Knowledge Representation in Cognitive Systems

Thesis

Course:
Cognitive Systems

Course Code:
CICPE08

Prepared by:
Shubhveer Singh Chaudhary (2021UCI8034)
Kushagra Lakhwani (2021UCI8036)
Vansh Kandwal (2021UCI8041)
Sushant Lavania (2021UCI8061)

Supervised by:
Dr.Shobha Bhatt

Computer Science and Engineering
Netaji Subhas University of Technology
August 31, 2024

Contents

Abstract

The rapid advancement of big cognitive AI models and the exponential growth of data necessitate efficient information storage and retrieval mechanisms. Knowledge Graphs have emerged as a pivotal tool for representing and reasoning about intricate relationships between entities within a domain. Their applications are diverse, encompassing social networks, recommendation systems, fraud detection, and network analysis, thereby underscoring their versatility and utility in complex information systems.

List of Figures

Chapter 1

What is Knowledge

1.1 Introduction

The concept of 'Knowledge' has been the topic of extensive research for many centuries for scientists and philosophers. For a Cognitive System, knowledge provides the basic blocks on which the system will function. The Cognitive Systems itself is incapable of taking in new data and without the knowledge of previous actions and works, it can not learn and produce results.

Data, being a major part of knowledge, and it's type will indicate what type of methods and algorithms are needed to build a particular Cognitive System.

- For example, if the system is provided with text-based data, then some key features that the machine learning will have to consider includes-sentiment analysis, PoS Tagging, identification and analysis of non-conventional text forms like emojis, etc. Natural Language Processing (NLP) techniques extract the meaning from a given text by identifying the grammar and it's various rules and using it to uncover the meaning behind the text. Linguistic Analysis helps in breaking down the text in order to get the meaning behind it.

- For Image based data, some key machine learning techniques include Principle Component Analysis (PCA), developed by Kirby and Servich in the late 1980s; Eigenface-Fisher Linear Discriminant (EFLD) and Dynamic Fuzzy Neural Networks (DFNN). The latter two algorithms are extensively used in face recognition, in terms of identifying dimensions of facial features. In recent times, Facebook has come up with it's own face recognition software DeepFace.

- For Speech based data, various stochastic models and Hidden Markov Model(HMM) have been used, especially for solving the problem of Automatic Speech Recognition. Nowadays, Automated Speech Recognition is widely in talks due to it's ability to identify patterns and recognise the context and meaning behind the speech. As technology further progresses, Speech Analysis becomes more and more important.

For a Cognitive System to be able to have 'Cognition' and be a part of the society, it is imperative to build the system so that it can capture and assess knowledge like humans, so that it can make it's own decisions on it's own.

1.2 The Importance of Knowledge in Cognitve Systems

History of Knowledge has become a discipline of itself, getting it's presence acknowledged since the 2000s, with the advent of the internet and later Big Data. The already existing knowledge and it's applications have been helpful for scientists to build further theories and also to rectify or challenge already existing models. This is the way our civilisation has been able to produce all the scientific and technological marvels.

The Cognitive System is expected to function in a similar fashion. The Cognitive System, depending on whether it utilises Supervised or Unsupervised Learning, will require external sources to able to perform it's tasks, whether it be predicting future outcomes, being able to provide solutions to everyday problems.

Taking the example of current AI tools like Gemini and ChatGPT, which utilise the enormous collection of data present on the internet, alongside it's efficient learning algorithms which not only allows them to provide answers to questions and hold on a full-fledged conversations, but also allows them to rectify the errors made by them in future.

All this requires an extensive and sophisticated collection and management of existing data. This is where Knowledge Representation comes in. The various methods involved in Knowledge Representation are essential in providing the Cognitive System a clear picture of the current situation and also to learn further.

Knowledge in Cognitive Systems is what allows it function efficiently. Without knowledge from external sources, most of the systems would become irrelevant and obsolete. Thus providing the Cognitive System with updated and precise data is of great importance.

In the Healthcare Sector,for example, new treatments and medicines are being constantly developed. There are several tools which are being made which provide more accurate representation of a patient's issues. Cognitive Systems are being developed which can interact with patients and thus help the experts in diagnosis. All this cannot be possible if constant updation and addition of data and knowledge is not present.

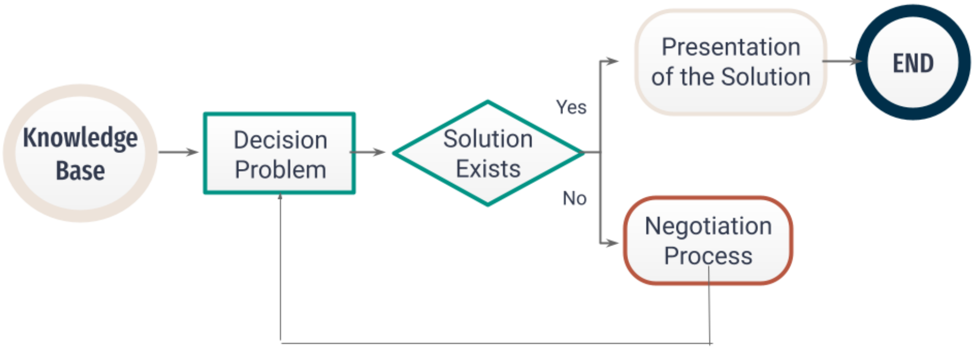

Figure 1.1: Role of Knowledge in decision making

Chapter 2

Knowledge in Cognitive Systems

2.1 The Concept of Learning

For a Cognitive System, the main aspect of knowledge gathering and it's management boils down to it's learning. Learning is the most important aspect of a Cognitive System. A Cognitive System which cannot learn from it's own actions and the data which is constantly updated to it can not be helpful to us. Just like how humans learn from their actions and develop changes within their thought process according to the situation present in front of them, the Cognitive System must be able to deal with the challenges and updations regularly.

2.2 Corpus Building

Corpus refers to the **machine readable** format of the data provided. It includes all the various types of data necessary for a Cognitive System for working in a format which is easily identified by the system.

Corpus is extensively used by many experts of various fields as it helps in Linguistic Analysis and various other research oriented tasks. Take for example a medical professional(the analogies and examples of correlation between Cognitive Systems and Healthcare will be encountered a lot in this chapter), who wants to recommend medicines for his/her patients. If the professional has access to the corpus, then it becomes easier to recommend new and effective medicines and treatment to patients since the corpus provides access to the various research advancements and discoveries to the experts.

Corpus also needs to be maintained in a proper fashion and has to be made in such a way that it remains useful to experts after every iteration of updation. If the new data added contains a lot of vague and unnecessary information, then it makes the corpus less useful and overtime it will loose it's credibility and will become obscure. On the other hand, if the data added is limiting it's full potential of precise data, then it can not be used for Hypothesis Scoring and will loose precious insights. The Machine Leaning of the Cognitive Systems will depend upon the quality of the corpus provided.

A Cognitive System uses several layers of Analytical and Extraction Services which act upon the data provided to the Cognitive System. These layers act as an intermediary for the outside world and the working models of the Cognitive System. They provide tools for representing data into ways which highlight the properties.

Electronic Medical Records(EMR) are a good example, cause they provide the experts with valuable information that may not be available to them through sheer memory and personal experiences only. They also provide an apt representation of the various methods and tools accessible to the experts.

Corpus Building also leads to various security issues as well. These include:-

- The way data is added to the Cognitive Systems is usually through the ETL method, i.e, **E**xtract-**T**ransform-**L**oad method, which can lead to management and security risks.

- Developers need to be certain about the compliance of data and metadata present in the corpus.

- The availability of tools does not relieve the developers from verification of the data coming from various sources and to check it's authenticity.

2.3 Introduction of Big Data in Cognitive Systems

With the advent of the internet and later all the new technological platforms, the amount of data keeps on increasing. This data is very beneficial for the

Cognitive Systems, since it keeps it updated to the new trends and helps in increasing it's efficiency, but the main issue arises with it's management. This enormous data cannot be handled with the traditional methods of handling conventional data.

This calls for devising new methods and techniques for handling such data and this type of data is called **Big Data**.

Big Data already is and also will be an important aspect of Cognitive Computing since we already live in an age where loads of data is being produced everyday, either through social media apps, IoT and smart devices, legacy data, storage units for Banking and Healthcare Sector, etc.

Big Data can be classified as :

- Structured Data
- Unstructured Data
- Semi-Structured Data

These types of data constitute most of the data we deal with. For further explaining Structured Data, here are it's few types:-

- **Key-Value Pair**: Includes data that has a pointer(Key) and it's assigned dataset(Value). eg - XML Based documents and EDI Systems
- **Columnar Database**: Data which is stored in columns instead of rows, which makes it easier for reading and writing data on the disk. eg - HBase and Google's BigTable
- **Graph Database**: Data which is sotred in a single structure,i.e, Graph,which includes nodes and edges.eg - Neo4j

There are also several Data Services Tools which are helpful for the developers in maintaining Cognitive Systems and it's Corpus. These include :-

- ETL (Extract-Transform-Load) that is mainly used for Hadoop and it's features, mainly for managing structured and unstructured data.
- Distributed File System, mainly used for managing data coming from varied sources
- Coordination Services, used for taking leverage on distributed data.

Chapter 3

Knowledge Representation

Data is like Lego blocks for the cognitive system, if our cognitive system can learn to organize or put these blocks together into correct places then it is termed as information, if this structure of data (blocks) can be used for future use it is termed as knowledge.

1. **Data**: Raw facts, figures, data that does not mean anything independently, much like individual LEGO blocks.

2. **Information**: When data is organized, structured, or given context, it becomes information. Much like fitting LEGO blocks into a structure. Information gives logic, meaning, and allows us/cognitive systems to understand how to use these blocks. This is like creating a house or a spaceship model using the LEGO.

3. **Knowledge**: When this structure of blocks (Information) can be understood, applied, and can be used for future use, it becomes knowledge. This is similar to having a LEGO creation that can be changed, extended, or reused in different ways.

4. **Knowledge Representation**: Imagine now writing detailed instructions on how to assemble the LEGO structure just like before. This step-by-step instructions will guide us to assemble the structure just as before by providing instructions on how to put the LEGO together to construct that structure. They store the relation between the blocks, the proper sequence for assembly, and any other relevant details required to reconstruct the structure.

3.1 How is Knowledge Represented

Let's consider the example of the healthcare industry, where a huge volume of data is generated and needs to be stored on a daily basis. This data includes digital health records (DHRs), Medical imaging data (MRI, X-Ray, ECG, etc.), lab reports, prescriptions, clinical notes, genetic and hereditary information, and much more.

In this section, we discuss how this data is represented in such a way that it becomes knowledge and can be used in the future.

3.1.1 Structured Data

Data such as DHR, lab reports, medical history is stored in a structured manner, that is, they follow a certain predefined structure. This data can be easily queried and analyzed as it can be stored easily in a RDBMS (Relational Database Management System). This data allows a cognitive system to derive info easily, extract insights to support diagnosis.

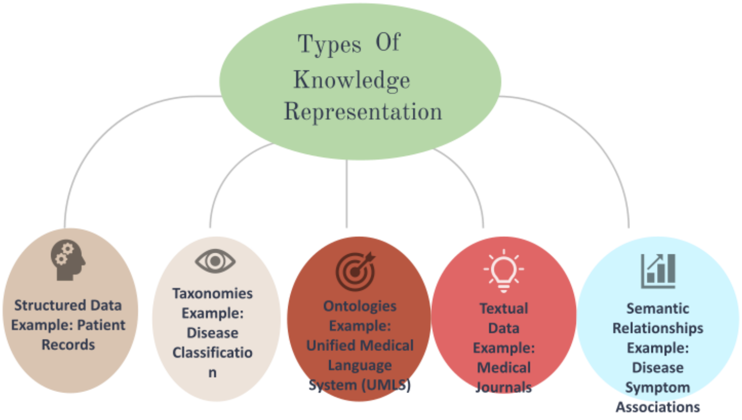
Types Of Knowledge Representation
Structured Data Example: Patient Records
Taxonomies Example: Disease Classificatio n
Ontologies Example: Unified Medical Language System (UMLS)
Textual Data Example: Medical Journals
Semantic Relationships Example: Disease Symptom Associations

Figure 3.1: Types of Knowledge Representation

3.1.2 Taxonomies

Definition: Taxonomy is a hierarchical classification model that is used to categorize entities/objects based on their similarities and differences.

Example

Disease Taxonomies:

Taxonomies organize diseases and medical conditions based on the hierarchy of diseases and affected organ systems. Disease classifications assist in patient condition analysis, support clinical decision, and disease management, enhancing patient care and outcomes.

3.1.3 Ontologies

We need structured mechanisms to derive relationships, concepts, and logic. Ontologies allow us to store and represent concepts and their relationships within a given domain.

Ontologies in Familial History:

When a patient like John visits a healthcare provider and mentions his family's medical history, ontologies play a crucial role in organizing and utilizing this information. For example, John might mention that his father had diabetes and his mother had cardiovascular disease. An ontology designed for familial history would define relationships between family members, genetic predispositions, and medical conditions. By annotating John's family history data according to this ontology, it becomes semantically interoperable. Healthcare providers can then assess John's risk factors for certain medical conditions based on his familial history, guiding preventive measures and treatment decisions. Additionally, aggregating familial history data using standardized ontologies enables population-level analysis and research, leading to a better understanding of genetic patterns and more effective healthcare strategies.

3.1.4 Textual Data

A lot of data such as clinical notes, medical literature, research papers, etc., are in text-based formats which are unstructured and may contain hidden patterns and context. This data requires Natural Language Processing (NLP) for it to be fully utilized and extract relevant info such as patients' medical context, or research findings, or expert opinions, and historical moves/comments about a similar situation.

3.2 Applications of Knowledge Representation

Certainly, here are five applications of knowledge representation with concise explanations:

1. **Robotics**: Robots need to understand their environment; this allows them to perform tasks and collaborate with humans. Knowledge representation techniques allow robots to represent this data efficiently.

2. **Knowledge Management**: Many systems need information from various sources, with different formats. Using ontologies and knowledge graphs allows these systems to use this data efficiently and make decisions based on them.

3. **Recommendation Systems**: Using knowledge representation techniques such as knowledge graphs, a recommendation system uses various attributes for recommending different products, services, or content such as movies and series.

Chapter 4

Knowledge Representation in Cognitive Systems

As discussed in the previous chapter, Knowledge is data that has been given structure and can be used in the future. Knowledge representation is a crucial part of any cognitive system. The knowledge base, where all this knowledge is stored in such a way that it is interlinked, and their relationship between the data is clearly mentioned. Cognitive systems continuously learn with experience and from the environment, hence a knowledge base system for it must support **Continuous Learning**.

4.1 Data Collection and Storage Strategies for Cognitive Systems

Data is the fundamental building block, as discussed in Chapter 3. Cognitive systems collect this data from various sources such as sensors, existing databases, paid/free public APIs, web scraping, and crowdsourcing.

4.1.1 Data Collection

Using the example of a healthcare Cognitive system, let's see how these various sources contribute to the data for the CS:

1. *Sensors*: Metrics like heart-rate, blood pressure, and sleep patterns can be collected using wearable devices, which allow deeper insights into

patients' health. Chronic diseases can be monitored in a much better way using devices such as continuous glucose monitors for diabetes.

2. *Existing Databases*: EHR and public patient databases allow hospitals and medical personnel to access medical history, running medication, and ongoing treatment resulting in faster and more efficient management of patient care.

3. *APIs (Application Programming Interfaces)*: Public APIs allow systems to collect data directly from third-party data providers and healthcare applications and devices.

4. *Web Scraping*: This technique allows systems to collect data from websites such as online medical journals, research papers, and medical publications.

5. *Crowdsourcing*: Platforms use forms and questionnaires to collect user data such as symptoms, doctor ratings/reviews, and experiences such as side effects to create an extensive knowledge base.

After this data is collected, it needs to be cleaned and stored.

4.1.2 Data Pre-processing

This collected data needs to be processed before storing it in a knowledge graph. This may include cleaning missing values, removing outliers, normalization, and feature selection.

4.1.3 Data Storage

Following steps may be taken to store this data:

- **Schema Matching**: A schema must be decided, and data needs to be fixed into it. A schema must be defined for each individual source. This can be done by mapping similar attributes together.

- **Semantic Analysis**: Data contains context and relationships among itself. This needs to be represented in the knowledge base to allow the cognitive system to find the correct entity and its relation with other data. For example, consider a clinical note that mentions a patient.

The cognitive system must be able to understand this real-life entity as the patient it knows from its knowledge base. This would allow it to get all the related data to the patient themselves and better extract insights from the note.

4.2 Knowledge Representation in Cognitive System Databases

After data collection and preprocessing, it needs to be represented in the cognitive system databases. This can be done in the following ways:

1. **Semantic Representation**: Traditional databases like RDBMS cannot represent complex relationships and capture the context of the data. Hence, we need a semantic representation of the data using techniques such as Knowledge Graphs. For example, the relationship between disease, symptoms, and treatment may be stored in ontologies. This allows the cognitive systems to draw diseases from symptoms, find treatments allowing better care for the patients.

2. **Custom Knowledge Schemas**: Custom schemas create a better way to organize knowledge to better utilize this knowledge. Taxonomies can be used to store familial data allowing to store and see what disease someone may be predisposed to because of their parents and grandparents.

3. **Multi-format Data**: Various formats may be needed to store such as textual reports, Medical images such as X-Rays MRIs etc. Patient's medical reports, X-rays, and other reports need to be stored along with their diagnosis; this is done using vector embedding, deep learning Techniques.

4.3 Enabling Continuous Learning in Cognitive Systems

Healthcare is an ever-growing industry; every day new research, diseases, treatments, and medications are found, hence we need to create a cognitive system that can adapt to such an environment. Hence, we need to deploy adaptive systems.

- **Continuous Learning Algorithm**: Neural networks and other deep learning techniques are employed in cognitive systems, but these should be incrementally learning. That is, they should be able to update the knowledge base without requiring the model to retrain algorithms as an industry like healthcare is very dynamic. For example, a CS monitoring blood sugar should be able to learn new trends in blood sugar of patients to maintain accuracy hence the new data must be stored and learned incrementally.

- **Dynamic Memory Structure**: Traditional storage and memory architecture may become overwhelmed with the ever-growing data that is generated in real-time. Hence, we need a memory system that can handle and store the data based on its relevance, recency, and contextual importance. We may use priority-based storage systems witch store the recent data like recent patient history or symptoms based on the current case in the faster memory like cache. For example, patients that are currently admitted to the hospital, their data would be given more priority in the memory.

4.4 Security Concerns

Knowledge base and cognitive system databases store private data of users (Medical data), important research papers, and other related data. To maintain the privacy of users and prevent malicious entities from accessing the data, security measures such as encryption, access control, and view management.

To maintain security and privacy of data and users can be done by following methods:

- **Encryption**: Methods like end-to-end encryption can be used to prevent accessing of data without authorization.

- **Minimal Data Collection**: The minimum amount of data, that is only necessary data for the functionality of the cognitive system should be collected from users.

- **Adding Noise while Storing Data**: Is done so that the data can still be used to train models but not make out any personal data.

- **Anonymization of Data**: Personal data of users must be protected to ensure that any data that is vulnerable should not contain users' personal details. This can also be done by differential privacy methods which use mathematical frameworks to ensure information about a specific person/user is not revealed.

- **Proper Access Control Mechanisms**: Role-based access control should be employed, i.e., people with certain roles/posts can only access certain data. For example, a doctor can access the data of a patient, but only when the patient is in the appointment with the doctor.

- **Auditing and Threat Detection**: Regular audits and security checks should be done to ensure that there is no suspicious activity happening. Threat detection should be done actively to detect any intrusions or attacks.

Chapter 5

Case Study - Knowledge Graphs

A Knowledge Graph (KG), represents a collection of interlinked descriptions of entities - real-world objects, events, situations or abstract concepts. The entities are represented as nodes and the relationships between them as edges. The nodes and edges are labeled with attributes that describe the entities and relationships.

Just like any Graph, the KG is made up of nodes and edges, where the edges have a direction, indicating the direction of the relationship between a pair of concepts.

A KG is a structured representation of knowledge, which can be used understood as a 'Graph of Concepts', to infer new knowledge, answer questions, and discover new relationships between entities. It is a powerful tool for representing and reasoning about complex relationships between entities in a domain.

5.1 Knowledge Graphs in the Real World

Knowledge Graphs are used in a variety of applications. Any text corpus can be visualised in an interactive and intuitive manner.

Graph Databases like Neo4j[2], Amazon Neptune[3], and JanusGraph[4] are a way to store and query data in the form of a graph. They are used in a variety of applications, including social networks, recommendation systems, fraud detection, and network analysis.

5.2 Creating Knowledge Graphs using LLMs

Large Language Models (LLMs) like GPT-3.5[8] and BERT[9] can be used to create Knowledge Graphs from unstructured text. Through the use of fine-tuning and prompt engineering on a specific domain or dataset these models can be used to generate a KG from the text.

5.2.1 Our Approach

We use the Google's generative AI suite to access the 'Gemini-1.0-pro' model to generate a a structured representation of Knowledge Graph from a text corpus through a user given prompt or a an already existing text corpus.

We use create tools to implement generation of KGs through function calling and provide a user-friendly interface to interact with the model.

Examples

Consider the following verse taken from the Hollow Knight game:

No cost too great.
No mind to think.
No will to break.
No voice to cry suffering,

Born of God and Void.
You shall seal the blinding light that plagues their dreams.
You are the Vessel.
You are the Hollow Knight.

– *The Pale King*

This entire text can be converted into a Knowledge Graph, which can be used to infer new knowledge, and discover new relationships between entities.

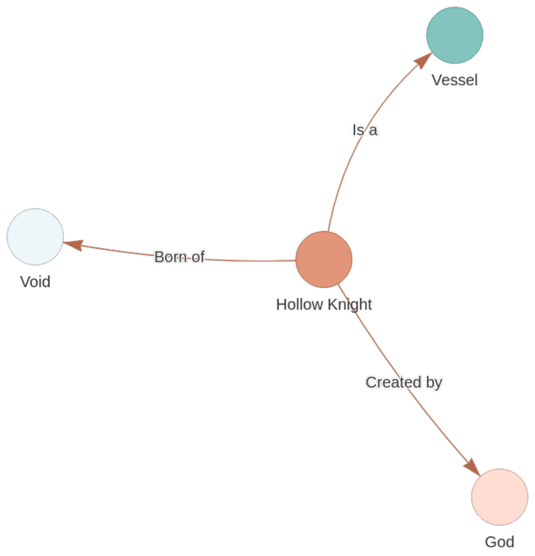

Figure 5.1: Corpus based Knowledge Graph

Our approach[5] can also create the following Knowledge Graph based on a user prompt:

```
User Prompt: "How does the Transformer model work?"
```

Which generates the following Knowledge Graph:

```
Knowledge Graph:
{
    "name": "add_to_database",
    "args": {
        "entities": [
            {
                "description": "Neural network model for natural
                language processing tasks",
                "type": "Model",
                "name": "Transformer Model Architecture"
            },
            ...
        ],
        "relationships": [
            {
                "to_entity_name": "Encoder",
                "relationship": "Composed of",
                "from_entity_name": "Transformer Model Architecture"
            },
            ...
        ]
    }
}
```

The generated Knowledge Graph represents the relationships between the entities in the text, and can be used to answer questions, infer new knowledge, and discover new relationships between entities.

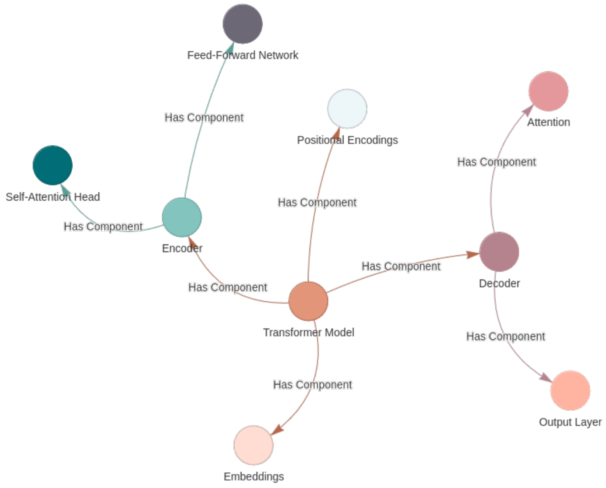

Figure 5.2: Prompted Knowledge Graph: Transformer

5.3 Applications

Knowledge Graphs have been used in a slew of different industries and applications. These are used by researchers, marketing professionals, and data scientists to understand the relationships between entities in a domain, and to infer new knowledge from the data.

Recommendation systems and search engines use Knowledge Graphs to provide more relevant and accurate results to users. They can be used to answer questions, infer new knowledge, and discover new relationships between entities in a domain.

Inspite of these large scale applications, KGs can be used in a more personal and individual level. They can be used to create a personal knowledge base, to store and organize information, and to help you remember and recall information more easily. They allow us to understand concepts at a glance and can be used to create a visual representation of a text corpus.

6

Conclusion and emerging trends

As we have shown, knowledge representation is a crucial aspect of cognitive systems. It allows us to store, organize, and use data efficiently. Knowledge representation techniques such as ontologies, taxonomies, and knowledge graphs are used in various industries and applications. They help us understand the relationships between entities in a domain, infer new knowledge from the data, and discover new relationships between entities. Knowledge representation is a rapidly evolving field, and we can expect to see more advancements in the future.

References

[1] J. T. Shreeja Das, Santanu Mahapatra and D. Saha, “Machine learning assisted search of figure of merit,” *Workshop on Spintronics and Magnetism*, 23rd to 27th August 2021.

[2] Neo4j, “Neo4j,” https://neo4j.com/, 2021, accessed: 2024-04-05.

[3] A. W. Services, “Amazon neptune,” https://aws.amazon.com/neptune/, 2021, accessed: 2024-04-06.

[4] JanusGraph, “Janusgraph,” https://janusgraph.org/, 2021, accessed: 2024-04-06.

[5] K. Lakhwani, “Knowledge graphs through GENAI,” https://github.com/KorigamiK/knowledge-graphs, 2024.

[6] J. Evermann, “Towards a cognitive foundation for knowledge representation,” 2005.

[7] A. B. Judith S. hurwitz, Marcia Kaufman, “Cognitive computing and big data.”

[8] OpenAI, “Gpt-3,” https://www.openai.com/gpt-3/, 2021, accessed: 2024-04-06.

[9] Google, “Bert,” https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html, 2018, accessed: 2024-04-06.