comprehend KGs	Node-link visualization	75%	8%
	Semantic Maps	67%	13%

Strongly agree Agree Neutral Disagree Strongly Disagree

agreed that semantic maps effectively found the requested item. Similarly, in the `MOVIES-AND-ACTORS.NT` dataset, 24% found semantic maps useful for search tasks. In contrast, classical visual representation garnered 8% and 12% endorsement for the same tasks in the

respective datasets. For the `CITIES.NT` dataset, 20% favored semantic maps, while only 8% found classical visualization helpful.

b) **Identification of the Main Term:**s For identifying the main topic of the KG, 44% of participants in the `DISEASES-AND-DRUGS.NT` dataset found classical visual representation effective, while 36% favored semantic maps. In the `MOVIES-AND-ACTORS.NT` dataset, 48% preferred semantic maps, whereas 32% relied on classical visual representation. For the `CITIES.NT` dataset, 44% leaned toward semantic maps, while 24% opted for classical visual representation.

c) **Exploration and Comprehension of KGs:** In this task, semantic maps facilitated understanding, with 20.8% of participants in the `DISEASES-AND-DRUGS.NT` dataset strongly agreeing. Similarly, in the `MOVIES-AND-ACTORS.NT` and `CITIES.NT` datasets, 41.7% of participants endorsed semantic maps for this task. In contrast, classical visual representation received varying levels of support: 12.5%, 25%, and 16.7% for the respective datasets.

These findings highlight the varied utility of semantic maps and visual representation in identifying the main topics and facilitating exploration of RDF vocabularies across diverse domains.

TABLE 7.4. EFFECTIVENESS ASSESSMENT RESULTS FOR VISUALIZATION ON DATASET `MOVIES-AND-ACTORS.NT`

Task	Visualization Strategy	Rating		
Search for one specific item	Node-link visualization	60%		20%
	Semantic Maps	64%		24%
Identification of main term	Node-link visualization	68%		12%
	Semantic Maps	68%		16%
Explore and comprehend KGs	Node-link visualization	79%		9%
	Semantic Maps	80%		8%

Conclusiones Generales

En este documento de tesis doctoral se ha presentado una estrategia para extraer mapas semánticos de un grafo de conocimiento y de esta forma facilitar el proceso de análisis y visualización de este tipo de estructuras.

El uso de los grafos de conocimiento en sistemas de inteligencia artificial, sistemas de respuestas automatizadas y procesamiento de lenguaje natural se ha incrementado en los últimos años. La extracción de versiones reducidas de estas estructuras es indispensable para realizar tareas de análisis y visualización de grafos de conocimiento que cada día incrementan su tamaño.

La propuesta descrita en este documento de tesis doctoral consiste en extraer mapas semánticos de un grafo de conocimiento. Se han descrito experimentos empleando diversos grafos de conocimiento extraídos de DBPedia para validar la estrategia propuesta en este trabajo de tesis doctoral. Basándonos en los resultados de estos experimentos, podemos concluir lo siguiente:

a) Los mapas semánticos actúan como herramientas para ilustrar las relaciones entre conceptos o términos dentro de un dominio o campo de estudio específico. Al organizar y mostrar visualmente las conexiones semánticas entre varias entidades, estos mapas destacan tanto las similitudes como las diferencias, proporcionando una visión de cómo están interconectadas las ideas.

b) El proceso de generación de mapas semánticos ofrece una solución para simplificar un KG dado al identificar similitudes semánticas entre todos los pares de entidades dentro del KG.

c) La evaluación de la calidad de los mapas semánticos revela que el algoritmo PAM supera al algoritmo de Propagación de Afinidad. En todos los conjuntos de datos, el algoritmo PAM consistentemente logra puntajes de silueta más altos, lo que indica que produce grupos más claros y distintos en comparación con el algoritmo de Propagación de Afinidad.

d) La encuesta presentada para validar la efectividad de los mapas semánticos muestra que son útiles para representar visualmente e identificar temas clave y facilitar la exploración de vocabularios RDF en diversos dominios.

Como trabajo futuro, planeamos evaluar nuestra técnica de resumén propuesta mediante la reducción del error de reconstrucción ℓ_p y las métricas de calidad del error de norma de corte, tal como lo propusieron Riondato et al.[78].

En resumen, nuestro trabajo integra eficazmente algoritmos de agrupación basados en centroides y el método de inferencia de términos principales para generar mapas semánticos para la visualización de KGs. Este enfoque proporciona una comprensión integral de los datos al combinar evaluaciones cualitativas y cuantitativas. Creemos que nuestros hallazgos hacen una contribución significativa al campo, permitiendo a los investigadores y profesionales visualizar y analizar KGs con mayor claridad e interpretabilidad.

General Conclusions

This doctoral thesis document presents a strategy for extracting semantic maps from a KG, thereby facilitating the analysis and visualization of such structures.

The use of knowledge graphs in artificial intelligence systems, automated response systems, and natural language processing has increased in recent years. The extraction of reduced versions of these structures is indispensable for performing analysis and visualization tasks on knowledge graphs that are growing in size each day.

The proposal described in this doctoral thesis consists of extracting semantic maps from a KG. Experiments using various KGs extracted from DBPedia have been described to validate the strategy proposed in this doctoral thesis. Based on the results of these experiments, we can conclude the following:

a) Semantic maps act as tools to illustrate the relationships between concepts or terms within a specific domain or field of study. By visually organizing and displaying the semantic connections among various entities, these maps highlight both similarities and distinctions, providing insights into how ideas are interconnected.

b) The process of producing a semantic map offers a solution for simplifying a given KG by identifying semantic similarities among all pairs of entities within the KG.

c) The evaluation of semantic map quality reveals that the PAM algorithm outperforms the Affinity Propagation algorithm. Across all datasets, the PAM algorithm consistently achieves higher silhouette scores, indicating clearer and more distinct clusters compared to those produced by the Affinity Propagation algorithm.

d) The survey presented to validate the effectiveness of semantic maps shows that they are useful for visually representing and identifying key topics and facilitating the exploration of RDF vocabularies across various domains.

As future work, we plan to evaluate our proposed summarization technique by reducing the ℓ_p-reconstruction error and the cut-norm error quality metrics, as proposed by Riondato et al.[78].

In summary, our work effectively integrates centroid-based clustering algorithms and and main term inference method to generate semantic maps for visualizing KGs. This approach provides a comprehensive understanding of the data by combining both qualitative and quantitative assessments. We believe our findings make a significant contribution to the field, enabling researchers and practitioners to visualize and analyze KGs with enhanced clarity and interpretability.

Appendix

A. List of Internal Research Reports

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Towards a visualization technique for large time-evolving graphs: challenges and opportunities," Internal Report PhDEngScITESO-18-33-R, ITESO, Tlaquepaque, Mexico, Dec. 2018.

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Designing a strategy to evaluate the visual representation of Dynamic networks," Internal Report PhDEngScITESO-20-03-R, ITESO, Tlaquepaque, Mexico, Mar. 2020.

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Exploratory data analysis for temporal networks: tools and libraries," Internal Report PhDEngScITESO-20-05-R, ITESO, Tlaquepaque, Mexico, Apr. 2020.

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Evaluation of the visual representation of dynamic networks provided by Gephi and Cytsocape," Internal Report PhDEngScITESO-20-15-R, ITESO, Tlaquepaque, Mexico, Aug. 2020.

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Graph mining; clustering analysis," Internal Report PhDEngScITESO-21-05-R, ITESO, Tlaquepaque, Mexico, Aug. 2021.

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Dectect comorbities clusters associated with Covid-19 in Mexico by using a graph-based approach," Internal Report PhDEngScITESO-21-25-R, ITESO, Tlaquepaque, Mexico, Dec. 2021.

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Knowledge graphs: Summarization and visualization," Internal Report PhDEngScITESO-21-26-R, ITESO, Tlaquepaque, Mexico, Dec. 2021.

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Exploring the semantic properties of knowledge graphs using semantic similarity and semantic maps," Internal Report PhDEngScITESO-22-20-R, ITESO, Tlaquepaque, Mexico, Nov. 2022.

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Defining the process to extract a semantic map from a knowledge graph," Internal Report PhDEngScITESO-23-02-R, ITESO, Tlaquepaque, Mexico, Feb. 2023.

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Quantitative assessment of the process to generate semantic maps of knowledge graphs," Internal Report PhDEngScITESO-23-07-R,, ITESO, Tlaquepaque, Mexico, Nov. 2023.

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, "Evaluating the effectiveness of semantic maps to summarize knowledge graphs," Internal Report PhDEngScITESO-23-12-R, ITESO, Tlaquepaque, Mexico, Dec. 2023.

B. List of Publications

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, ”A task-based evaluation methodology for visual representation of dynamic networks,” EDBT: 23rd International Conference on Extending Database Technology, Copenhagen, Denmark, 2020, [Online]. Available: `https://ceur-ws.org/Vol-2578/BigVis10.pdf`

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, ”A graph-based approach to cluster comorbidities from Mexican patients of COVID-19,” 19th Mexican International Conference on Artificial Intelligence (MICAI), Mexico, 2020, Vol 149 (9).

- P. Camarillo-Ramírez, F. Cervantes-Álvarez, and L. F. Gutiérrez-Preciado, ”Semantic Maps for Knowledge Graphs: A Semantic-based Summarization Approach,” in *IEEE Access*, vol. 12, pp. 6729-6744, 2024, doi: 10.1109/ACCESS.2024.3351170

C. Semantic Maps

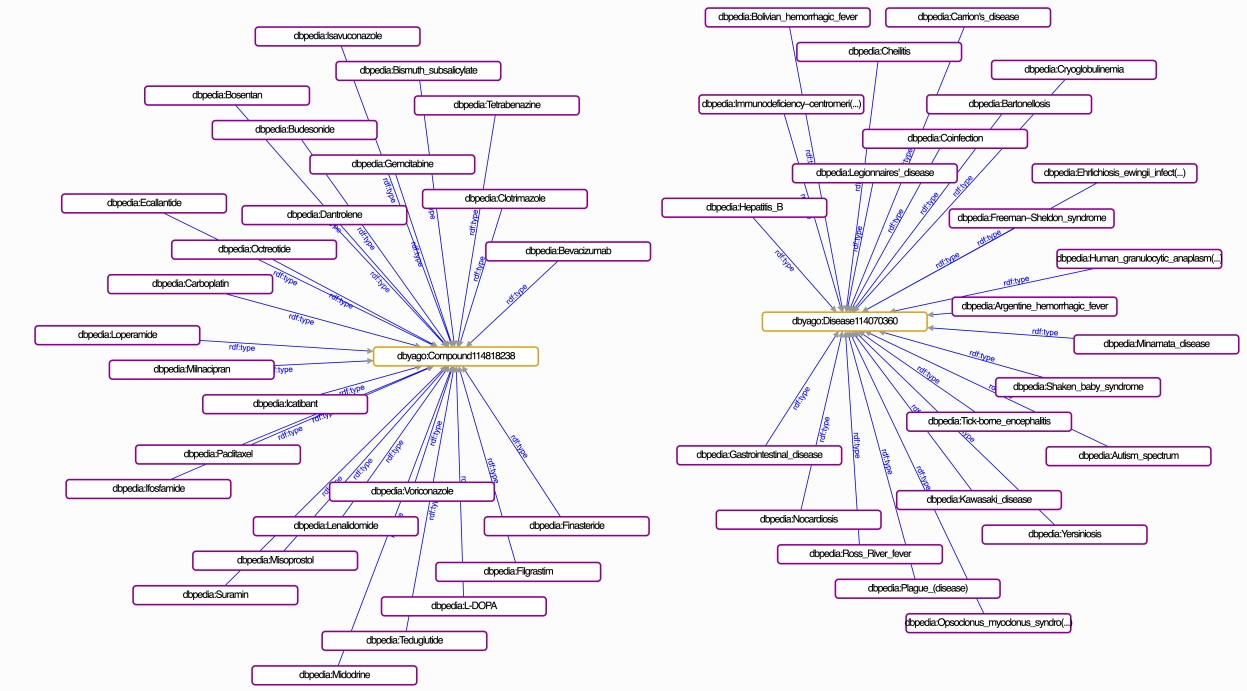


Figure 7.1: Dataset `DISEASES-AND-DRUGS.NT` represented using RDF Visualizer online tool.

Figure 7.2: Dataset `DISEASES-AND-DRUGS.NT` represented using the generated semantic map.

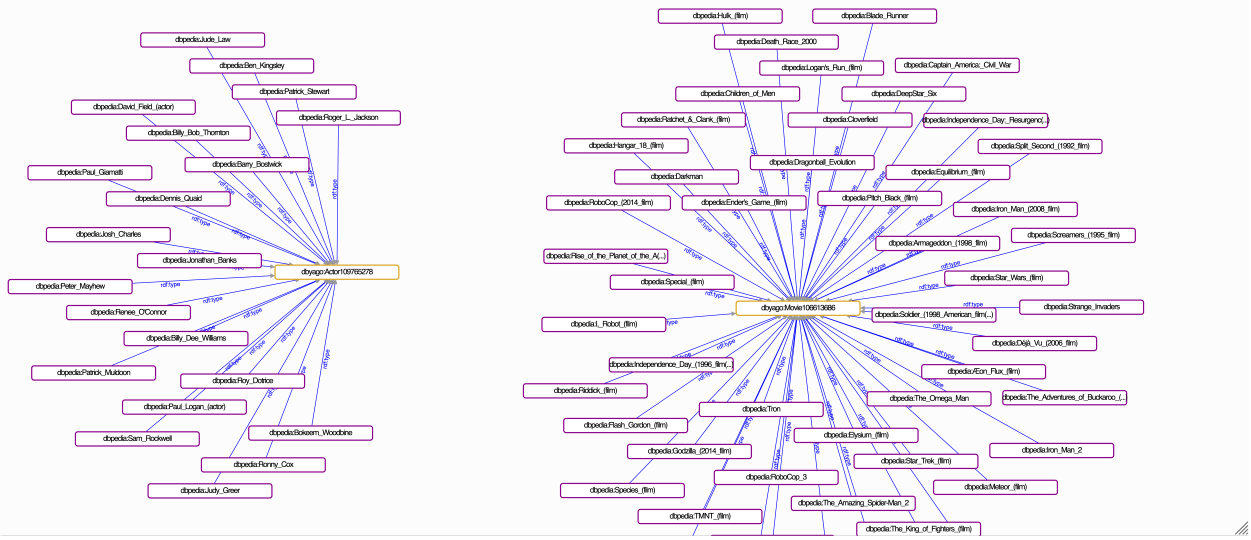


Figure 7.3: Dataset `ACTORS-AND-MOVIES.NT` represented using RDF Visualizer online tool.

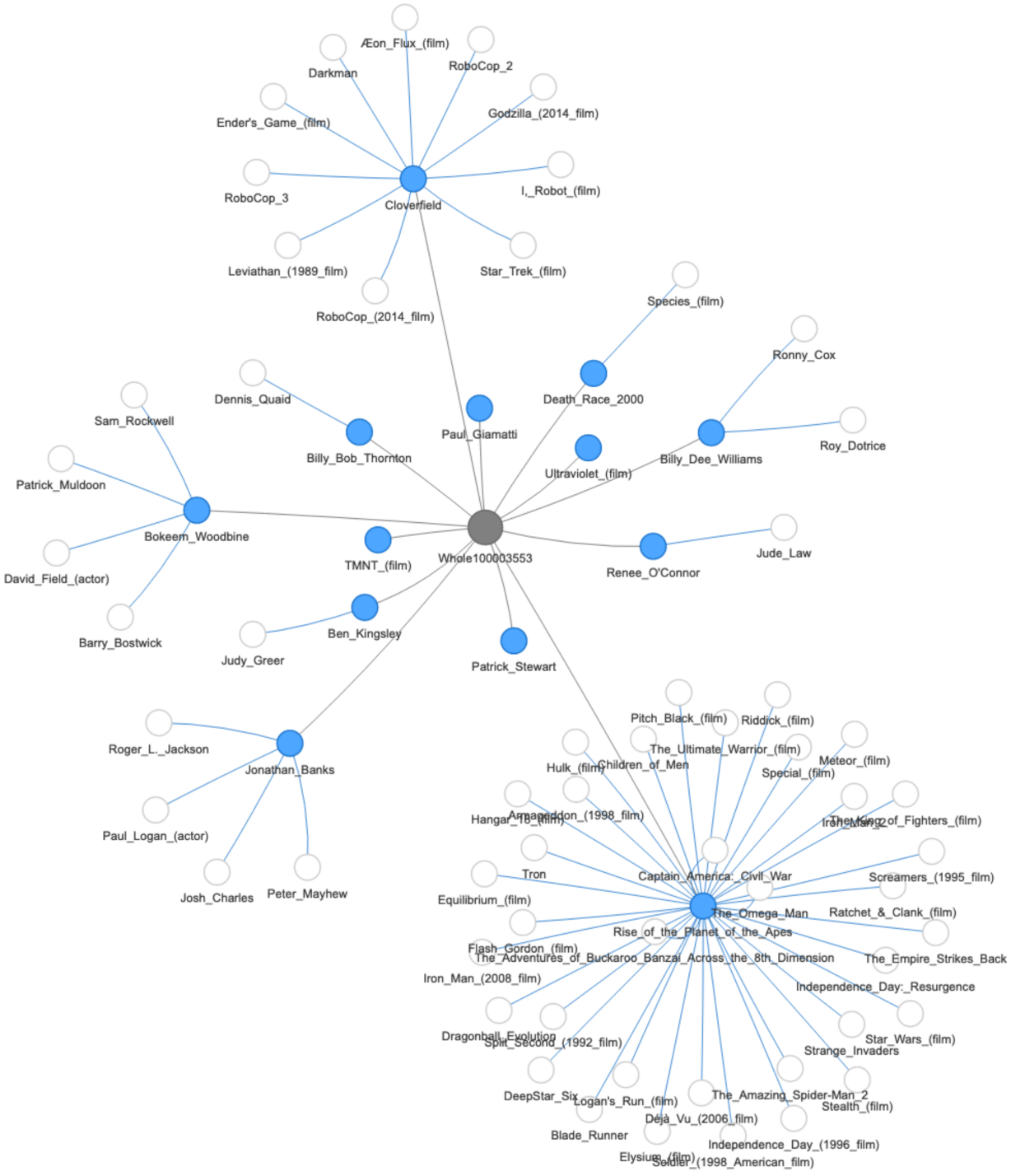


Figure 7.4: Dataset `ACTORS-AND-MOVIES.NT` represented using the generated semantic map.

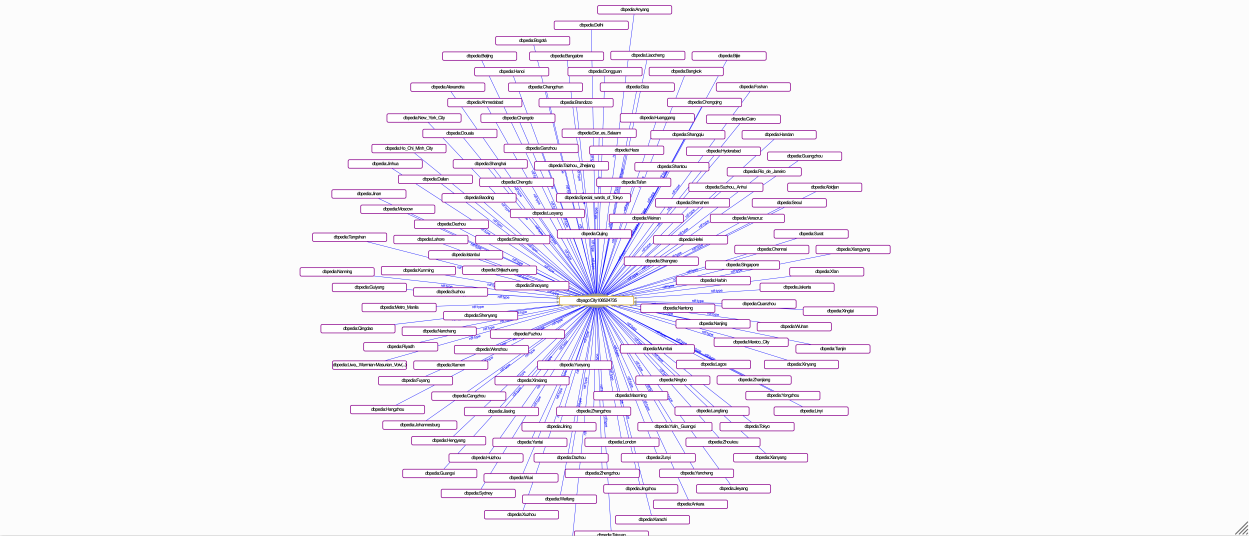


Figure 7.5: Dataset `CITIES.NT` represented using RDF Visualizer online tool.

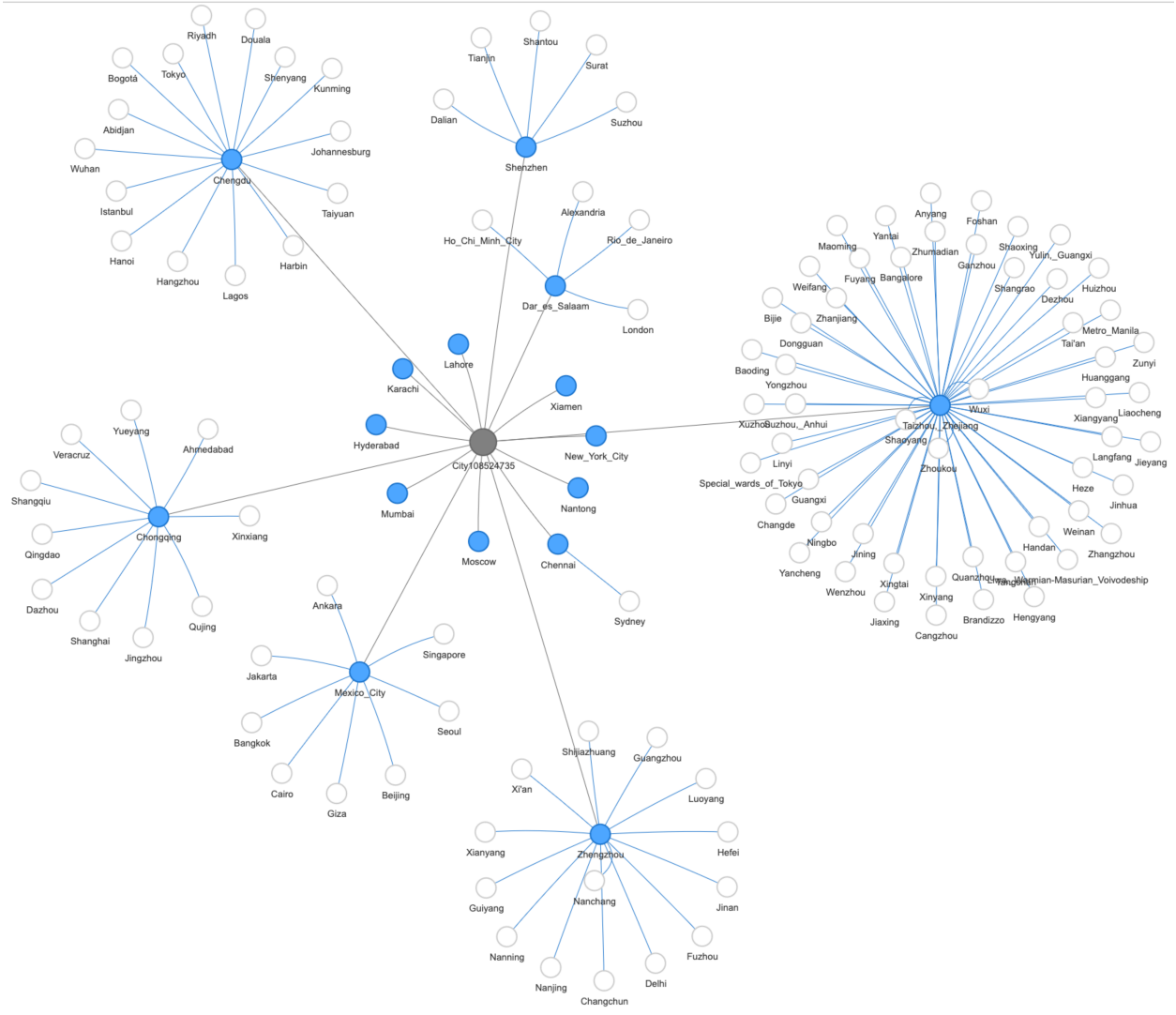


Figure 7.6: Dataset `CITIES.NT` represented using the generated semantic map.

Bibliography

[1] Charu C. Aggarwal and Chandan K. Reddy, eds. *Data Clustering: Algorithms and Applications*. CRC Press, 2014. ISBN: 978-1-46-655821-2. URL: http://www.crcpress.com/product/isbn/9781466558212.

[2] Daniel Archambault and Helen C. Purchase. "The "Map" in the mental map: Experimental results in dynamic graph drawing". In: *International Journal of Human-Computer Studies* 71.11 (2013), pp. 1044–1055. ISSN: 1071-5819. DOI: j.ijhcs.2013.08.004.

[3] Sören Auer et al. "DBpedia: A Nucleus for a Web of Open Data". In: *The Semantic Web*. Ed. by Karl Aberer et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 722–735. ISBN: 978-3-540-76298-0.

[4] Franz Baader. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003. ISBN: 9780521781763.

[5] Franz Baader, Ian Horrocks, and Ulrike Sattler. "Description Logics". In: *Handbook of Knowledge Representation*. Ed. by F. van Harmelen, V. Lifschitz, and B. Porter. Amsterdam, The Netherlands: Elsevier, 2008. Chap. 3, pp. 135–179.

[6] Franz Baader and Ulrike Sattler. "An Overview of Tableau Algorithms for Description Logics". In: *Studia Logica* 69.1 (2001), pp. 5–40. DOI: 10.1023/A:1013882326814. URL: https://doi.org/10.1023/A:1013882326814.

[7] David Baxter et al. "An engineering design knowledge reuse methodology using process modelling". In: *Research in Engineering Design* 18.1 (2007), pp. 37–48. DOI: 10.1007/s00163-007-0028-8. URL: https://doi.org/10.1007/s00163-007-0028-8.

[8] François Belleau et al. "Bio2RDF: towards a mashup to build bioinformatics knowledge systems". In: *Journal of biomedical informatics* 41.5 (2008), pp. 706–716. DOI: 10.1016/j.jbi.2008.03.004.

[9] Tim Berners-Lee, James Hendler, and Ora Lassila. "The Semantic Web". In: *Scientific American* 284.5 (2001), pp. 34–43. ISSN: 00368733, 19467087. URL: http://www.jstor.org/stable/26059207 (visited on 09/03/2023).

[10] Kurt Bollacker et al. "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge". In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. SIGMOD '08. Vancouver, Canada: Association for Computing Machinery, 2008, pp. 1247–1250. ISBN: 9781605581026. DOI: 10.1145/1376616.1376746.

[11] Agustín Borrego et al. "Completing Scientific Facts in Knowledge Graphs of Research Concepts". In: *IEEE Access* 10 (2022), pp. 125867–125880. DOI: 10.1109/ACCESS.2022.3220241.

[12] Josep Maria Brunetti et al. "Formal linked data visualization model". In: *Proceedings of International Conference on Information Integration and Web-based Applications & Services*. 2013, pp. 309–318.

[13] Tadeusz Calinski and Jerzy Harabasz. "Dendrite method for cluster analysis". In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[14] Diego Calvanese et al. "Reasoning in Expressive Description Logics". In: *Handbook of Automated Reasoning (in 2 volumes)*. Ed. by John Alan Robinson and Andrei Voronkov. Amsterdam, The Netherlands: Elsevier and MIT Press, 2001. Chap. 23, pp. 1581–1634. DOI: 10.1016/B978-044450813-3/50025-4. URL: https://doi.org/10.1016/b978-044450813-3/50025-4.

[15] Pablo Camarillo-Ramirez, Francisco Cervantes-Alvarez, and Luis Fernando Gutiérrez-Preciado. "A task-based evaluation methodology for visual representation of dynamic networks." In: *EDBT/ICDT Workshops*. Copenhagen, Denmark, Mar. 2020. URL: http://ceur-ws.org/Vol-2578/BigVis10.pdf.

[16] Andrew Carlson et al. "Toward an Architecture for Never-Ending Language Learning". In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI'10. Atlanta, Georgia: AAAI Press, 2010, pp. 1306–1313.

[17] Yi-Hui Chen, Eric Jui-Lin Lu, and Ying-Yen Lin. "Efficient SPARQL Queries Generator for Question Answering Systems". In: *IEEE Access* 10 (2022), pp. 99850–99860. DOI: 10.1109/ACCESS.2022.3206794.

[18] Kenneth Church and Patrick Hanks. "Word association norms, mutual information, and lexicography". In: *Computational linguistics* 16.1 (1990), pp. 22–29.

[19] *Clustering Algorithms*. Last visited: 2024-03-29. July 2022. URL: https://developers.google.com/machine-learning/clustering/clustering-algorithms.

[20] William Croft. "On two mathematical representations for semantic maps". In: *Zeitschrift für Sprachwissenschaft* 41.1 (2022), pp. 67–87. DOI: https://doi.org/10.1515/zfs-2021-2040.

[21] David L Davies and Donald W Bouldin. "A cluster separation measure". In: *IEEE transactions on pattern analysis and machine intelligence* 1.2 (1979), pp. 224–227.

[22] J. C. Dunn. "Well-separated clusters and optimal fuzzy partitions". In: *Journal of Cybernetics* 4.1 (1974), pp. 95–104.

[23] Facebook Engineering. *Under the Hood: The Entities Graph*. 2013. URL: https://www.facebook.com/notes/facebook-engineering/under-the-hood-the-entities-graph/10151490531588920.

[24] Fredo Erxleben et al. "Introducing Wikidata to the Linked Data Web". In: *The Semantic Web – ISWC 2014*. Ed. by Peter Mika et al. Cham: Springer International Publishing, 2014, pp. 50–65. ISBN: 978-3-319-11964-9.

[25] Sabhia Firdaus and Md Ashraf Uddin. "A survey on clustering algorithms and complexity analysis". In: *International Journal of Computer Science Issues (IJCSI)* 12.2 (2015), p. 62.

[26] Brendan J Frey and Delbert Dueck. "Clustering by passing messages between data points". In: *science* 315.5814 (2007), pp. 972–976.

[27] Lars Marius Garshol. "Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it all". In: *Journal of Information Science* 30 (2004), pp. 378–391.

[28] Thanasis Georgakopoulos. *Semantic Maps*. Last visited: 2023-11-16. URL: https://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0229.xml.

[29] John S. Gero and Udo Kannengiesser. "A function–behavior–structure ontology of processes". In: *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 21.4 (2007), pp. 379–391. DOI: 10.1017/S0890060407000340.

[30] Ashok K. Goel, Spencer Rugaber, and Swaroop Vattam. "Structure, behavior, and function of complex systems: The structure, behavior, and function modeling language". In: *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 23.1 (2009), pp. 23–35. DOI: 10.1017/S0890060409000080.

[31] Juan Gómez-Romero et al. "Visualizing large knowledge graphs: A performance analysis". In: *Future Generation Computer Systems* 89 (2018), pp. 224–238.

[32] Kainan Guan, Liang Du, and Xinhua Yang. "Relationship Extraction and Processing for Knowledge Graph of Welding Manufacturing". In: *IEEE Access* 10 (2022), pp. 103089–103098. DOI: 10.1109/ACCESS.2022.3209066.

[33] Kalpa Gunaratna. "Semantics-based summarization of entities in knowledge graphs". PhD thesis. Wright State University, 2017.

[34] Liang Guo et al. "An automatic method for constructing machining process knowledge base from knowledge graph". In: *Robotics and Computer-Integrated Manufacturing* 73 (2022), p. 102222. ISSN: 0736-5845. DOI: https://doi.org/10.1016/j.rcim.2021.102222. URL: https://www.sciencedirect.com/science/article/pii/S0736584521001058.

[35] Aidan Hogan et al. *Knowledge Graphs*. Vol. 54. 4. New York, NY, USA: Association for Computing Machinery, 2021.

[36] Ian Horrocks. "Ontologies and the semantic web". In: *Communications of the ACM* 51.12 (2008), pp. 58–67.

[37] Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. "Collaboratively built semi-structured content and Artificial Intelligence: The story so far". In: *Artificial Intelligence* 194 (2013), pp. 2–27. ISSN: 0004-3702. DOI: 10.1016/j.artint.2012.10.002.

[38] Ioana Hulpuş, Narumol Prangnawarat, and Conor Hayes. "Path-based semantic relatedness on linked data and its use to word and entity disambiguation". In: *Proc. 14th Int. Semantic Web Conference 2015*. Springer. Bethlehem, PA, USA, 2015, pp. 442–457.

[39] ISO. *Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts*. Standard. Geneva, CH: International Organization for Standardization, 2018.

[40] Rricha Jalota et al. "LAUREN - Knowledge Graph Summarization for Question Answering". In: *2021 IEEE 15th Int. Conf. on Semantic Computing (ICSC)*. Laguna Hills, CA, USA, Jan. 2021, pp. 221–226. DOI: 10.1109/ICSC50631.2021.00047.

[41] Jay J Jiang and David 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.

[42] Dale D Johnson, Susan D Pittelman, and Joan E Heimlich. "Semantic mapping". In: *The reading teacher* 39.8 (1986), pp. 778–783.

[43] Leonard Kaufman and Peter J Rousseeuw. "Finding Groups in Data: An Introduction to Cluster Analysis." In: Hoboken, New Jersey: John Wiley & Sons, Inc., 1990. Chap. Partitioning around medoids (program PAM).

[44] Leonard Kaufman and Peter J Rousseeuw. "Finding Groups in Data: An Introduction to Cluster Analysis." In: Hoboken, New Jersey: John Wiley & Sons, Inc., 1990. Chap. Clustering Large Applications (program CLARA).

[45] Catherine Kearns. *Credit card fraud detection using machine learning*. 2024. URL: https://cambridge-intelligence.com/detect-credit-card-fraud-with-network-visualization/ (visited on 05/01/2024).

[46] Daniel A Keim. "Visual exploration of large data sets". In: *Communications of the ACM* 44.8 (2001), pp. 38–44.

[47] Danai Koutra et al. “Vog: Summarizing and understanding large graphs”. In: *Proceedings of the 2014 SIAM international conference on data mining*. SIAM. Philadelphia, PA, USA, 2014, pp. 91–99.

[48] Thomas K Landauer and Susan T Dumais. “A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.” In: *Psychological review* 104.2 (1997), p. 211.

[49] Claudia Leacock and Martin Chodorow. “Combining local context and WordNet similarity for word sense identification”. In: *WordNet: An electronic lexical database* 49.2 (1998), pp. 265–283.

[50] Edward A. Lee. “Model-Driven Development - From Object-Oriented Design to Actor-Oriented Design”. In: *Proceedings of the Workshop for Software Engineering for Embedded Systems, From Requirements to Implementation*. Chicago, IL, US, Sept. 2003. URL: https://api.semanticscholar.org/CorpusID:17172424.

[51] Haotian Li et al. “KG4Vis: A Knowledge Graph-Based Approach for Visualization Recommendation”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2022), pp. 195–205. DOI: 10.1109/TVCG.2021.3114863.

[52] Vladimir Lifschitz, Leora Morgenstern, and David Plaisted. “Knowledge Representation and Classical Logic”. In: *Handbook of Knowledge Representation*. Ed. by F. van Harmelen, V. Lifschitz, and B. Porter. Amsterdam, The Netherlands: Elsevier, 2008. Chap. 1, pp. 3–74.

[53] Rensis Likert. In: *Archives of Psychology* 22.1932 (1932), pp. 5–55.

[54] Dekang 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.

[55] Yanru Lin et al. “A Recommendation Strategy Integrating Higher-Order Feature Interactions With Knowledge Graphs”. In: *IEEE Access* 10 (2022), pp. 119290–119300. DOI: 10.1109/ACCESS.2022.3220322.

[56] Yike Liu et al. “Graph summarization methods and applications: A survey”. In: *ACM Computing Surveys (CSUR)* 51.3 (2018), pp. 1–34. DOI: 10.1145/3186727.

[57] J. MacQueen. "Some methods for classification and analysis of multivariate observations". In: *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. The Regents of the University of California. Berkeley, CA, US, Jan. 1967, pp. 281–297.

[58] Michael Martin et al. "CubeViz: Exploration and Visualization of Statistical Linked Data". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: Association for Computing Machinery, 2015, pp. 219–222. ISBN: 9781450334730. DOI: 10.1145/2740908.2742848. URL: https://doi.org/10.1145/2740908.2742848.

[59] John McCarthy. "Programs with Common Sense". In: *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*. London: Her Majesty's Stationary Office, 1959, pp. 75–91. URL: http://www-formal.stanford.edu/jmc/mcc59.html.

[60] Dan McGinn et al. "Visualizing dynamic bitcoin transaction patterns". In: *Big data* 4.2 (2016), pp. 109–119.

[61] Deborah L. McGuinness and Frank van Harmelen. *5 Trends Appear on the Gartner Hype Cycle for Emerging Technologies, 2019*. 2004. URL: https://www.w3.org/TR/owl-features/.

[62] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. "Corpus-based and knowledge-based measures of text semantic similarity". In: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*. AAAI'06. Boston, Massachusetts: AAAI Press, 2006, pp. 775–780. ISBN: 9781577352815.

[63] Miguel Molina-Solana, David Birch, and Yi-ke Guo. "Improving data exploration in graphs with fuzzy logic and large-scale visualisation". In: *Applied Soft Computing* 53 (2017), pp. 227–235.

[64] Vassil Momtchev et al. "Expanding the pathway and interaction knowledge in linked life data". In: *Proc. of International Semantic Web Challenge*. 2009.

[65] John Mylopoulos. "An overview of knowledge representation". In: *ACM SIGART Bulletin* 74 (1980), pp. 5–12.

[66] David Newman et al. "Visualizing search results and document collections using topic maps". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 8.2 (2010), pp. 169–175. ISSN: 1570-8268. DOI: https://doi.org/10.1016/j.websem.2010.03.005. URL: https://www.sciencedirect.com/science/article/pii/S1570826810000211.

[67] Raymond T. Ng and Jiawei Han. "CLARANS: A method for clustering objects for spatial data mining". In: *IEEE transactions on knowledge and data engineering* 14.5 (2002), pp. 1003–1016.

[68] Maximilian Nickel et al. "A review of relational machine learning for knowledge graphs". In: *Proceedings of the IEEE* 104.1 (2015), pp. 11–33.

[69] Charalampos Nikolaou et al. "Sextant: Visualizing time-evolving linked geospatial data". In: *Journal of Web Semantics* 35 (2015), pp. 35–52. ISSN: 1570-8268. DOI: https://doi.org/10.1016/j.websem.2015.09.004. URL: https://www.sciencedirect.com/science/article/pii/S1570826815000864.

[70] Jeff Pan et al. "Large Language Models and Knowledge Graphs: Opportunities and Challenges". In: *Transactions on Graph Data and Knowledge* (June 2023). DOI: 10.48550/arXiv.2308.06374. URL: https://hal.science/hal-04370111.

[71] David Paulius and Yu Sun. "A Survey of Knowledge Representation in Service Robotics". In: *Robotics and Autonomous Systems* 118 (2019), pp. 13–30. ISSN: 0921-8890. DOI: https://doi.org/10.1016/j.robot.2019.03.005. URL: https://www.sciencedirect.com/science/article/pii/S0921889018303506.

[72] Ciyuan Peng et al. "Knowledge Graphs: Opportunities and Challenges". In: *Artificial Intelligence Review* 56.11 (2023), pp. 13071–13102. DOI: 10.1007/s10462-023-10465-9. URL: https://doi.org/10.1007/s10462-023-10465-9.

[73] Hemant Purohit, Valerie L. Shalin, and Amit P. Sheth. "Knowledge Graphs to Empower Humanity-Inspired AI Systems". In: *IEEE Internet Computing* 24.4 (2020), pp. 48–54. DOI: 10.1109/MIC.2020.3013683.

[74] Richard Qian. *Understand Your World with Bing*. 2013. URL: https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/.

[75] R. Rada et al. "Development and application of a metric on semantic nets". In: *IEEE Transactions on Systems, Man, and Cybernetics* 19.1 (1989), pp. 17–30. DOI: 10.1109/21.24528.

[76] P. Resnik. "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language". In: *Journal of Artificial Intelligence Research* 11 (July 1999), pp. 95–130. ISSN: 1076-9757. DOI: 10.1613/jair.514. URL: http://dx.doi.org/10.1613/jair.514.

[77] Philip 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.

[78] Mateo Riondato, David García-Soriano, and Francesco Bonchi. "Graph summarization with quality guarantees". In: *Proc. of 2014 IEEE Int. Conf. on Data Mining*. Shenzhen, China, Dec. 2014, pp. 947–952. DOI: 10.1109/ICDM.2014.56.

[79] J. A. Robinson. "A Machine-Oriented Logic Based on the Resolution Principle". In: *J. ACM* 12.1 (Jan. 1965), pp. 23–41. ISSN: 0004-5411. DOI: 10.1145/321250.321253. URL: https://doi.org/10.1145/321250.321253.

[80] Alan Ruttenberg et al. "Life sciences on the Semantic Web: The Neurocommons and beyond". In: *Briefings in bioinformatics* 10 (Apr. 2009), pp. 193–204. DOI: 10.1093/bib/bbp004.

[81] T Safavi et al. "Personalized Knowledge Graph Summarization: From the Cloud to Your Pocket". In: Beijing, China, Nov. 2019, pp. 528–537.

[82] Satu Elisa Schaeffer. "Graph clustering". In: *Computer Science Review* 1.1 (2007), pp. 27–64. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2007.05.001.

[83] Ansgar Scherp et al. "Structural Summarization of Semantic Graphs Using Quotients". In: *Transactions on Graph Data and Knowledge* 1.1 (2023), 12:1–12:25. ISSN: 2942-7517. DOI: 10.4230/TGDK.1.1.12. URL: https://drops-dev.dagstuhl.de/entities/document/10.4230/TGDK.1.1.12.

[84] Guus Schreiber and Yves Raimond. *RDF 1.1 Primer*. 2014. URL: https://www.w3.org/TR/rdf11-primer/.

[85] Boon-Siew Seah et al. "FUSE: a profit maximization approach for functional summarization of biological networks". In: *BMC Bioinformatics* 13.3 (2012), S(10). DOI: 10.1186/1471-2105-13-S3-S10.

[86] Ahmed Seffah et al. "Usability measurement and metrics: A consolidated model". In: *Software Quality Journal* 14 (June 2006), pp. 159–178. DOI: 10.1007/s11219-006-7600-8.

[87] Stephan Seufert et al. "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.

[88] L. Shi et al. "Hierarchical Focus+Context Heterogeneous Network Visualization". In: *2014 IEEE Pacific Visualization Symposium*. Yokohama, Japan, 2014, pp. 89–96. DOI: 10.1109/PacificVis.2014.44.

[89] 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, Sept. 1996, pp. 336–343. DOI: 10.1109/VL.1996.545307.

[90] Amit Singhal. *Introducing the Knowledge Graph: things, not strings*. 2012. URL: https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.

[91] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge". In: *Proceedings of the 16th international conference on World Wide Web*. Singapore , Singapore, 2007, pp. 697–706.

[92] Mayesha Tasnim et al. "Chapter 8 Context-Based Entity Matching for Big Data". In: *Knowledge Graphs and Big Data Processing*. Ed. by Valentina Janev et al. Cham: Springer International Publishing, 2020, pp. 122–146.

[93] Robert L. Thorndike. "Who belongs in the family?" In: *Psychometrika* 18.4 (1953), pp. 267–276. DOI: 10.1007/BF02289263.

[94] Peter D Turney and Patrick Pantel. "From frequency to meaning: Vector space models of semantics". In: *Journal of artificial intelligence research* 37 (2010), pp. 141–188.

[95] Zhibiao Wu and Martha Palmer. "Verb semantics and lexical selection". In: *Proc. 32nd Annu. Meeting Assoc. Comput. Linguistics*. Las Cruces, New Mexico, USA, 1994, pp. 133–138.

[96] Dongkuan Xu and Yingjie Tian. "A Comprehensive Survey of Clustering Algorithms". In: *Annals of Data Science* 2.2 (2015), pp. 165–193. DOI: 10.1007/s40745-015-0040-1.

[97] Bo Yan. "Geographic Knowledge Graph Summarization". Available at https://escholarship.org/uc/item/04h696z4. PhD thesis. Santa Barbara, CA: UC Santa Barbara, June 2019.

[98] Shuo Yang et al. "Efficiently Answering Technical Questions — A Knowledge Graph Approach". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (Feb. 2017). DOI: 10.1609/aaai.v31i1.10956. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10956.

[99] Zeqian Shen, Kwan-Liu Ma, and T. Eliassi-Rad. "Visual Analysis of Large Heterogeneous Social Networks by Semantic and Structural Abstraction". In: *IEEE Transactions on Visualization and Computer Graphics* 12.6 (2006), pp. 1427–1439. DOI: 10.1109/TVCG.2006.107.

[100] Kaicheng Zhang et al. "A Knowledge Graph Completion Method for Telecom Metadata Based on the Spherical Coordinate System". In: *IEEE Access* 10 (2022), pp. 122670–122678. DOI: 10.1109/ACCESS.2022.3223432.

[101] Yang Zhao et al. "Knowledge Graph Enhanced Neural Machine Translation via Multi-task Learning on Sub-entity Granularity". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4495–4505. DOI: 10.18653/v1/2020.coling-main.397. URL: https://aclanthology.org/2020.coling-main.397.

[102] Z. Zhou et al. “Semantic-Aware Visual Abstraction of Large-Scale Social Media Data With Geo-Tags”. In: *IEEE Access* 7 (2019), pp. 114851–114861. DOI: 10.1109/ACCESS.2019.2935471.

[103] Zili Zhou, Yanna Wang, and Junzhong 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.

[104] Ganggao Zhu and Carlos A. Iglesias. “Computing Semantic Similarity of Concepts in Knowledge Graphs”. In: *IEEE Transactions on Knowledge and Data Engineering* 29.1 (2017), pp. 72–85. DOI: 10.1109/TKDE.2016.2610428.

[105] Ganggao Zhu and Carlos A. Iglesias. “Sematch: Semantic similarity framework for Knowledge Graphs”. In: *Knowledge-Based Systems* 130 (2017), pp. 30–32. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2017.05.021.

[106] Xiaohan Zou. “A Survey on Application of Knowledge Graph”. In: *Journal of Physics: Conference Series* 1487.1 (Mar. 2020), p. 012016. DOI: 10.1088/1742-6596/1487/1/012016. URL: https://dx.doi.org/10.1088/1742-6596/1487/1/012016.

Index of Authors

Subject Index