PEOPLE'S DEMOCRATIC REPUBLIC OF ALGER

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY MOHAMED BOUDIAF - M'SILA

Faculty of mathematics and informatics

DEPARTMENT OF COMPUTER SCIENCE

A Dissertation in Fulfillment

For the Requirements of the Degree of Master

In Information Systems and Software Engineering

By :

BOURAS AMINA BOCHRA

BELHADJAMI MESSAOUD

SUBJECT

PREDICTION OF MISSING VALUES USING KNOWLEDGE GRAPH

Supervised by:

Dr. MEHENNI TAHAR University of M'sila

Academic year: 2019-2020

بسم الله الرحمن الرحيم

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGER

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY MOHAMED BOUDIAF - M'SILA

Faculty of mathematics and informatics

DEPARTMENT OF COMPUTER SCIENCE

A Dissertation in Fulfillment

For the Requirements of the Degree of Master

In Information Systems and Software Engineering

By :

BOURAS AMINA BOCHRA

BELHADJAMI MESSAOUD

PREDICTION OF MISSING VALUES USING KNOWLEDGE GRAPH

Supervised by:

Dr. MEHENNI TAHAR University of M'sila

Academic year: 2019-2020

Dedication

To our dear parents and all closes persons.

To my sisters Bouras: K, H, Kh *for their support* and

my best friends.

Acknowledgment

We thank god for helping us to complete this modest achievement.

*For me **Bouras Amina**, I would like to thank at first my dear sister B.Khouloud for all beautiful things and specially her help and my mother. Also, thanks to my aunt Bouras Rachida for supporting me. For sure, thanks to my partner Belhadjami Messaoud for all things and B.Khaled to help me.*

*For me, **Belhadjami Messaoud**, I would like to thank my family and all the people who have helped me.*

*Without forgetting our supervisor **Dr. Mehenni Tahar** great thanks to him.*

TABLE OF CONTENTS

List of figures

List of tables

LIST OF ACRONYMS

ACID	Atomicity, Consistency, Isolation, Durability
AI	Artificial Intelligence
ARIMA	Autoregressive Integrated Moving Average
DAD	Développement d'Applications Distribuées (in French)
DB	Data base
DMRI	Data Mining et Recherche d'Information (in French)
EM	Expectation-Maximization
HTML	HyperText Mark-up Language
IBM	International Business Machines
KG	Knowledge graph
KGs	Knowledge graphs
k-NN	k-Nearest Neighbors
KS	Knowledge Schema
LOCF	Last Observation Carried Forward
MAE	Mean Absolute Error
MAR	Missing at Random
MCAR	Missing Completely at Random
NELL	Never Ending Language Learning
NMAR	Not Missing at Random
NoSQL	Not only SQL
OWL	Web Ontology Language.
POC	Programmation Orientée Composants (in French)
RDF	Resource Description Framework
RMSE	Root Mean Squared Error
SKOS	Simple Knowledge Organization System
SPARQL	Acronym for SPARQL Protocol and RDF Query Language
TestLog	Test du Logiciel (in French)
URIs	Uniform Resource Identifiers
YAGO	Yet Another Great Ontology

INTRODUCTION

When we go back a little bit, exactly in 2012 when Google introduce the knowledge graph in search motor. The reason why Google introduce it is to improve semantic search and offers effective search results. This concept was used as one of the most important instruments to represent information in data science. Then it was used by many companies in different forms from the smallest to the biggest according to this companies.

Knowledge graph represents the real word object as entities and its relations as edges. Any form of data (Text, Database, …) can be represented as knowledge graph.

In the other side, the data needs analyzing and processing before saving it. Sometimes database may contain missing values detected in analyzing phase. When we find missing values, it affects the processing phase, so it won't be completed with the missing values. We will be obliged to detect it based on data mining domain which has many prediction techniques like Mean, Linear Interpolation, K- nearest neighbors…etc.

In the same context of data mining, we will implement knowledge graph as technique to predict missing values in database relational that will be represented as graph model (entities are nodes which are connected by relations). In this case, missing values will be represented by special symbol or empty node then we will use link prediction techniques to complete missing information.

Our thesis contains four chapters structured as follows.

The first chapter explains knowledge graph concept and its representations with some examples. As well, it contains overview about companies which used knowledge graphs, history of this concept and its domains.

The second chapter presents data mining and specifically missing values. There are many prediction techniques used to expect these values. We explain, in this part, types of missing values and mention some prediction techniques to handling missing values.

The third chapter introduces our contribution which applies knowledge graph as a technique to predict missing values. Where we used links prediction techniques of KG to find the missing links, then we expected missing values by these links.

Chapter four contains different evaluations of the technique using accuracy rate of prediction model and the discussion of the results.

CHAPTER 1 KNOWLEDGE GRAPH

Chapter 1 Knowledge graph

1.1 Introduction

In the recent years, the knowledge graph term has appeared. It was used in the web. In this chapter, we identify the concept and the representation of knowledge graph. Knowledge graph generally is tied with web when google introduces it in search, we searched as much information as possible about knowledge graph and we found how and where it was used. Then we mentioned KG history.

1.2 Definition

Def 1: **"**A Knowledge graph (i) mainly describes real world entities and their interrelations, organized in a graph,(ii) defines possible classes and relations of entities in a schema,(iii) allows for potentially interrelating arbitrary entities with each other,(iv) covers various topical domains.". [31]

Def 2: **"**Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities." [51]

Def 3: "Knowledge graphs could be envisaged as a network of all kind things which are relevant to a specific domain or to an organization. They are not limited to abstract concepts and relations but can also contain instances of things like documents and datasets." [51]

1.3 Description of Knowledge graph

KGs come in different shapes and sizes; emerging from both companies and open source communities; human curated and automatically generated; with fixed ontology or continuously expanded. Regardless of their differences, most KGs will follow a simple principle: organizing information in a structured way by explicitly describing the relations among entities.[8]

Knowledge graph represented as entities, edges and attributes.

Entity**:** Represents something in the real world.
Edge: Represents relationship.
Attribute: Represents something about an entity [23]

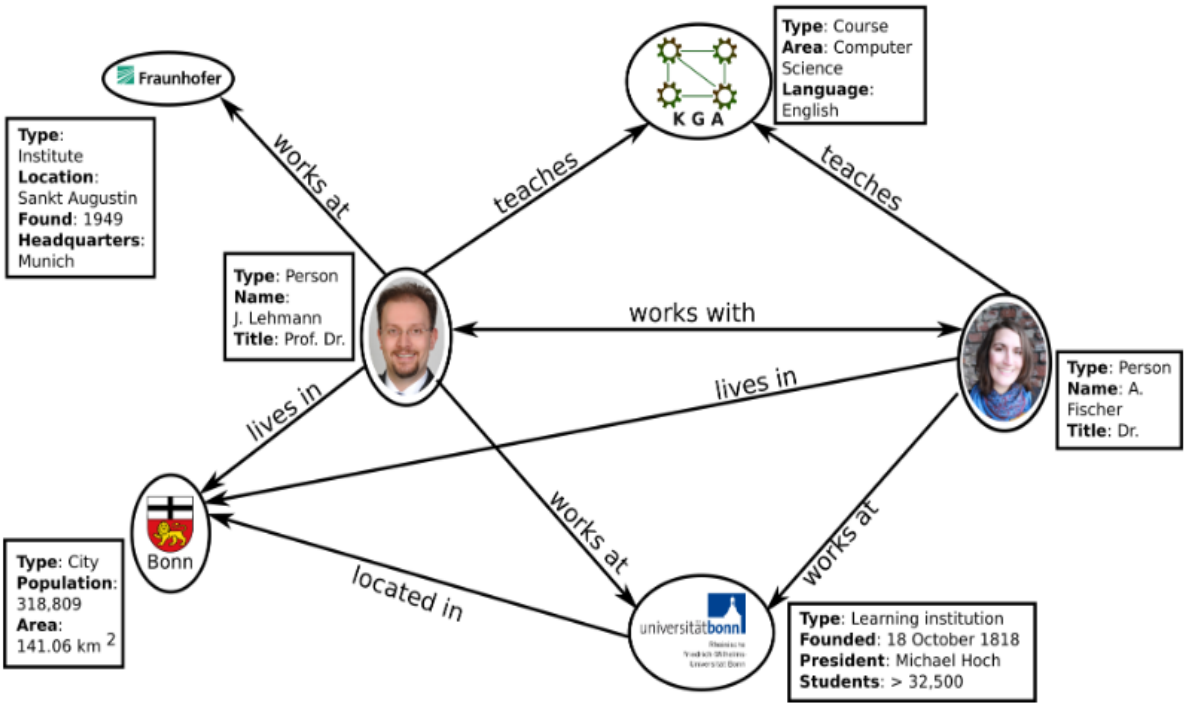

Figure 1.1 Example of a Knowledge graph [47]

1.4 Knowledge graphs generally

The phenomena of a Knowledge Graph first became known worldwide when, in 2012, Google [63] started to use such a Graph in their search engine, allowing users to search for things, people or places [62] Inspired by Google, Knowledge Graphs are being developed by the world's leading information companies. For instance, DBpedia [3] an open knowledge graph, available to anyone on the Web. According to the statistics [61] the lasts DBpedia release 2016 consists of 13 billion pieces of information (RDF triples) out of which 1.7 billion were extracted from the English edition of Wikipedia, 6.6 billion were extracted from other language editions and 4.8 billion – from Wikipedia Commons [65] and Wikidata [64]. [27]

Knowledge Graph is a model of information entities inter-relation. It is a database which stores knowledge in accordance with the particular Knowledge Schema (KS). Such a Knowledge Schema, in turn:

1) Constructs the meta-layer of a KG and provides its internal structure;

2) Defines classes as abstract containers for similar types of entities;

3) Contains a set of potentially available describing elements and potential relationships between classes;

4) Serves as a reference point for integrating new data or constructing new queries;

5) Contains only structural information;

6) Does not contain data about real units of the chosen domain of knowledge;

7) Can be considered as a visual representation of a Knowledge Graph.[10]

1.5 Knowledge graphs in the web

Knowledge graph applications on the web are often viewed as attempts to get closer to the vision of the Semantic Web.[28]

Understandably, knowledge graphs on the web are mostly viewed as a tool to transform web from a collection of websites and links to a full-scale knowledge base, where searching the web is equal to reading a book where the information flow is predefined by someone else in order for a reader to gain the best possible understanding of the subject and where an answer to the initial question already includes answers to questions that might follow.[24]

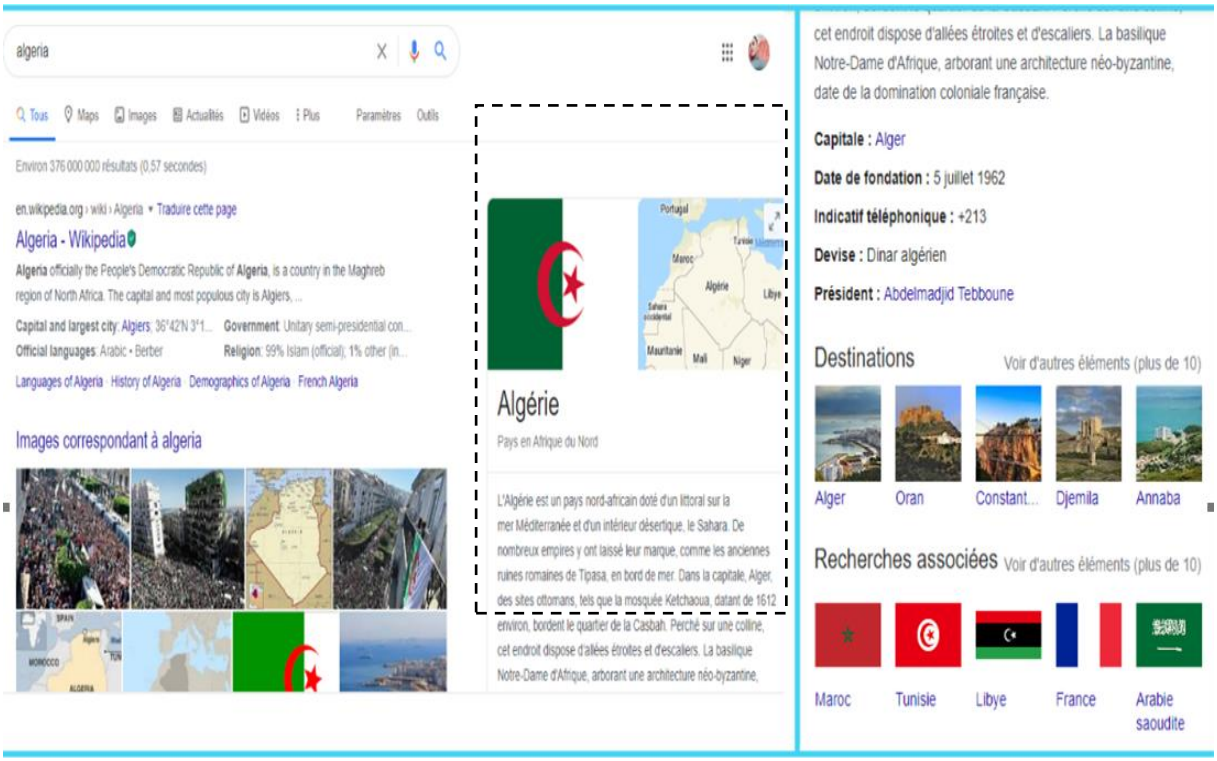

Figure 1.2 Example of search in the web

In accordance with the RDF standard, information is represented in so-called “triples”, or “triplets” (subject-predicate-object), where the two entities (subject and object) are related to

each other by a predicate. Each triple indicates a particular fact, showing interrelations of two selected entities.[27]

To build a Knowledge Schema (and then a Knowledge Graph), the triples should be combined together into an actual multi-graph which will have entities as nodes, relations as edges and predicates as edge labels for each particular relation.[27]

1.6 Semantic web knowledge graph

"The Semantic Web is a Web of Data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. The collection of Semantic Web technologies (RDF, OWL, SKOS, SPARQL, etc.) provides an environment where application can query that data, draw inferences using vocabularies, etc." [10]

However, to make the Web of Data a reality, it is important to have the huge amount of data on the Web available in a standard format, reachable and manageable by Semantic Web tools. Furthermore, not only does the Semantic Web need access to data, but relationships among data should be made available, too, to create a Web of Data (as opposed to a sheer collection of datasets). This collection of interrelated datasets on the Web can also be referred to as Linked Data.[10]

From the early days, the Semantic Web has promoted a graph-based representation of knowledge, e.g., by pushing the RDF standard. In such a graph-based knowledge representation, entities, which are the nodes of the graph, are connected by relations, which are the edges of the graph (e.g., Shakespeare has written Hamlet), and entities can have types, denoted by the relation *is* (e.g., Shakespeare is a writer, Hamlet is a play). In many cases, the sets of possible types and relations are organized in a schema or ontology, which defines their interrelations and restrictions of their usage.[31]

With the advent of Linked Data, it was proposed to interlink different datasets in the Semantic Web. By means of interlinking, the collection of datasets could be understood as one large, global knowledge graph (although very heterogeneous in nature). To date, roughly 1,000 datasets are interlinked in the Linked Open Data cloud, with the majority of links connecting identical entities in two datasets [31].

There are different technologies to build and operate a knowledge graph. In DataFabric we employ Semantic Web standards and technologies, so in this case a dataset following these

characteristics and using Semantic Web standards is called a Semantic Web Knowledge Graph**.** [56]

The foundations of these standards are:

- Usage of URIs for referring to entities, i.e. URIs that point to resources on the Web,
- Usage of RDF for representing the graph,
- Usage of RDF Schema and/or OWL for representing the schema of the graph. [56]

Knowledge graphs on the Semantic Web are typically provided using Linked Data as a standard. They can be built using different methods: they can be curated by an organization or a small, closed group of people, crowd-sourced by a large, open group of individuals, or created with heuristic, automatic or semi-automatic means. In the following, we give an overview of existing knowledge graphs, both open and company-owned.[31]

1.7 Popular Knowledge Graphs

There are many different types of knowledge graphs developed by different companies that are used for different purposes. While many companies use an internal or smaller knowledge graph for online functions, some of the biggest ones are being used by many people all over the world. Below lists a selection of some of the largest knowledge graphs to date from Microsoft, Google, Facebook, IBM and eBay (Fig 1.3). [33]

Developer	Purpose & Function	Stage of Development
Microsoft	Uses knowledge graph for the Bing search engine, LinkedIn data & Academics.	Actively used in products
Google	Knowledge graph is used as a massive categorization function across Google's devices and directly imbedded in the search engine.	Actively used in products
Facebook	Develops connections between people, events and ideas, mainly focusing on news, people and events related to the social network.	Actively used in products
IBM	Provides a framework for other companies and/or industries to develop internal knowledge graphs.	Actively used by clients
eBay	Currently developing a knowledge graph that functions to provide connections between users and products provided on the website.	Early Stage of Development

Figure 1.3 Popular knowledge Graphs, its purpose and function [33]

1.7.1 **Freebase**

Curating a universal knowledge graph is an endeavor which is infeasible for most individuals and organizations. To date, more than 900 person years have been invested in the creation of Cyc, with gaps still existing. Thus, distributing that effort on as many shoulders as possible through crowdsourcing is a way taken by Freebase, a public, editable knowledge graph with schema templates for most kinds of possible entities (i.e., persons, cities, movies, etc.). [31]

If you have a Wikipedia page, then your Freebase entry will be created by itself. You just need to populate the page with more information, if necessary. With Freebase listing, your brand and business niche will be easily understood by Google. This way, your chances of showing up in the Knowledge Graph results will increase. [46]

After MetaWeb, the company running Freebase, was acquired by Google, Freebase was shut down on March 31st, 2015. The last version of Freebase contains roughly 50 million entities

and 3 billion facts9. Freebase's schema comprises roughly 27,000 entity types and 38,000 relation types.[31]

1.7.2 Wikidata

Like Freebase, Wikidata is a collaboratively edited knowledge graph, operated by the Wikimedia foundation that also hosts the various language editions of Wikipedia. After the shutdown of Freebase, the data contained in Freebase is subsequently moved to Wikidata. A particularity of Wikidata is that for each axiom, provenance metadata can be included – such as the source and date for the population figure of a city.[31]

1.7.3 DBpedia

DBpedia is a knowledge graph which is extracted from structured data in Wikipedia. The main sources for this extraction are the key-value pairs in the Wikipedia info boxes. In a crowd-sourced process, types of infoboxes are mapped to the DBpedia ontology, and keys used in those infoboxes are mapped to properties in that ontology. Based on those mappings, a knowledge graph can be extracted.[31]

The most recent version of the main DBpedia (i.e., DBpedia 2015-04, extracted from the English Wikipedia based on dumps from February/March 2015) contains 4.8 million entities and 176 million statements about that entities.17 The ontology comprises 735 classes and 2,800 relations.[31]

The DBpedia knowledge base has several advantages knowledge it covers many domains; it represents real community agreement; it automatically evolves as Wikipedia changes, and it is truly multilingual.[3]

The DBpedia knowledge base allows you to ask quite surprising queries against Wikipedia, for instance "Give me all cities in New Jersey with more than 10,000 inhabitants" or "Give me all Italian musicians from the 18th century". Altogether, the use cases of the DBpedia knowledge base are widespread and range from enterprise knowledge management, over Web search to revolutionizing Wikipedia search.[3]

1.7.4 Yago

Like DBpedia, YAGO is also extracted from DBpedia. YAGO builds its classification implicitly from the category system in Wikipedia and the lexical resource WordNet, with infobox properties manually mapped to a fixed set of attributes. While DBpedia creates

different interlinked knowledge graphs for each language edition of Wikipedia, YAGO aims at an automatic fusion of knowledge extracted from various Wikipedia language editions, using different heuristics.[31]

The latest release of YAGO, i.e., YAGO3, contains 4.6 million entities and 26 million facts about those types. The schema comprises roughly 488,000 types and 77 relations.[31]

1.7.5 NELL

While DBpedia and YAGO use semi-structured content as a base, methods for extracting knowledge graphs from unstructured data have been proposed as well. One of the earliest approaches working at web-scale was the Never Ending Language Learning project. The project works on a largescale corpus of web sites and exploits a coupled process which learns text patterns corresponding type and relation assertions, as well as applies them to extract new entities and relations. Reasoning is applied for consistency checking and removing inconsistent axioms. The system is still running today, continuously extending its knowledge base. While not published using Semantic Web standards, it has been shown that the data in NELL can be transformed to RDF and provided as Linked Open Data as well.[31]

In its most recent version (i.e., the 945th iteration), NELL contains roughly 2 million entities and 433,000 relations between those. The NELL ontology defines 285 classes and 425 relations.[31]

1.7.6 Google's Knowledge Graph

Google's Knowledge Graph was introduced to the public in 2012, which was also when the term knowledge graph as such was coined. Google is rather secretive about how their Knowledge Graph is constructed. There are only a few external sources that discuss some of the mechanisms of information flow into the Knowledge Graph based on experience.[31]

The introduction of Google Knowledge Graph has greatly enhanced the search experience by showing a quick summary of the subject queried, important facts, images, and links. In addition, it provides related searches that allow users to explore additional information on what they were looking for. Rather than one specific result, you are now provided with a collection of answers. In fact, the carousal at the top of search results is one of the best parts to the new algorithm because it makes the search process much faster, quicker, and detailed.[46]

In the Google Knowledge Graph, entities, such as people, places, things, concepts, etc., may be stored as nodes and the edges between those nodes may indicate the relationship between the nodes. [13]

1.7.7 Google's Knowledge Vault

The Knowledge Vault is another project by Google. It extracts knowledge from different sources, such as text documents, HTML tables, and structured annotations on the Web with Microdata or MicroFormats. Extracted facts are combined using both the extractor's confidence values, as well as prior probabilities for the statements, which are computed using the Freebase knowledge graph. From those components, a confidence value for each fact is computed, and only the confident facts are taken into Knowledge Vault. the Knowledge Vault contains roughly 45 million entities and 271 million fact statements, using 1,100 entity types and 4,500 relation types.[31]

1.7.8 Facebook's Entities Graph

Although the majority of the data in the online social network Facebook is perceived as connections between people, Facebook also works on extracting a knowledge graph which contains a larger variety of entities. The information people provide as personal information (e.g., their home town, the schools they went to), as well as their likes (movies, bands, books, etc.), often represent entities, which can be linked both to people as well as among each other. By parsing textual information and linking to Wikipedia, the graph also contains links among entities, e.g., the writer of a book. Although not many public numbers about Facebook's Entities Graph exist, it is said to contain more than 100 billion connections between entities.[31]

1.7.9 Yahoo's Knowledge Graph

Like Google, Yahoo! also has their internal knowledge graph, which is used to improve search results. The knowledge graph builds on both public data (e.g., Wikipedia and Freebase), as well as closed commercial sources for various domains. It uses wrappers for different sources and monitors evolving sources, such as Wikipedia, for constant updates. Yahoo's knowledge graph contains roughly 3.5 million entities and 1.4 billion relations. Its schema, which is aligned with schema.org, comprises 250 types of entities and 800 types of relations. [31]

Name	Instances	Facts	Types	Relations
DBpedia (English)	4,806,150	176,043,129	735	2,813
YAGO	4,595,906	25,946,870	488,469	77
Freebase	49,947,845	3,041,722,635	26,507	37,781
Wikidata	15,602,060	65,993,797	23,157	1,673
NELL	2,006,896	432,845	285	425
OpenCyc	118,499	2,413,894	45,153	18,526
Google's Knowledge Graph	570,000,000	18,000,000,000	1,500	35,000
Google's Knowledge Vault	45,000,000	271,000,000	1,100	4,469
Yahoo! Knowledge Graph	3,443,743	1,391,054,990	250	800

Table 1.1 Number of entities and relations graph [31]

1.8 Approaches for Completion of Knowledge Graphs

Completion of knowledge graphs aims at increasing the coverage of a knowledge graph. Depending on the target information, methods for knowledge graph completion either predict missing entities, missing types for entities, and/or missing relations that hold between entities. [31]

1.8.1 Internal Methods

Internal methods use only the knowledge contained in the knowledge graph itself to predict missing information.[31]

1.8.1.1 Methods for Completing Type Assertions

Predicting a type or class for an entity given some. Characteristics of the entity are a very common problem in machine learning, known as classification. The classification problem is supervised, i.e., it learns a classification model based on labeled training data, typically the set of entities in a knowledge graph (or a subset thereof) which have types attached. In machine learning, binary and multi-class prediction problems are distinguished.

In the context of knowledge graphs, in particular the latter are interesting, since most knowledge graphs contain entities of more than two different types. Depending on the graph at hand, it might be worthwhile distinguishing multi-label classification, which allows for assigning more than one class to an instance (e.g., Arnold Schwarzenegger being both an Actor

and a Politician), and single-label classification, which only assigns one class to an instance. [31]

For internal methods, the features used for classification are usually the relations which connect an entity to other entities, i.e., they are a variant of link-based classification problems. For example, an entity which has a director relation is likely to be a Movie.[31]

Since many knowledge graphs come with a class hierarchy, e.g., defined in a formal ontology, the type prediction problem could also be understood as a hierarchical classification problem. Despite a larger body of work existing on methods for hierarchical classification, there are, to the best of our knowledge, no applications of those methods to knowledge graph completion.[31]

In data mining, association rule mining is a method that analyzes the co-occurrence of items in item sets and derives association rules from those co-occurrences. For predicting missing information in knowledge graphs, those methods can be exploited, e.g., in the presence of redundant information. For example, in DBpedia, different type systems (i.e., the DBpedia ontology and YAGO, among others) are used in parallel, which are populated with different methods (Wikipedia infoboxes and categories, respectively). This ensures both enough overlap to learn suitable association rules, as well as a number of entities that only have a type in one of the systems, to which the rules can be applied. In we exploit such association rules to predict missing types in DBpedia based on such redundancies.[31]

1.8.1.2 Methods for Predicting Relations

While primarily used for adding missing type assertions, classification methods can also be used to predict the existence of relations. To that end, Socher et al. propose to train a tensor neural network to predict relations based on chains of other relations, e.g., if a person is born in a city in Germany, then the approach can predict (with a high probability) that the nationality of that person is German. The approach is applied to Freebase and WordNet. where the authors show that refining such a problem with schema knowledge – either defined or induced can significantly improve the performance of link prediction.[31]

1.8.2 External Methods

External methods use sources of knowledge – such as text corpora or other knowledge graphs which are not part of the knowledge graph itself. Those external sources can be linked from the knowledge graph, such as knowledge graph interlinks or links to web pages, e.g., Wikipedia

pages describing an entity, or exist without any relation to the knowledge graph at hand, such as large text corpora.[31]

1.8.2.1 Methods for Completing Type Assertions

For type prediction, there are also classification methods that use external data. In contrast to the internal classification methods described above, external data is used to create a feature representation of an entity.[31]

Nuzzolese et al. Propose the usage of the Wikipedia link graph to predict types in a knowledge graph using a k-nearest neighbors classifier. Given that a knowledge graph contains links to Wikipedia, interlinks between Wikipedia pages are exploited to create feature vectors, e.g., based on the categories of the related pages. Since links between Wikipedia pages are not constrained, there are typically more interlinks between Wikipedia pages than between the corresponding entities in the knowledge graph.[31]

Apriosio et al. use types of entities in different DBpedia language editions (each of which can be understood as a knowledge graph connected to the others) as features for predicting missing types. The authors use a k-NN classifier with different distance measures (i.e., kernel functions), such as the overlap of two articles' categories. In their setting, a combination of different distance measures is reported to provide the best results.[31]

1.8.2.2 Methods for Predicting Relations

Like types, relations to other entities can also be predicted from textual sources, such as Wikipedia pages. Lange et al. learn patterns on Wikipedia abstracts using Conditional Random Fields.[31]

1.9 Knowledge graphs and Artificial Intelligence

For Artificial Intelligence to make its intelligent decisions, it is necessary to emulate 'context'. It must be possible to represent the relationships, complexity and, most importantly, the meaning of data in such a way that the AI's seemingly intelligent conclusions and decision are sensible in the real-world. Knowledge graphs represent data and the meaning of that data. This is how information is transformed into the 'knowledge' in knowledge graphs. GraphDB's use of Linked Data captures the relationships between the data, which is the 'graph' part. This representation of data's semantics and its relationships create a contextual understanding that is used by AI to make those seemingly intelligence decisions. [7]

1.10 History of knowledge graph

In the 1980s, researchers from the University of Groningen and the University of Twente in the Netherlands initially introduced the term knowledge graph to formally describe their knowledge-based system that integrates knowledge from different sources for representing natural language. The authors proposed KGs with a limited set of relations and focus on qualitative modeling including human interaction, which clearly contrasts with the idea of KGs that has been widely discussed in recent years. In 2012, Google introduced the Knowledge Graph as a semantic enhancement of Google's search function that does not match strings, but enables searching for \things", in other words, real-world objects. Since 2012, the term knowledge graph is also used to describe a family of applications. Frequently mentioned implementations are DBpedia, YAGO (Yet Another Great Ontology), Freebase, Wikidata, Yahoo's semantic search assistant tool Spark, Google's Knowledge Vault, Microsoft's Satori and Facebook's entity graph. Those applications differ in their characteristics, such as architecture, operational purpose, and technology used.[51]

1.11 Conclusion:

In this section, we explained the notion of Knowledge graph using some definitions and explanation, we gave examples about it like DBpedia, and its relations with web, semantic web and artificial intelligence. Knowledge graph has completion approaches which work to find missing information. This information could be entities, type of entities or relation between two entities.

Therefore, based on these approaches, we will try to use Knowledge graph in database to expect missing values.

CHAPTER 2 MISSING VALUES PREDICTION TECHNIQUES

Chapter 2 Missing values prediction techniques

2.1 Introduction

When we talk about detecting missing values, we need prediction techniques. There are many techniques of prediction. Each one of these techniques has special way to work in like deletion techniques, imputation (Time-series problem, general problem).

In this section we saw the effect of missing values in database we explained deferent missing values prediction techniques.

2.2 Categories of missing values

Missing values can be categorized in three types:

2.2.1 Missing Completely at Random (MCAR)

Values in a data set are said to be missing completely at random (MCAR) if the events that lead to any particular data item being missing are independent of both the observable variables and unobservable parameters of interest, and hence occurs entirely at random. Unbiased analysis is performed on the MCAR data. The data sets rarely have MCAR data. It signifie the maximum level of randomness. [43]

2.2.2 Missing at Random (MAR)

When the missing values of some attribute are not randomly distributed across the observations but are distributed within one or more samples, they are said to be missing at random (MAR). Or in other MAR does not depend on that particular attribute but depends on the values of another attribute. MAR is more common than the previous MCAR type. [43]

2.2.3 Not Missing at Random (NMAR)

NMAR is also known as non-ignorable missing value. In this case the missing data is dependent on the values of the attribute. NMAR signifies the least level of randomness. It is the most problematic form as it involves missing values that are not randomly distributed across the observations. The only way to deal with NMAR data is to attain an estimate of the parameters by modelling the miss. [43]

2.3 Challenges in prediction of missing values

Missing data directly impacts the data quality. If the missing data in a dataset is less than 1% of its total data then it does not cause a significant problem for knowledge discovery in

database. Also 1%-5% missing data is manageable to some extent while 5-15% needs sophisticated methods for handling. But if the missing data exceeds the 15% of total data, it severely affects the interpretation and the mining tasks in a negative way.

Handling of missing data through imputation methods have their own issues like loss of information or reduced efficiency, difficulty in data handling because of irregular data and systematic difference in the data. Thus, such issues make the prediction task a challenging one.[43]

The following challenges arise when we apply any technique to predict the missing values:

a) The missing data prediction method should not alter the distribution of data.
b) The relationships between the attributes of the data set must be retained by the prediction method deployed.
c) The prediction method should not be too complex and should not have high time cost factor.
d) The missing values should be predicted and replaced in such a way that all the data mining analytical procedures can be applied to the newly completed dataset easily.[43]

2.4 Prediction techniques

There are many prediction techniques, we list the most of them in the following.

2.4.1 Instance Deletion

The method of instance deletion is the most primitive approach used for handling the missing values in a dataset. It involves the complete deletion of the instances with missing data and analyzing the remaining complete data of the dataset.

Though it is easy to implement this technique but it has several consequences. First, it leads to reduction in the size of data set available for analysis which in turn gives inaccurate results for mining. Secondly, deleting entire instance causes biasing in the distribution of data and its statistical analysis because the data is not always missing at random. An improved version of this method can be deleting the attributes with high missing rate but only after running the relevance analysis.[43]

2.4.1.1 Listwise

Listwise deletion (complete-case analysis) removes all data for an observation that has one or more missing values. Particularly if the missing data is limited to a small number of observations, you may just opt to eliminate those cases from the analysis.

However, in most cases, it is often disadvantageous to use listwise deletion. This is because the assumptions of MCAR (Missing Completely at Random) are typically rare to support. As a result, listwise deletion methods produce biased parameters and estimates.[17]

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	N/A	N/A	95%
8	Lite	76	77%
9	Fast+	180	N/A

Deletion 4, 7 and 9:

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
5	Lite	76	70%
6	Fast+	155	10%
8	Lite	76	77%

Table 2.1 Example Listwise [5]

2.4.1.2 **Pairwise**

Pairwise deletion analyses all cases in which the variables of interest are present and thus maximizes all data available by an analysis basis. A strength to this technique is that it increases power in your analysis but it has many disadvantages. It assumes that the missing data are MCAR. If you delete pairwise then you'll end up with different numbers of observations contributing to different parts of your model, which can make interpretation difficult.[17]

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	N/A	N/A	95%
8	Lite	76	77%
9	Fast+	180	N/A

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+		80%
5	Lite	76	70%
6	Fast+	155	10%
7			95%
8	Lite	76	77%
9	Fast+	180	

Table 2.2 Example pairwise [5]

2.4.1.3 **Dropping Variables**

If there are too many data missing for a variable it may be an option to delete the variable or the column from the dataset. There is no rule of thumbs for this but depends on situation and a proper analysis of data is needed before the variable is dropped all together. This should be the last option and need to check if model performance improves after deletion of variable.[17]

Delete ↓

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	N/A	80%
2	Lite	N/A	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	N/A	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

→

Mobile ID	Mobile Package	Data Limit Usage
1	Fast+	80%
2	Lite	70%
3	Fast+	10%
4	Fast+	80%
5	Lite	70%
6	Fast+	10%
7	Fast+	95%
8	Lite	77%
9	Fast+	95%

Table 2.3 Example dropping variables.[5]

Advantage: Convenient to apply.[43]

Disadvantages: a) Reduction of size of dataset. b) Induction of bias. c) Reduction of accuracy for data mining.[43]

2.4.2 **Hot Deck Imputation**

This method involves a two-stage processing of data set with missing values. First the data set is partitioned into clusters. Then within every cluster the missing value are replaced with predicted values.[43]

Advantages:

1) Can predict both quantitative and qualitative attributes. [43]
2) Doesn't require a predictive model for each attribute with missing data.[43]

Disadvantage: Difficult to predict in non-related samples available for prediction.[43]

2.4.3 Mean, Median and Mode

In this imputation technique goal is to replace missing data with statistical estimates of the missing values. Mean, Median or Mode can be used as imputation value**.** [49].

In a mean substitution, the mean value of a variable is used in place of the missing data value for that same variable. This allows the researchers to utilize the collected data in an incomplete dataset. The theoretical background of the mean substitution is that the mean is a reasonable estimate for a randomly selected observation from a normal distribution.

However, with missing values that are not strictly random, especially in the presence of a great inequality in the number of missing values for the different variables, the mean substitution method may lead to inconsistent bias. Furthermore, this approach adds no new information but only increases the sample size and leads to an underestimate of the errors. Thus, mean substitution is not generally accepted. [49]

Mean= (157+99+167+76+155+76+180)/7 = **130**

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	130	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	130	95%
8	Lite	76	77%
9	Fast+	180	95%

Table 2.4 Example of Mean [5]

Median can be used when variable has a skewed distribution: {76,76,99,155,157,167,180}

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	155	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	155	95%
8	Lite	76	77%
9	Fast+	180	95%

Table 2.5 Example of Median [5]

Mode is to replace the population of missing values with the most frequent value.

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	N/A	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	N/A	95%
8	Lite	200	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	200	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	200	95%
8	Lite	200	77%
9	Fast+	180	95%

Table 2.6 Example of Mode [5]

Advantages:

1) Ease of Application.[43]

2) Predicts realistic values.[43]

3) Avoids Distortion in Imputation.[43]

Disadvantage:

1) Replacing all missing values with same mean changes the characteristic of original data set

2) Induces bias.[43]

2.4.4 Prediction using K- nearest neighbor algorithm (KNN)

The KNN prediction algorithm searches for the most similar instances of the data for predicting and replacing the missing values. It can be used for both numeric and nominal values. The value selected for 'k' has a huge impact on results.[43]

K-NN is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution i.e. the model structure determined from the dataset.[22]

It is called Lazy algorithm because it does not need any training data points for model generation. All training data is used in the testing phase which makes training faster and testing phase slower and costlier.[22]

K-Nearest Neighbor (K-NN) is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. [60]

Advantages:

a) Can predict both quantitative and qualitative attributes.[43]

b) Doesn't require a predictive model for each attribute with missing data.[43]

Disadvantage:

With the increase in size of data set, the selection of value of K becomes a critical issue.[43]

2.4.5 Regression Imputation

In Regression based imputation, it is assumed that a variable changes its value linearly with other variables or there is a linear relationship between the attributes of the data set. So, the missing data can be replaced by the linear regression function. But the drawback of this method is that generally the relationship among the attributes is not linear. Support Vector Machine Imputation (SVMI) is an example of regression-based imputation technique that takes condition and decision attributes. It is then applied for the prediction of values for the missed condition. [43]

Advantage: Works well with large size datasets.[43]

Disadvantage: Performance degrades if the samples taken are less than the features in the dataset.[43]

2.4.6 Prediction using Bayesian Iteration

The Naive Bayesian Classifier is a simple and popular classifier which gives good performance in terms of accuracy. Prediction using Bayesian Iteration also consists of two phases. Initially the order of attribute with missing values is decided based on parameters such

as information gain, weighted index etc. Thereafter the Naive Bayesian Classifier is used to predict the missing data in an iterative process. In first iteration, the algorithm replaces the missing value in the first attribute in the order and then goes to next attributes in further iterations.[43]

Advantage: Provides good performance by using Naïve Bayesian classifier.[43]

Disadvantage: Can deal with only the nominal attributes for prediction.[43]

2.4.7 Linear Interpolation

Interpolation is a mathematical method that adjusts a function to data and uses this function to extrapolate the missing data. The simplest type of interpolation is the linear interpolation, that makes a mean between the values before the missing data and the value after. Of course, we could have a pretty complex pattern in data and linear interpolation could not be enough. There are several different types of interpolation. Just in Pandas we have the following options like : 'linear', 'time', 'index', 'values', 'nearest', 'zero', 'slinear', 'quadratic', 'cubic', 'polynomial', 'spline', 'piece wise polynomial' and many more . [5]

N/A 1= (90+150)/2=120 N/A2= (160+180)/2=170

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	70%
3	3-Jan	167	10%
4	4-Jan	90	80%
5	5-Jan	N/A	70%
6	6-Jan	150	10%
7	7-Jan	160	95%
8	8-Jan	N/A	77%
9	9-Jan	180	95%

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	200	80%
2	2-Jan	100	70%
3	3-Jan	200	10%
4	4-Jan	90	80%
5	5-Jan	120	70%
6	6-Jan	200	10%
7	7-Jan	160	95%
8	8-Jan	170	77%
9	9-Jan	180	95%

Table 2.7 Example of Linear Interpolation. [5]

2.4.8 Last Observation Carried Forward (LOCF)

If data is time-series data, one of the most widely used imputation methods is the last observation carried forward (LOCF). Whenever a value is missing,

it is replaced with the last observed value. This method is advantageous as it is easy to understand and communicate. Although simple, this method strongly assumes that the value of the outcome remains unchanged by the missing data, which seems unlikely in many settings.[5]

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	70%
3	3-Jan	167	10%
4	4-Jan	90	80%
5	5-Jan	N/A	70%
6	6-Jan	150	10%
7	7-Jan	N/A	95%
8	8-Jan	N/A	77%
9	9-Jan	180	95%

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	70%
3	3-Jan	167	10%
4	4-Jan	90	80%
5	5-Jan	90	70%
6	6-Jan	150	10%
7	7-Jan	150	95%
8	8-Jan	150	77%
9	9-Jan	180	95%

Table 2.8 Example of last observation.[5]

2.4.9 Fuzzy k- means clustering imputation

In Fuzzy k- means clustering imputation method, every data object is assigned a membership function. This membership function signifies the degree of belongingness of that data object to any particular cluster. Then on the basis of these membership function and cluster centroid values, this method substitutes the missing values in the dataset.[43]

Advantage:

Better performance output than simple K means prediction as data objects can belong to more than one cluster.[43]

Disadvantage:

a) High implementation cost in terms of computation time.[43]

b) Highly sensitive to noise.[43]

2.4.10 Prediction using C4.5 algorithm

Internal methods C4.5 is a tree-based classifier used widely. It has been further improved for handling missing data by developing some internal algorithms. C4.5 uses the probabilistic approach to treat missing data. It first selects the attribute from the dataset based on correctional gain ratio and then distributes all the missing data instances into subsets based on probability of size of subset. The decision tree classifies the instances and searches all possible paths. Finally, a classification result is obtained in terms of probability. [43]

Advantage:

a) It can be applied to both nominal and numeric attributes
b) Searches all possible paths to give result in form of classification. [43]

Disadvantage:

a) Computation time increases significantly with increase in size of dataset. [43]

2.4.11 Maximum likelihood

There are a number of strategies using the maximum likelihood method to handle the missing data. In these, the assumption that the observed data are a sample drawn from a multivariate normal distribution is relatively easy to understand. After the parameters are estimated using the available data, the missing data are estimated based on the parameters which have just been estimated.[49]

When there are missing but relatively complete data, the statistics explaining the relationships among the variables may be computed using the maximum likelihood method. That is, the missing data may be estimated by using the conditional distribution of the other variables.[49]

2.4.12 Expectation-Maximization

Expectation-Maximization (EM) is a type of the maximum likelihood method that can be used to create a new data set, in which all missing values are imputed with values estimated by the maximum likelihood methods. This approach begins with the expectation step, during which the parameters (e.g., variances, covariances, and means) are estimated, perhaps using the listwise deletion. Those estimates are then used to create a regression equation to predict the missing data. The maximization step uses those equations to fill in the missing data. The expectation step is then repeated with the new parameters, where the new regression equations are determined to "fill in" the missing data. The expectation and maximization steps are repeated until the system stabilizes, when the covariance matrix for the subsequent iteration is virtually the same as that for the preceding iteration. [49]

An important characteristic of the expectation-maximization imputation is that when the new data set with no missing values is generated, a random disturbance term for each imputed value is incorporated in order to reflect the uncertainty associated with the imputation. However, the expectation-maximization imputation has some disadvantages. This approach can take a long time to converge, especially when there is a large fraction of missing data, and it is too complex to be acceptable by some exceptional statisticians. This approach can lead to the biased parameter estimates and can underestimate the standard error. [49]

For the expectation-maximization imputation method, a predicted value based on the variables that are available for each case is substituted for the missing data. Because a single imputation omits the possible differences among the multiple imputations, a single imputation will tend to underestimate the standard errors and thus overestimate the level of precision. Thus, a single imputation gives the researcher more apparent power than the data in reality.[49]

2.5 Techniques for Handling the Missing Data

The best possible method of handling the missing data is to prevent the problem by well-planning the study and collecting the data carefully.[49]

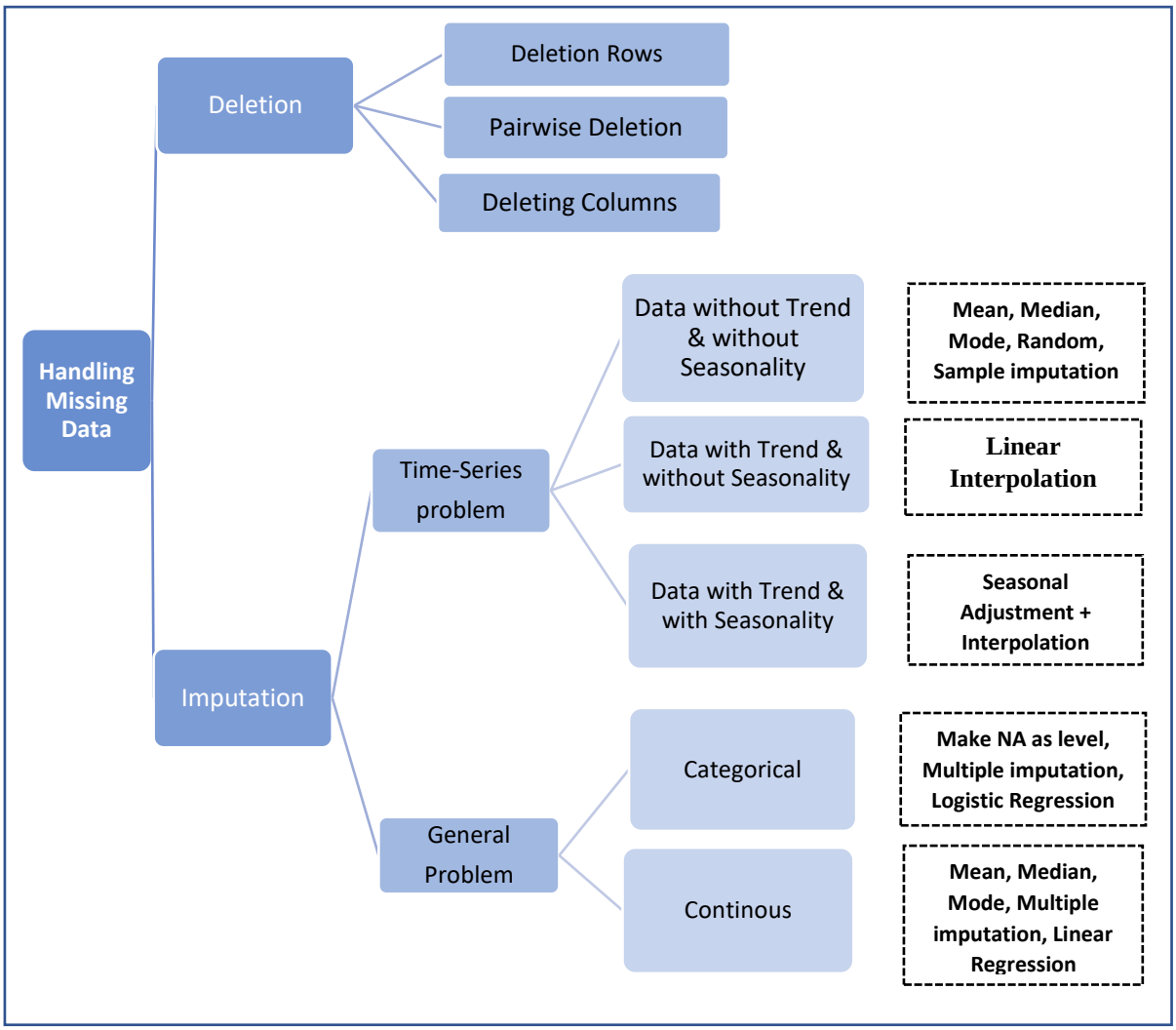

Figure 2.1 Handling missing data diagram. [17]

2.6 Conclusion

Categories of missing values are shown in this chapter and challenges in prediction of missing values are clarified. We have mentioned the prediction techniques which are supported by some examples. Then, we have search and pick out some of its advantages and disadvantages. In the end, we explained handling missing data in a diagram.

CHAPTER 3 PREDICTION OF MISSING VALUES USING KNOWLEDGE GRAPH

3.1 Introduction:

This chapter presents an application case of knowledge graphs which we used it to detect and calculate the missing value in a data table. We take data base of students then transform it to graph. This data base has some missing value that we will search about it.

Knowledge graph has different prediction techniques concerning link as common neighbors, Jaccard's coefficient, Adamic/Adar…, we apply common neighbors concept to find the missing value.

3.2 Reminder on knowledge graphs

The knowledge graph (KG) represents a collection of interlinked descriptions of entities – real-world objects, events, situations or abstract concepts.[55]

3.2.1 What is the knowledge graph?

When say knowledge graph we do not mean bar charts, pie charts, and line plots when we say graphs. Here, we are talking about interconnected entities which can be people, locations, organizations, or even an event. [57]

We can define a graph as a set of nodes and edges.

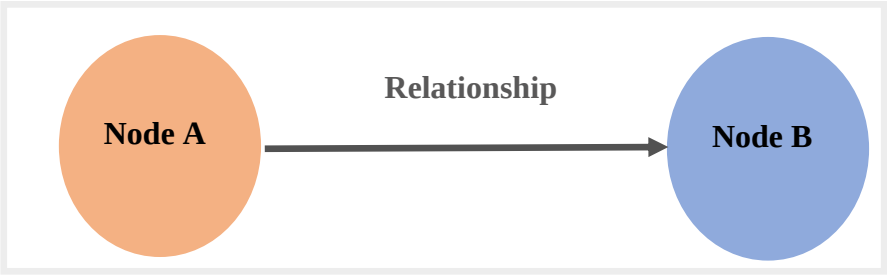

Figure 3.1 example of graph

Node A and Node B here are two different entities. These nodes are connected by an edge that represents the relationship between the two nodes.[57]

3.2.2 How to represent Knowledge in a graph?

Before we get started with building Knowledge Graphs, it is important to understand how information or knowledge is embedded in these graphs.

Example: *Abdelmadjid Tebboune is an Algerian politician currently serving as the president of Algeria since December 2019, and was born on 17 November 1945 in Mécheria, Algeria.*

The rule when extraction entities like form (subject-object-predicate)

- Subject : Abdelmadjid Tebboune.
- Predicate : presidentOf.
- Object : Algeria.

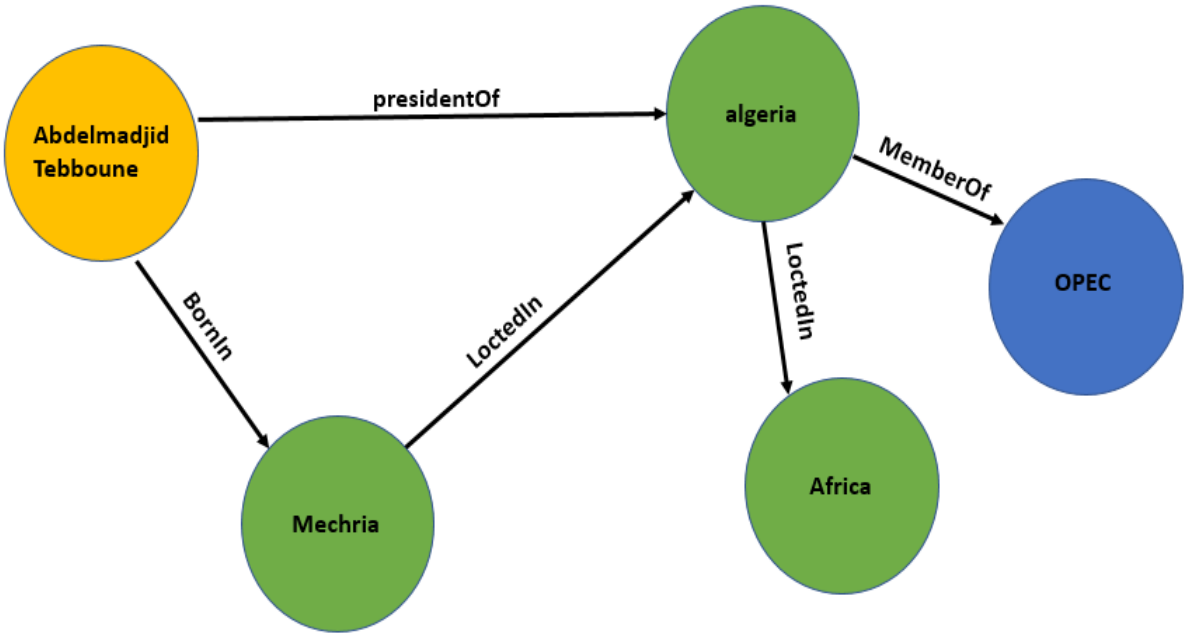

Figure 3.2 An example of Knowledge graph.

3.3 Converting a Database relational to a graph

Steps to convert a relational database model using the schema database model:

- Table to Node Label each entity table in the relational model becomes a label on nodes in the graph model.
- *Row to Node* each row in a relational entity table becomes a node in the graph.
- *Column to Node Property* – columns (fields) on the relational tables become node properties in the graph.
- Foreign keys to Relationships – replace foreign keys to the other table with relationships(edge), remove them afterwards.
- *Join tables to Relationships* join tables are transformed into relationships, columns on

those tables become relationship properties.[40]

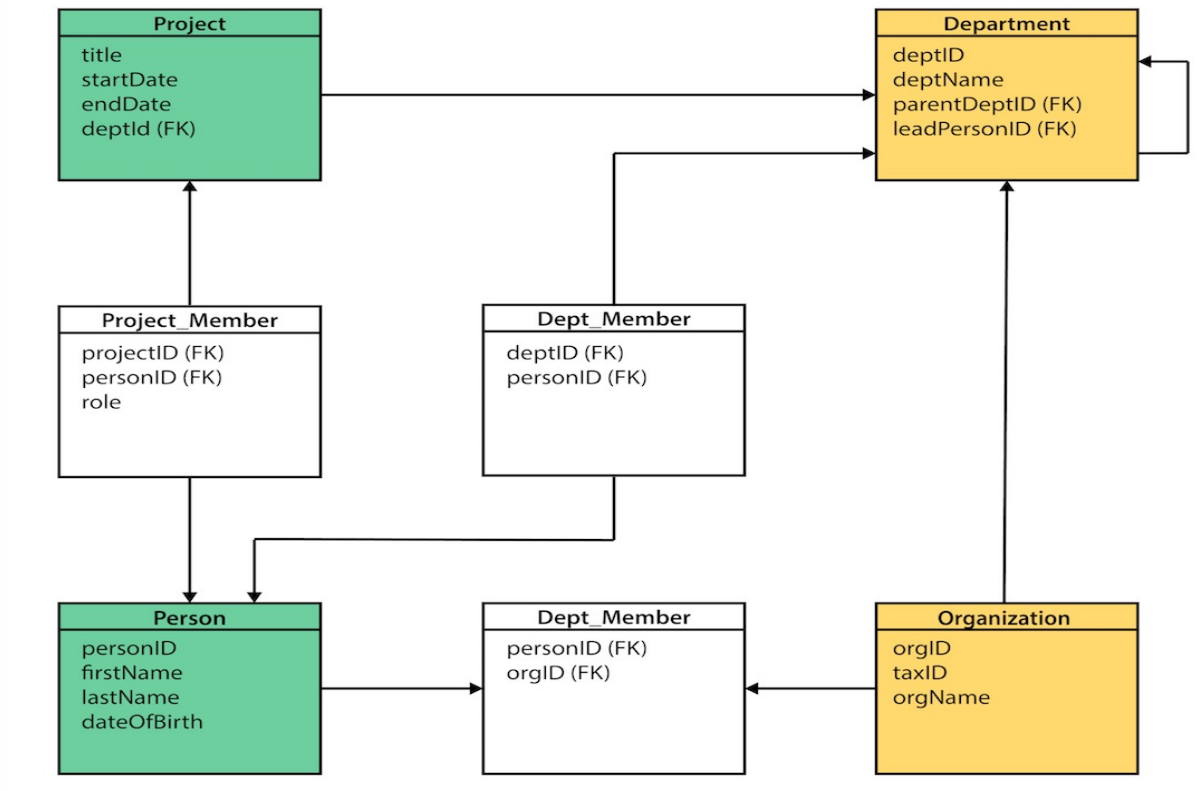

Figure 3.3 A relational database model of a domain with people and projects within an organization with several departments.[40]

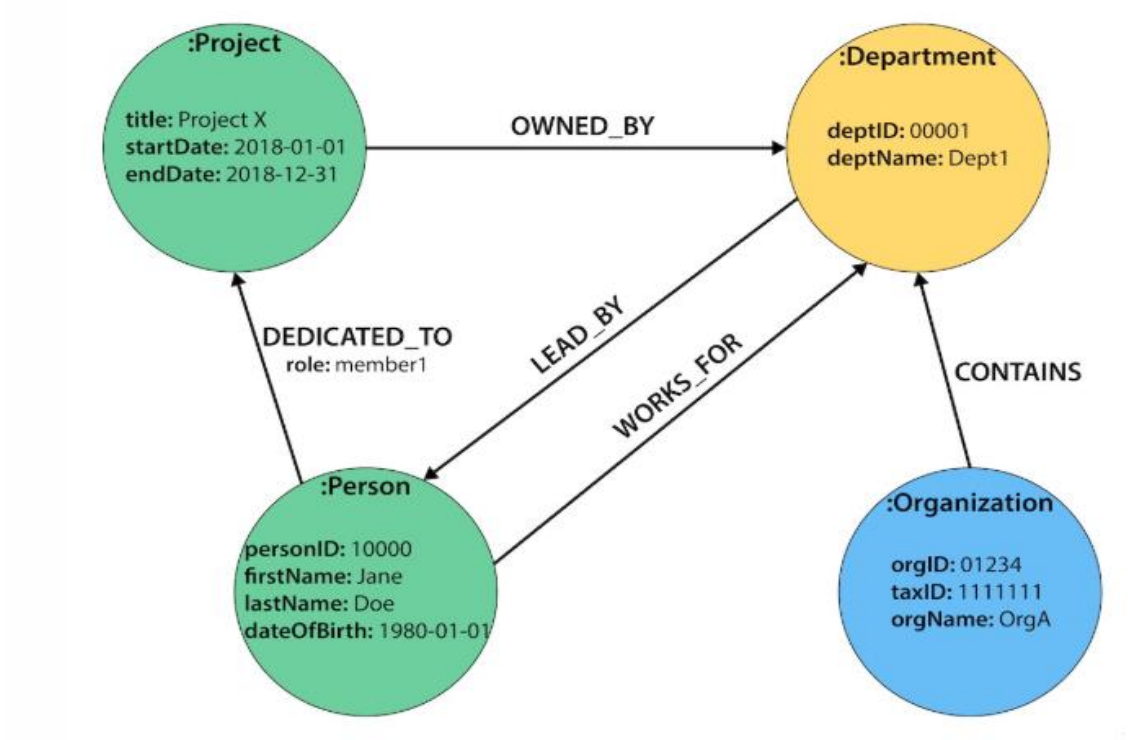

Figure 3.4 Knowledge graph of database [40]

3.4 Implementation of the knowledge graph

In this stage, we have table of students and their marks. We will transform them into graphs with Neo4j tool.

3.4.1 Presentation of Noe4j Tool

Neo4j is an open-source, NoSQL native graph database that provides an ACID compliant transactional backend for your applications. Initial development began in 2003, but it has been publicly available since 2007. The source code was written in Java and Scala, Neo4j is referred to as a native graph database because it efficiently implements the property graph model down to the storage level. This means that the data is stored exactly as you whiteboard it, and the database uses pointers to navigate and traverse the graph.[54]

It also has a variety of extension libraries and developer tools that can be added to existing products to enhance functionality.[41]

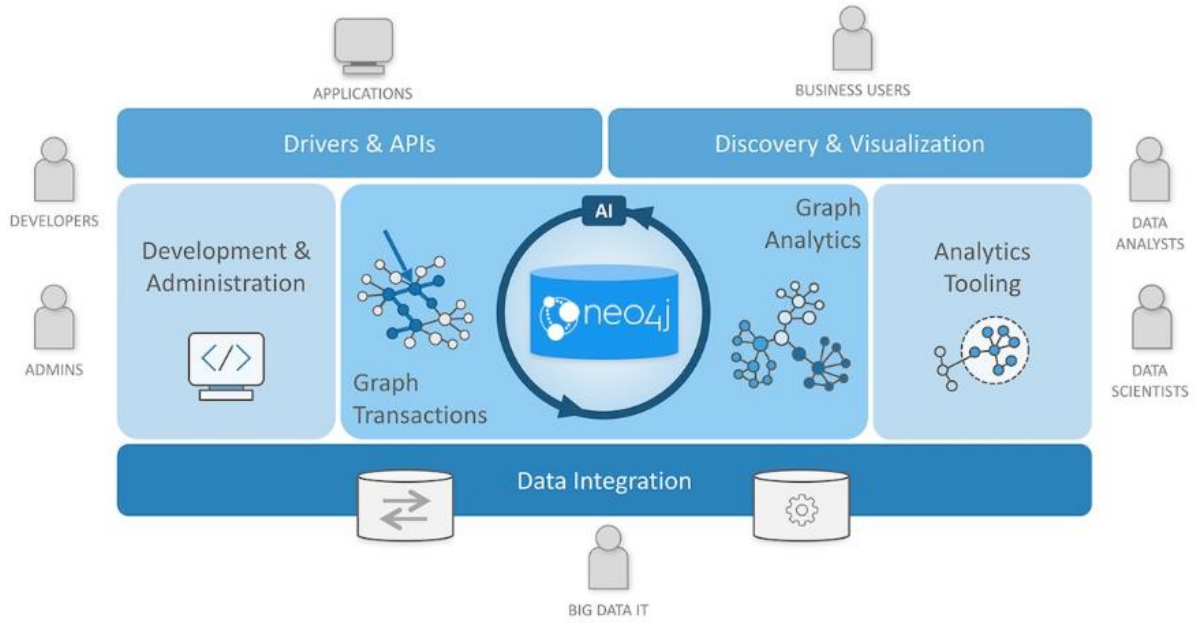

Figure 3.5 Components of the neo4j graph platform. [41]

3.4.2 **Presentation of the student database**

The student database is composed of the following tables:

Table student: contain students Id, First name, Last name.

Table module: contain module Id and Name module.

Table marks: contain Student Id and Module Id.

ModId	ModName
1	DAD
2	IA
. .	. .
n	TestLog

Table 3.1 Modules table

StudId	First name	Last name
1	F1	L1
2	F2	L2
. . .	. . .	. . .
n	Fn	Ln

Table 3.2 Students table

StudId	ModId	Marks
1	1	14
1	2	11
2 . .	1 . .	13 . .
N	1	16

Table 3.3 Marks table

We transformed student database into entity-relationship model.

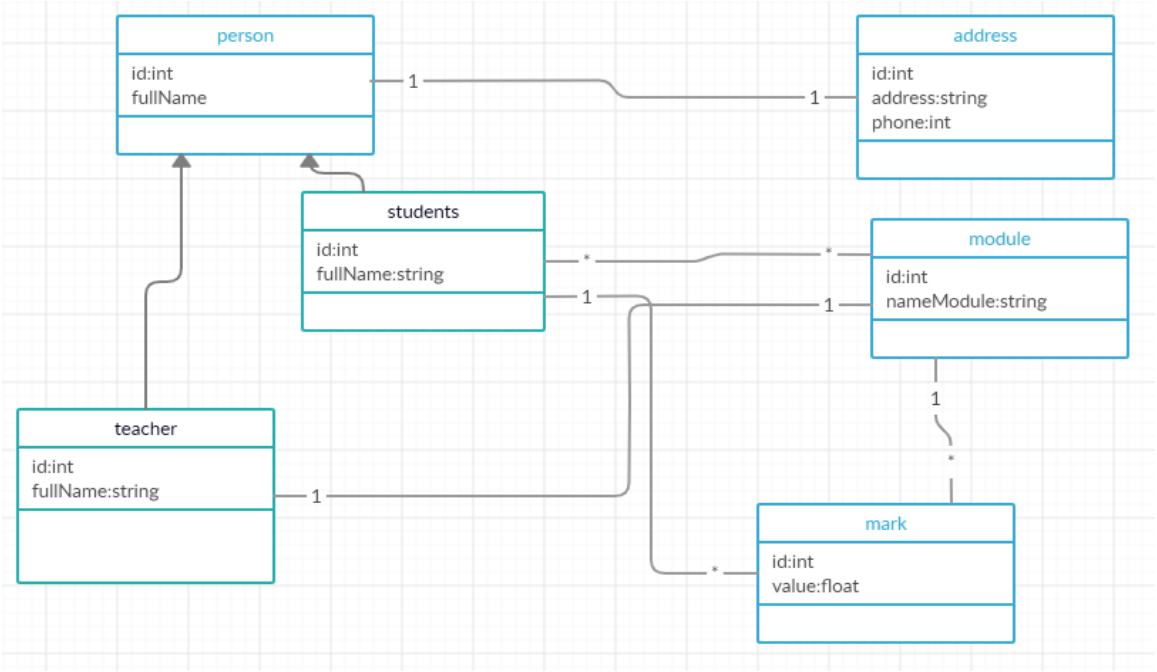

Figure 3.6 Student database as an entity-relationship model.

When we represent students database in Neo4j we will get diagram bellow:

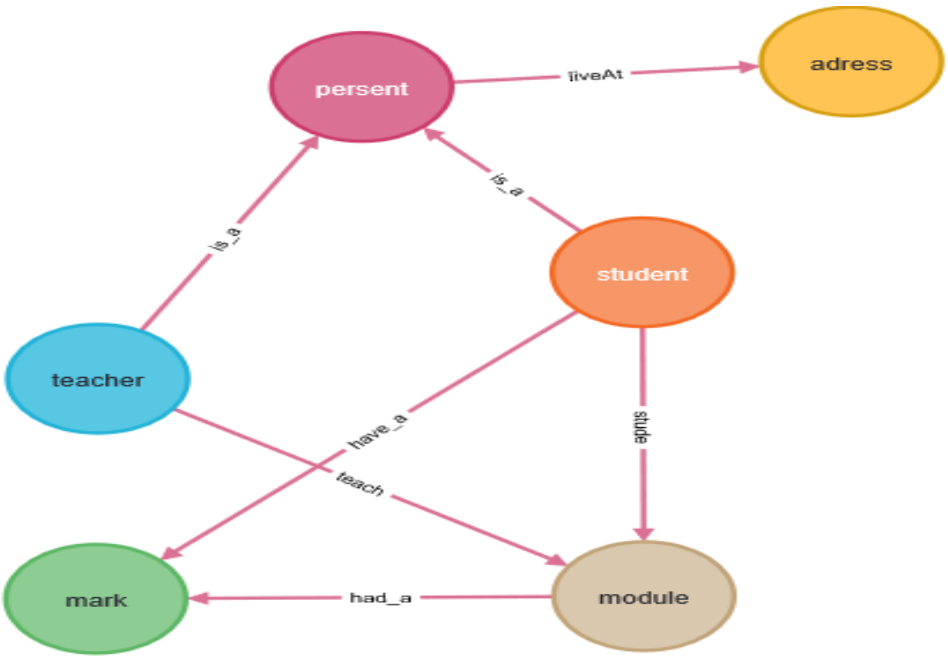

Figure 3.7 Student database in Neo4j.

3.4.3 **Our work**

We take table3.4 of students then we will represent it as a graph. After that, we will delete some values from both in order to apply one of the missing values prediction techniques.

N	Last name	First name	Inscription.N	DAD marks
1	ABDESLAM	WISSAME	19115066444	14.5
2	ABDOUNE	SOUAD	M201533071893	15
3	ABLI	MOHAMMED	M201535119912	11
4	ADOUANE	NOUR EL HOUDA	M201535100739	14
5	BAKHTI	ISLAM NADJIB	19054101720	12
6	BAKRI	FATIHA	M201535102826	16
7	BARKATI	NASSIBA	M201535105892	12
8	BELHADJAMI	MESSAOUD	M201435085251	12
9	BENAZOUZ	KHAOULA	M291535099854	16
10	BOURAS	AMINA BOCHRA	M201535097754	16

Table 3.4 DAD module marks.

To convert database to a graph using neo4j tool, we use the following colors for the nodes:

1. Orange node is Students.
2. Green node is Marks.
3. Brown node is Modules.
4. Blue node is Teacher.

If the table contains missing value, this value will be represented with X.

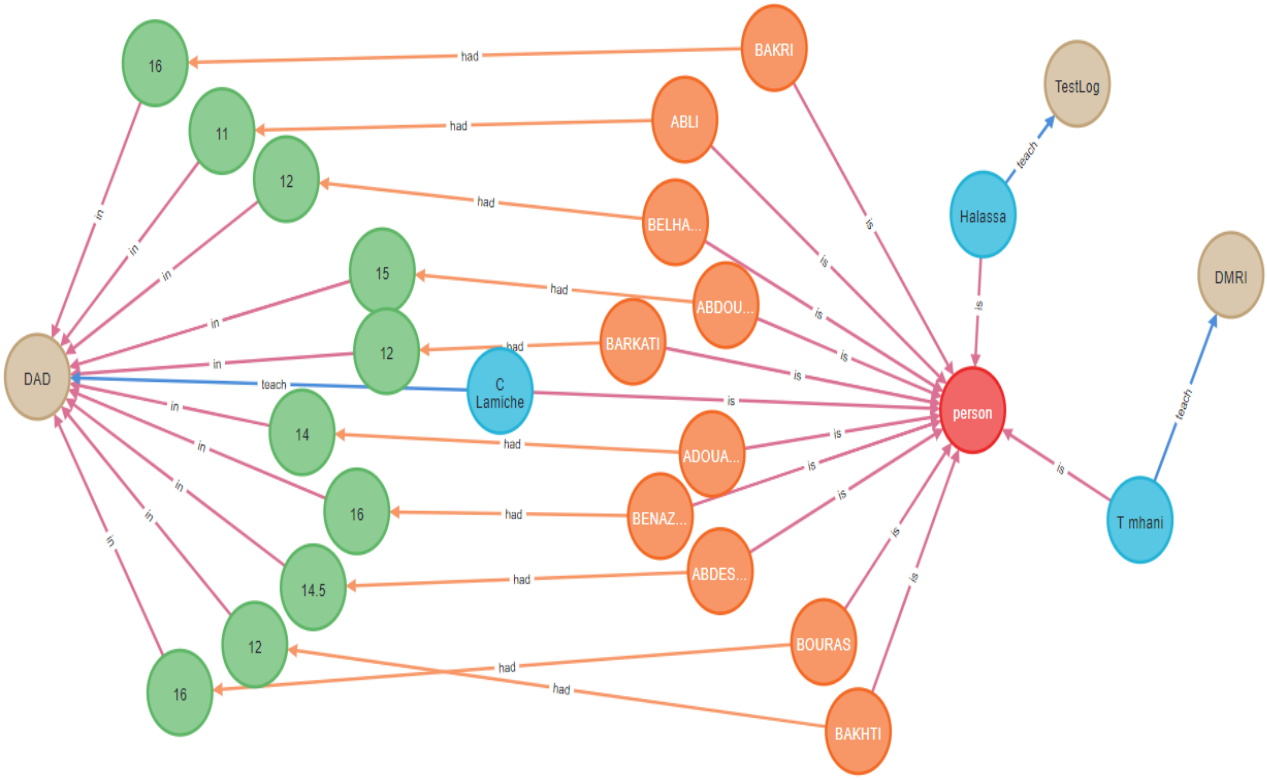

Figure 3.8 Students table as graph.

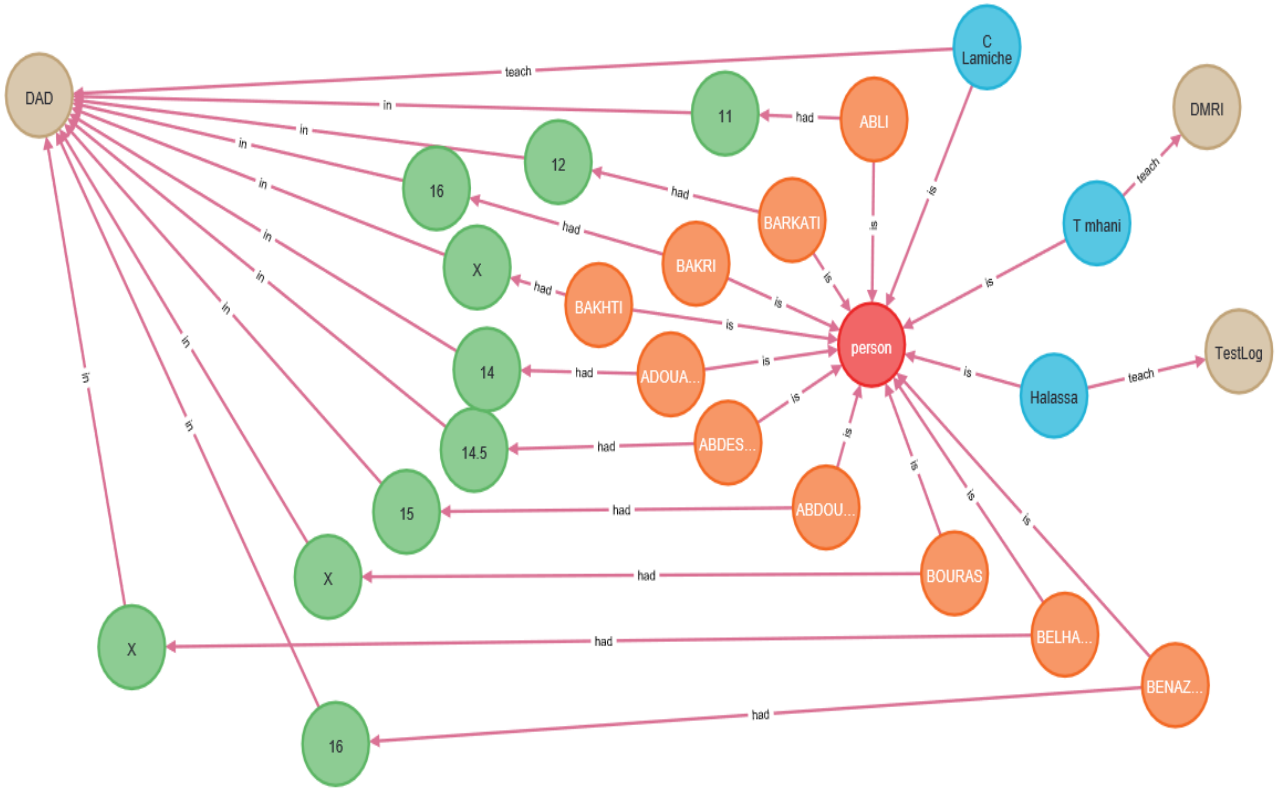

Figure 3.9 Graph with missing values.

3.5 Link prediction algorithms

The link prediction is an important research field in data mining. It has a wide range of scenarios. Many data mining tasks involve the relationship between the objects. Link prediction can be used for recommendation systems, social networks, information retrieval, and many other fields.[48]

Kleinberg and Liben-Nowell describe a set of methods that can be used for link prediction. Among these methods: graph distance, Jaccard's coefficient, adamic/Adar and common neighbors.[36]

3.5.1 Application of Common neighbors method

We choose this method (common neighbors) because it works to detect the existence of relationship between two nodes. Therefore, if there are empty nodes (in our case empty nodes represent missing values) we will find the missing values approximate to real values using the relationships expected by common neighbors.

3.5.1.1 Common neighbors

The common-neighbors predictor captures the notion that two strangers who have a common friend may be introduced by that friend. This introduction has the effect of "closing a triangle" in the graph and feels like a common mechanism in real life.[35]

We may find a relation between two nodes, if they have a common neighbor.

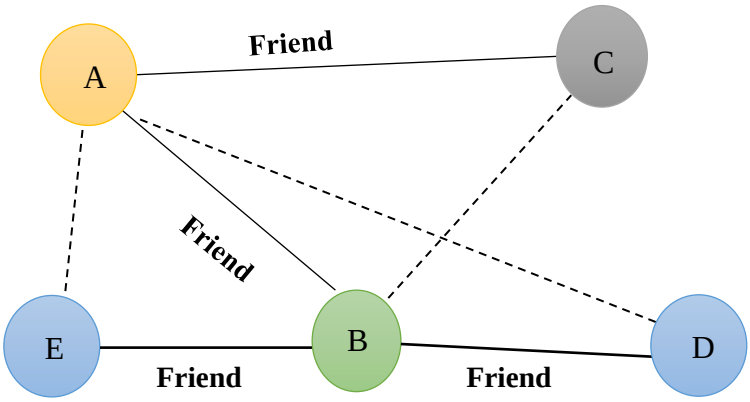

Figure 3.10 Example of common neighbors concept.

Node A is friend of B and C so maybe there is a relation between B and C. In addition, node B is friend of A and D so maybe there is a relation between A and D.

3.5.1.2 predicting missing values

To get missing value, we apply the mentioned idea as follows: We take part from table 3.4 without some values and we add another module (TestLog) to obtain best results.

N	Last name	First name	Inscription.N	DAD	TestLog
1	ABDESLAM	WISSAME	19115066444	14.5	13
2	ABDOUNE	SOUAD	M201533071893	15	11.5
3	ABLI	MOHAMMED	M201535119912	11	14
4	ADOUANE	NOUR EL HOUDA	M201535100739	X1	7
5	BAKHTI	ISLAM NADJIB	19054101720	12	9

Table 3.5 Student table in two modules

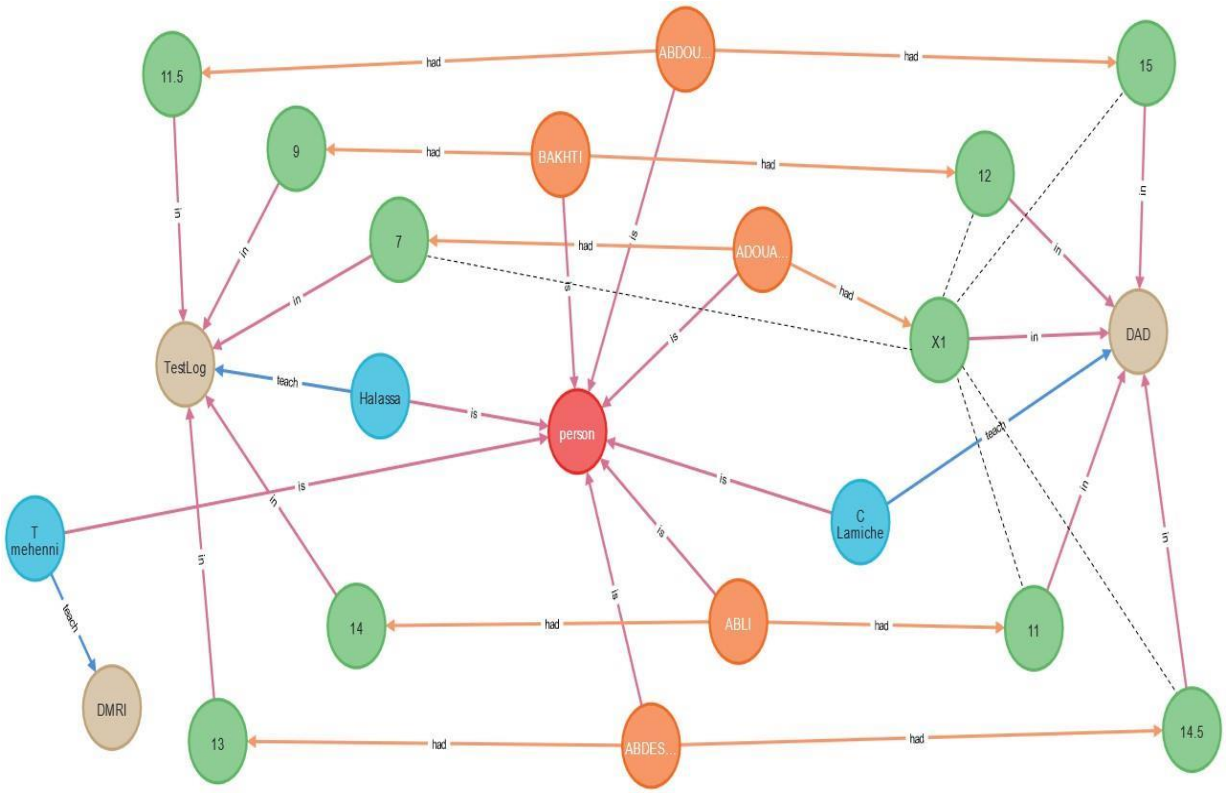

Figure 3.11 Table 3.5 into graph

In figure 3.11 we have missing value which represented by X1. Let's see neighbors of X1 and conclude what we are searching about.

So, X1 has DAD and Adouan as neighbors:

1. DAD has 15,12, 11 and 14.5 as neighbors.

2. Adouan has 7 as neighbors

There are two ways for predicting missing values: Mean of probabilities and Mean of neighbors

a) Predicting missing values by mean of probabilities method: where we take neighbors of X1 (DAD and Adouan) and we select neighbors of each one of them, then we choose each two neighbors (of DAD and Adouan) and calculate the mean of them. Therefore, we will get many probabilities of mean so we calculate the general mean (mean of probabilities) and put it in X1.

Example: X1 has DAD and Adouan as neighbors

DAD has as neighbors {15, 12,11,14.5} and Adouan has {7} as neighbors.

I. Probabilities of neighbors means of DAD are:

1) (15+12)/2 = 13.5
2) (15+11)/2 = 13
3) (15+14.5)/2 = 14.75
4) (12+11)/2 = 11.5
5) (12+14.5)/2 = 13.25
6) (11+14.5)/2 = 12.75

II. Probabilities of neighbors means of Adouan are: 7

General mean = (13.5 + 13 +14,75 + 11,5 + 13,25 + 12.75+7)/7 = 12.25

X1 = 12.25

b) **Predicting missing values by mean of neighbors method:** we calculate Mean of mean1 (of DAD) and mean2 (of Adouan)

where: (mean1= (15+12+11+14.5)/4 = 13.125)

(mean2 = 7)

X1= Mean ((mean1+mean2))

X1= (13.125 + 7)/2 = 10,06

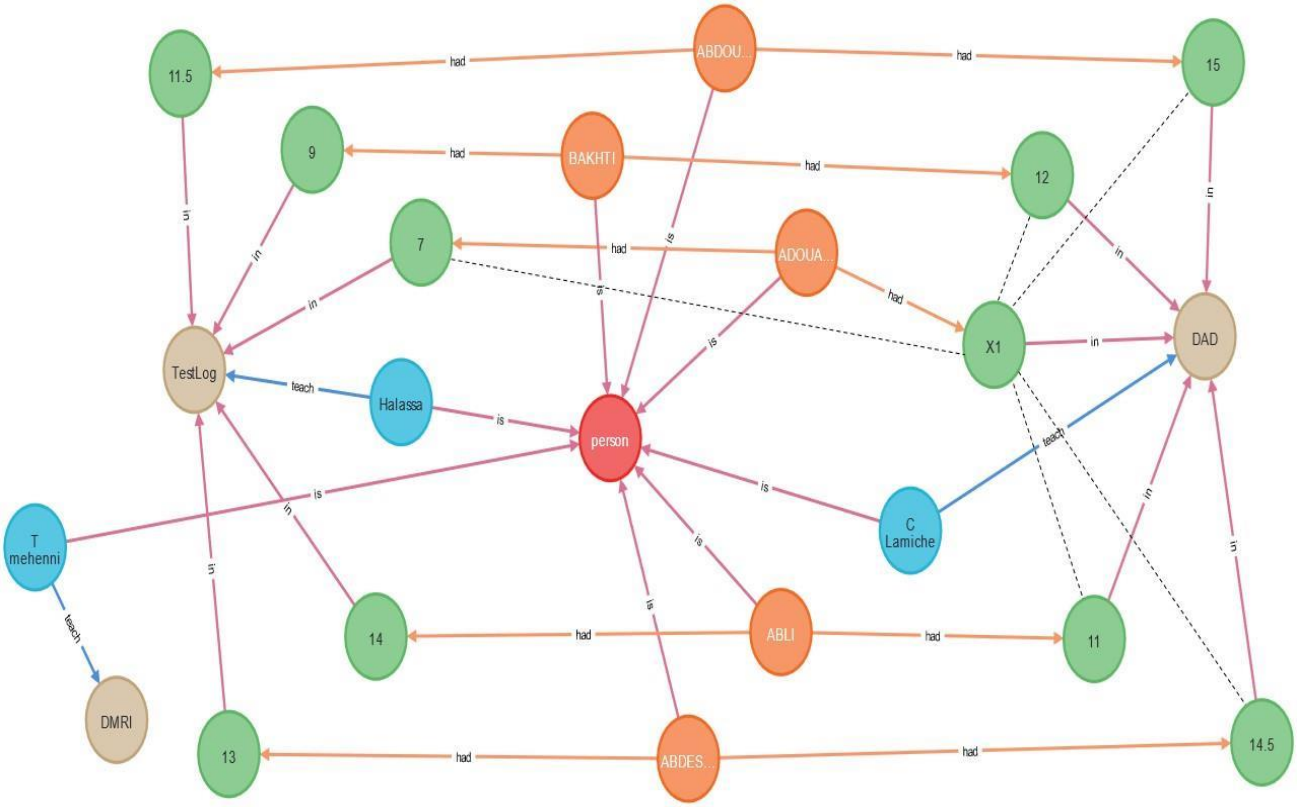

Figure 3.12 Neighbors of X1.

3.6 Predicting all the missing values in the database

3.6.1 Mean of probabilities:

Each step we will change X to get many missing values which we will predict. First, we have this information:

X1 → mark of Adouan Nour Elhouda in DAD.

X2 → mark of Abdoun Souad in DAD.

X3 → mark of Bakhti Islam Nadjib in DAD.

X4 → mark of Abli Mohemmed in DAD.

X5 → mark of Abdeslam Wissame in DAD.

X6 → mark of Adouan Nour Elhouda in TestLog.

X7 → mark of Abdoun Souad in TestLog.

X8 → mark of Bakhti Islam Nadjib in TestLog.

X9 → mark of Abli Mohemmed in TestLog.

X10 → mark of Abdeslam Wissame in TestLog.

1) Neighbors of X1 are DAD and Adouan

X1 = 12.25

2) Neighbors of X2 are DAD and Abdoun

DAD has as neighbors { 12,14,11,14.5} and Abdoun has {11.5} as neighbors.

- Probabilities of neighbors means of DAD are :

a) (12+14)/2 = 13 d) (14+11)/2 = 12.5

b) (12+11)/2 = 11.5 e) (14+14.5)/2 = 14.25

c) (12+14.5)/2 = 13.25 f) (11+14.5)/2 = 12.75

- Probabilities of neighbors means of Adouan are: 11.5

General mean = (13 + 11.5 + 13.25 + 12.5+14.25+12,75 +11.5)/7 = 12.67

X2 = 12.5

3) Neighbors of X 3 are DAD and Bakhti

DAD has as neighbors { 12,14,11,14.5} and Bakhti has {11.5} as neighbors.

- Probabilities of neighbors means of DAD are :

a) (15+14)/2 = 14.5 d) (14+11)/2 = 12.5

b) (15+11)/2 = 13 e) (14+14.5)/2 = 14.25

c) (15+14.5)/2 = 14.75 f) (11+14.5)/2 = 12.75

- Probabilities of neighbors means of Bakhti are: 9

General mean = (14.5 + 13 + 14.75 + 12.5 + 14.25 + 12,75 +9)/7 = 12.96

X3 = 13 (approximate value)

4) Neighbors of X4 are DAD and Abli

X4=(13.5+14.75+14.5+13+13.25+14.25+14)/7=13.89

X4 = 14 (approximate value)

5) Neighbors of X5 are DAD and AbdEslam

X5=(13.5+14.5+13+13+11.5+12.25+13)/7= 12.96

X5 = 13 (approximate value)

6) Neighbors of X6 are TestLog and Adouan

TestLog has as neighbors { 11.5, 9, 14,13} and Adouan has {14} as neighbors.

- Probabilities of neighbors means of DAD are :

a) (11.5+9)/2 = 10.25 d) (9+14)/2 = 11.5

b) (11.5+14)/2 = 12.75 e) (9+13)/2 = 11

c) (11.5+13)/2 = 12.25 f) (14+13)/2 = 13.5

- Probabilities of neighbors means of Abdoun are: 14

General mean = (10.25+ 12.75+ 12.25 + 11.5+ 11+ 13.5+14)/7 = 12.17

X6 = 12 (approximate value)

7) Neighbors of X7 are TestLog and Abdoun

X7 = 11.5 (approximate value)

8) Neighbors of X8 are TestLog and Bakhti

X8 = 11.5 (approximate value)

9) Neighbors of X9 are TestLog and Abli

X9 = 10.25 (approximate value)

10) Neighbors of X10 is TestLog and AbdEslam

X10 = 11 (approximate value)

After we expected missing values we obtain new student table :

N	Last name	First name	Inscription.N	DAD	TestLog
1	ABDESLAM	WISSAME	19115066444	13	11
2	ABDOUNE	SOUAD	M201533071893	12.5	11.5
3	ABLI	MOHAMMED	M201535119912	13	10.25
4	ADOUANE	NOUR EL HOUDA	M201535100739	12.25	12.25
5	BAKHTI	ISLAM NADJIB	19054101720	13	11.5

Table 3.6 Student table with new values.

3.6.2 **Mean of neighbors:**

In this way we will calculate mean of each neighbor then calculate general mean.

1) X1 has DAD and Adouan as neighbors, Mean1of DAD and Mean2 of Adouan

   Mean1=(14+12+11+14.5)/4 =13.12. Mean2= 7

   X1=General mean = (Mean1+ Mean2)/2 = (13.12 + 7)/2 = 10.06

   X1=10 (approximate value)

2) X2 has DAD and Abdoun as neighbors, Mean1of DAD and Mean2 of Abdoun

   Mean1=(12+14+11+14.5)/4 =12.875. Mean2= 11.5

   X2= (Mean1+ Mean2)/2 = (12.875+ 11.5)/2 = 12.18

   X2=12.25 (approximate value).

3) X3 has DAD and Bakhti as neighbors, Mean1of DAD and Mean2 of Bakhti

   Mean1=(15+14+11+14.5)/4 =13.12. Mean2= 9

   X3 = (13.62 + 9)/2 = 11,31

   X3 = 11.5 (approximate value)

4) X4 has DAD and Abli as neighbors, Mean1of DAD and Mean2 of Abli

   Mean1=(15+14+12+14.5)/4 =13.87. Mean2= 14

   X4 = (13.87 + 14)/2 = 13.93

   X4 = 14 (approximate value)

5) X5 has DAD and Abdeslam as neighbors, Mean1of DAD and Mean2 of Abdeslam

   Mean1=(15+14+12+11)/4 =13. Mean2= 13

   X5 = (13 + 13)/2 = 13

   X5 = 13

6) X6 has TestLog and Adouan as neighbors, Mean1of TestLog and Mean2 of Adouan

   Mean1=(14+12+11+14.5)/4 =12.87. Mean2= 14

   X6 (12.87 + 14)/2 = 13.43

   X6=13.5 (approximate value)

7) X7 has TestLog and Abdoun as neighbors, Mean1of TestLog and Mean2 of Abdoun

   Mean1=(9+7+14+13)/4 =10.75. Mean2= 15

   X7 = (10.75+ 15)/2 = 12.87

   X7 =13 (approximate value).

8) X8 has TestLog and Bakhti as neighbors, Mean1of TestLog and Mean2 of Bakhti

   Mean1=(11,5+7+14+13)/4 =11.37. Mean2= 12

   X8 = (11,37+ 12)/2 = 11.68

   X8 = 11.75 (approximate value)

9) X9 has TestLog and Abli as neighbors, Mean1of TestLog and Mean2 of Abli

   Mean1=(11.5+9+7+13)/4 =10.12. Mean2= 14

   X9 = (10.12 + 14)/2 = 12,06

   X9 = 12 (approximate value)

10) X10 has TestLog and Abdeslam as neighbors, Mean1of TestLog and Mean2 of Abdeslam

    Mean1=(11.5+9+7+14)/4 =10.37. Mean2= 14.5

    X10 = (10.37 + 14.5)/2 = 12.43

    X10 = 12.5 (approximate value)

Now we obtain new table with new expected value.

Table 3.7 New values by double mean

N	Last name	First name	Inscription.N	DAD	TestLog
1	ABDESLAM	WISSAME	19115066444	13	12.5
2	ABDOUNE	SOUAD	M201533071893	12.25	13
3	ABLI	MOHAMMED	M201535119912	14	12
4	ADOUANE	NOUR EL HOUDA	M201535100739	10	13.5
5	BAKHTI	ISLAM NADJIB	19054101720	11.5	11.75

3.7 Conclusion

In this chapter we used knowledge graph as technique of prediction missing values of database. To be more precise, we implemented concept of common neighbors which work to find missing link, then we selected all neighbors and we took values of nodes which had a common neighbor and we calculated all missing values which we put it and filled database with new values. In the next chapter we will evaluate and discuss this technique.

CHAPTER 4 EVALUATION AND DISCUSSION OF THE RESULTS

4.1 Introduction

When a new forecasting technique is discovered, the accuracy of its results must be calculated. One can said it is successful and implemented if its result is optimal. Based on this, we will calculate the accuracy of our method which we presented in the previous chapter.

This chapter introduces how the accuracy is measured and then, presents two techniques of measurement which are MAE and RMSE. We put all expected values we found by knowledge graph in tables (see chapter 3) and we compute the accuracy of this model using MAE and RMSE. Therefore, we will discuss the outcome and discover the advantages and disadvantages of our proposed technique.

4.2 Accuracy of Predictive Models

When developing predictive models and algorithms, whether linear regression or ARIMA models (**autoregressive integrated moving average)** it is important to quantify how well the model fits to the future observations. One of the simplest methods of calculating how correct a model is using the error between the predicted value and the actual value. From there, there are several methodologies that take this difference and further exploit meaning from it. Quantifying the accuracy of an algorithm is an important step to justifying the usage of the algorithm in product.[18]

4.2.1 Mean Absolute Error (MAE)

The mean absolute error (MAE) is defined as the sum of the absolute value of the differences between all the expected values and predicted values, divided by the total number of predictions. The equation form of MAE looks like this:[39]

Formula 1: Avg (Abs (Actual – Forecast)) [18]

Formula 2: $\frac{1}{n}\sum_{n}^{t=0}|At - Ft|$ [18]

The expected values are the answers you already know that are part of the training, validation or test sets, and the predicted values are the results predicted by the model for such inputs.[39]

4.2.2 Root Mean Squared Error (RMSE)

The root mean squared error (RMSE) seems somewhat similar to the MAE. They both take the difference between the actual and the forecast. However, the RMSE also then squares the difference, finds the average of all the squares and then finds the square root. Now it might seem like the action of squaring and then taking the square root may cancel each other out. This

isn't the case. The RMSE essentially punishes larger errors. Another way to phrase that is that it puts a heavier weight on larger errors.[18]

Formula 1: Sqrt (Avg (Power (Actual -Forecast))) [18]

Formula 2: $\sqrt{\frac{\sum |At-Ft|^2}{n}}$ [18]

4.2.3 Comparison between MAE and RMSE

Similarities: Both MAE and RMSE express average model prediction error in units of the variable of interest. Both metrics can range from 0 to ∞ and are indifferent to the direction of errors. They are negatively-oriented scores, which means lower values are better.[37]

Differences: Taking the square root of the average squared errors has some interesting implications for RMSE. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable.[37]

4.3 Expected values

After we explained how we used Knowledge graph to get missing values in database where we calculated missing values with two ways which each one based on:

First one used probability and second one depended on mean of neighbors. Then we will make a test to the expected values, and we will discuss the results.

We get many tables because of the implementation of our study on group of students with one module then we added other modules and, for best results, each case we will implement the test and compare the results with real values.

As we mentioned before, MAE will be applied on our database and its result will take for discuss it. So, we could evaluate our method.

4.3.1 Real marks of expected value

In the beginning we propose these marks of students in different modules before use it as missing values. Then use it to compare between it and expected values.

N	Last name	First name	DAD	TestLog	DMRI	POC	IA
1	ABDESLAM	WISSAME	14.5	13	16	16	9.75
2	ABDOUNE	SOUAD	15	11.5	12	15.5	8.75
3	ABLI	MOHAMMED	11	14	16.75	8.5	7
4	ADOUANE	HOUDA	14	7	8	13	8.5
5	BAKHTI	ISLAM	12	9	12	9	5
6	BAKRI	FATIHA	16	15.5	15.75	17	9
7	BARKATI	NASSIBA	12	13	17.25	8.5	5
8	BELHDJAMI	MESSAOUD	12	14.5	8	7.5	5.25

Table 0.1 Real marks of students

4.3.2 Evaluate the expected values

This is our example table which represents students marks in various modules. Let's take DAD and Testlog modules marks as missing values. So, our study consists of two parts.

4.3.2.1 Evaluate the mean of probabilities method

After we calculate all probabilities of DAD module and the other module marks of students, we obtain this table of probabilities mean:

	DAD	TestLog	Tlog+DMRI	Tlog+DMRI+POC	Tlog+DMRI+POC+IA
Mean of X1	13,4128571	13	14,5	15	13,6875
Mean of X2	13,0714286	11,5	11,75	13,33333333	12,1875
Mean of X3	13,6428571	14	15,375	13,08333333	11,5625
Mean of X4	13,2142857	7	7,5	9,333333333	9,125
Mean of X5	13,5	9	10,5	10	8,75
Mean of X6	12,9285714	15,5	15,625	16,08333333	14,3125
Mean of X7	13,5	13	15,125	12,91666667	10,9375
Mean of X8	13,5	14,5	11,25	10	8,8125

Table 1.2 Mean of probabilities of DAD

From these probabilities we calculated value of each X by this way, first case, mean of (DAD +Testlog), second mean of (DAD + Tlog + DMRI) then mean of (DAD + Tlog + DMRI +POC) Finally mean of (DAD + Tlog + DMRI + POC+ IA). So now we show the result of the test.

	Marks of DAD with TestLog				
	Full Name	Actual	Forecast	|Actual-Forecast|	|Actual-Forecast|²
X1	ABDESLAM WISSAME	14,5	13,0714286	1,428571429	2,040816327
X2	ABDOUNE SOUAD	15	12,2857143	2,714285714	7,367346939
X3	ABLI MOHAMMED	11	13,8214286	2,821428571	7,960459184
X4	ADOUANE NOUR HOUD	14	10,1071429	3,892857143	15,15433673
X5	BAKHTI ISLAM	12	9	3	9
X6	BAKRI FATIHA	16	14,2142857	1,785714286	3,18877551
X7	BARKATI NASSIBA	12	13,25	1,25	1,5625
X8	BELHADJAMI MESSAOU	12	14	2	4
MAE				2,361607143	6,284279337
RMSE					2,506846492

Table 1.3 Test result with two modules.

	Marks of DAD with TestLog and DMRI				
	Full Name	Actual	Forecast	|Actual-Forecast|	|Actual-Forecast|²
X1	ABDESLAM WISSAME	14,5	13,8214286	0,678571429	0,460459184
X2	ABDOUNE SOUAD	15	12,4107143	2,589285714	6,70440051
X3	ABLI MOHAMMED	11	14,5089286	3,508928571	12,31257972
X4	ADOUANE NOUR HOUD	14	10,3571429	3,642857143	13,27040816
X5	BAKHTI ISLAM	12	12	0	0
X6	BAKRI FATIHA	16	14,2767857	1,723214286	2,969467474
X7	BARKATI NASSIBA	12	14,3125	2,3125	5,34765625
X8	BELHADJAMI MESSAOU	12	12,375	0,375	0,140625
MAE				1,853794643	5,150699538
RMSE					2,269515265

Table 1.4 Test result with three modules

	Marks of DAD with TestLog and DMRI and Poc				
	Full Name	Actual	Forecast	|Actual-Forecast|	|Actual-Forecast|²
X1	ABDESLAM WISSAME	14,5	14,0714286	0,428571429	0,183673469
X2	ABDOUNE SOUAD	15	13,202381	1,797619048	3,23143424
X3	ABLI MOHAMMED	11	13,3630952	2,363095238	5,584219104
X4	ADOUANE NOUR HOUD	14	11,2738095	2,726190476	7,432114512
X5	BAKHTI ISLAM	12	11,75	0,25	0,0625
X6	BAKRI FATIHA	16	14,5059524	1,494047619	2,232178288
X7	BARKATI NASSIBA	12	13,2083333	1,208333333	1,460069444
X8	BELHADJAMI MESSAOU	12	11,75	0,25	0,0625
MAE				1,314732143	2,531086132
RMSE					1,590938758

Table 1.5 Test result with four modules.

	Marks of DAD with TestLog and DMRI and Poc +IA				
	Full Name	Actual	Forecast	|Actual-Forecast|	|Actual-Forecast|²
X1	ABDESLAM WISSAME	14,5	13,4151786	1,084821429	1,176837532
X2	ABDOUNE SOUAD	15	12,6294643	2,370535714	5,619439573
X3	ABLI MOHAMMED	11	12,9151786	1,915178571	3,66790896
X4	ADOUANE NOUR HOUD	14	11,1696429	2,830357143	8,010921556
X5	BAKHTI ISLAM	12	11,125	0,875	0,765625
X6	BAKRI FATIHA	16	13,6205357	2,379464286	5,661850287
X7	BARKATI NASSIBA	12	12,21875	0,21875	0,047851563
X8	BELHADJAMI MESSAOU	12	11,15625	0,84375	0,711914063
MAE				1,564732143	3,207793567
RMSE					1,791031425

Table 1.6 Test result with five modules.

As we note through the test tables, the values of the mean absolute error (MAE) is bordered between [1.31, 2.36] and the value of RMSE is bordered between [1.59, 2.50]. It means the result of the test is good.

Let's see test result of TestLog module below:

	TestLog	DAD	DAD+DMRI	DAD+DMRI+POC	DAD+DMRI+POC+IA
Mean of X1	12,07142857	14,5	15,25	15,5	14,0625
Mean of X2	12,28571	15	13,5	14,5	13,0625
Mean of X3	11,9285714	11	13,875	12,08333333	10,8125
Mean of X4	12,92857143	14	11	11,66666667	10,875
Mean of X5	12,64286	12	12	11	9,5
Mean of X6	11,7142857	16	15,875	16,25	14,4375
Mean of X7	12,25	12	14,625	12,58333333	10,6875
Mean of X8	11,8557142	12	10	9,166666667	8,1875

Table 1.7 Mean of probabilities of TestLog

	Marks of TestLog with DAD				
	Full Name	Actual	Forecast	|Actual-Forecast|	|Actual-Forecast|²
X1	ABDESLAM WISSAME	13	13,2857143	0,285714286	0,081632653
X2	ABDOUNE SOUAD	11,5	13,6428571	2,142857143	4,591836735
X3	ABLI MOHAMMED	14	11,4642857	2,535714286	6,429846939
X4	ADOUANE NOUR HOUD	7	13,4642857	6,464285714	41,7869898
X5	BAKHTI ISLAM	9	12,3214286	3,321428571	11,03188776
X6	BAKRI FATIHA	15,5	13,8571429	1,642857143	2,698979592
X7	BARKATI NASSIBA	13	12,125	0,875	0,765625
X8	BELHADJAMI MESSAOU	14,5	11,9285714	2,571428571	6,612244898
MAE				2,479910714	9,249880421
RMSE					3,041361606

Table 1.8 Test result with two modules.

	Marks of TestLog with DAD and DMRI				
	Full Name	Actual	Forecast	|Actual-Forecast|	|Actual-Forecast|²
X1	ABDESLAM WISSAME	13	13,6607143	0,660714286	0,436543367
X2	ABDOUNE SOUAD	11,5	12,8928571	1,392857143	1,94005102
X3	ABLI MOHAMMED	14	12,9017857	1,098214286	1,206074617
X4	ADOUANE NOUR HOUD	7	11,9642857	4,964285714	24,64413265
X5	BAKHTI ISLAM	9	12,3214286	3,321428571	11,03188776
X6	BAKRI FATIHA	15,5	13,7946429	1,705357143	2,908242985
X7	BARKATI NASSIBA	13	13,4375	0,4375	0,19140625
X8	BELHADJAMI MESSAOU	14,5	10,9285714	3,571428571	12,75510204
MAE				2,143973214	6,889180086
RMSE					2,624724764

Table 1.9 Test result with three modules.

	Marks of TestLog with DAD and DMRI and Poc				
	Full Name	Actual	Forecast	|Actual-Forecast|	|Actual-Forecast|²
X1	ABDESLAM WISSAME	13	13,7857143	0,785714286	0,617346939
X2	ABDOUNE SOUAD	11,5	13,3928571	1,892857143	3,582908163
X3	ABLI MOHAMMED	14	12,0059524	1,994047619	3,976225907
X4	ADOUANE NOUR HOUD	7	12,297619	5,297619048	28,06476757
X5	BAKHTI ISLAM	9	11,8214286	2,821428571	7,960459184
X6	BAKRI FATIHA	15,5	11,6904762	3,80952381	14,51247166
X7	BARKATI NASSIBA	13	12,4166667	0,583333333	0,340277778
X8	BELHADJAMI MESSAOU	14,5	10,5119048	3,988095238	15,90490363
MAE				2,646577381	9,369920103
RMSE					3,061032522

Table 1.10 Test result with four modules.

	Marks of TestLog with DAD and DMRI and Poc +IA				
	Full Name	Actual	Forecast	|Actual-Forecast|	|Actual-Forecast|²
X1	ABDESLAM WISSAME	13	13,0669643	0,066964286	0,004484216
X2	ABDOUNE SOUAD	11,5	12,6741071	1,174107143	1,378527583
X3	ABLI MOHAMMED	14	11,3705357	2,629464286	6,91408243
X4	ADOUANE NOUR HOUD.	7	11,9017857	4,901785714	24,02750319
X5	BAKHTI ISLAM	9	11,0714286	2,071428571	4,290816327
X6	BAKRI FATIHA	15,5	13,0758929	2,424107143	5,87629544
X7	BARKATI NASSIBA	13	11,46875	1,53125	2,344726563
X8	BELHADJAMI MESSAOU	14,5	10,0223214	4,477678571	20,04960539
MAE				2,409598214	8,110755142
RMSE					2,847938753

Table 1.11 Test result with five modules.

The result of the test is changed, as seen above the values of MAE became between [2.14, 2.64] as well the values of RMSE became between [2.62, 3.04]. That means the result is medium.

4.3.2.2 **Evaluate mean of neighbors**

In the same way of evaluation of probabilities mean method, we have evaluated the mean of neighbor's method. First, we calculated the mean of neighbors which we found it as follows:

	DAD	TestLog	TestLog+DMRI	TestLog+DMRI+Poc	TestLog+DMRI+Poc+IA
Mean of X1	13,1428571	13	14,5	15	13,6875
Mean of X2	12,8333333	11,5	11,75	13,33333333	12,1875
Mean of X3	13,2	14	15,375	13,08333333	11,5625
Mean of X4	13	7	7,5	9,333333333	9,125
Mean of X5	13,3333333	9	10,5	10	8,75
Mean of X6	12	15,5	15,625	16,08333333	14,3125
Mean of X7	13,25	13	15,125	12,91666667	10,9375
Mean of X8	14,75	14,5	11,25	10	8,875

Table 1.12 Mean of neighbors of DAD

	Marks of DAD with TestLog and DMRI				
	Name	Actual	Forecast	|Actual-Forecast|	|Actual − Forecast|^2
X1	ABDESLAM WISSAME	14,5	13,0714286	1,428571429	2,040816327
X2	ABDOUNE SOUAD	15	12,2857143	2,714285714	7,367346939
X3	ABLI MOHAMMED	11	13,8214286	2,821428571	7,960459184
X4	ADOUANE NOUR HOUD	14	10,1071429	3,892857143	15,15433673
X5	BAKHTI ISLAM	12	11,25	0,75	0,5625
X6	BAKRI FATIHA	16	14,2142857	1,785714286	3,18877551
X7	BARKATI NASSIBA	12	13,25	1,25	1,5625
X8	BELHADJAMI MESSOUAD	12	14	2	4
MAE				2,091836735	5,229591837
RMSE					2,286830085

Table 1.13 Test result with two modules

	Marks of DAD with TestLog and DMRI				
	Name	Actual	Forecast	|Actual-Forecast|	|Actual − Forecast|^2
X1	ABDESLAM WISSAME	14,5	13,8214286	0,678571429	0,460459184
X2	ABDOUNE SOUAD	15	12,4107143	2,589285714	6,70440051
X3	ABLI MOHAMMED	11	14,5089286	3,508928571	12,31257972
X4	ADOUANE NOUR HOUD	14	10,3571429	3,642857143	13,27040816
X5	BAKHTI ISLAM	12	12	0	0
X6	BAKRI FATIHA	16	14,2767857	1,723214286	2,969467474
X7	BARKATI NASSIBA	12	14,3125	2,3125	5,34765625
X8	BELHADJAMI MESSOUAD	12	12,375	0,375	0,140625
MAE				2,06505102	5,150699538
RMSE					2,269515265

Table 1.14 Test result with Three modules

	Marks of DAD with TestLog and DMRI and Poc				
	Name	Actual	Forecast	Actual-Forecast	|Actual − Forecast|^2
X1	ABDESLAM WISSAME	14,5	14,0714286	0,428571429	0,183673469
X2	ABDOUNE SOUAD	15	13,202381	1,797619048	3,23143424
X3	ABLI MOHAMMED	11	13,3630952	2,363095238	5,584219104
X4	ADOUANE NOUR HOUD	14	11,2738095	2,726190476	7,432114512
X5	BAKHTI ISLAM	12	11,75	0,25	0,0625
X6	BAKRI FATIHA	16	14,5059524	1,494047619	2,232178288
X7	BARKATI NASSIBA	12	13,2083333	1,208333333	1,460069444
X8	BELHADJAMI MESSOUAD	12	11,75	0,25	0,0625
MAE				1,466836735	2,531086132
RMSE					1,590938758

Table 1.15 Test result with four modules

	Marks of DAD with TestLog and DMRI and Poc +IA				
	Name	Actual	Forecast	|Actual-Forecast|	|Actual − Forecast|^2
X1	ABDESLAM WISSAME	14,5	13,4151786	1,084821429	1,176837532
X2	ABDOUNE SOUAD	15	12,6294643	2,370535714	5,619439573
X3	ABLI MOHAMMED	11	12,6026786	1,602678571	2,568578603
X4	ADOUANE NOUR HOUD	14	11,1696429	2,830357143	8,010921556
X5	BAKHTI ISLAM	12	11,125	0,875	0,765625
X6	BAKRI FATIHA	16	13,6205357	2,379464286	5,661850287
X7	BARKATI NASSIBA	12	12,21875	0,21875	0,047851563
X8	BELHADJAMI MESSOUAD	12	11,1875	0,8125	0,66015625
MAE				1,623086735	3,063907545
RMSE					1,75040211

Table 4.16 test result with five modules

Second method was giving this result: MAE limited between [1.46, 2.09] and RMSE limited between [1.59, 2,26]. We applied the same way to another module (TestLog) then we got MAE between [2.12, 2.48] and RMSE between [2.61, 3.04]

4.4 Discussion

So now, based on test results which are in the tables above, we start analyzing the suggested technique. Generally, the outcomes, in both methods, of MAE were 1.31 at least and 2.64 at most and the outcomes of RMSE were limited between 1.59 and 3.04. So, the average ratio error is reasonable.

Therefore, we can say that the idea of applying knowledge graph to find missing value is relatively logical, so we can use it as an adopted technique in the near future.

4.4.1 Advantages of model

One of the advantages of this model, is using the neighboring concept which is an interesting information of the actual node. In our example neighbors are modules or students:

a) While the neighbor is student the result will be near from the level of this student.
b) While the neighbor is module the result will be near from the level of the exam and students' marks.

In both cases, the outcome will be the average of them. As well, we noted in general whenever we add a module, the result will be improved. That means, if we increased information and edges then the result may become better than before.

4.4.2 **Disadvantage of model**

However, this model has also some disadvantages. In fact, we found that:

a) An exceptional value can change results and increase error rate. For example, a student got a weak mark in a module and other students got good marks, when we calculate mean, this mark give a big difference and it will change the average ratio error which will be large.

Conclusion

In this terminal chapter, we measured the accuracy of our model by two techniques MAE and RMSE. As well, we discussed the outcome of the test and we gave some advantages and disadvantage of our model. Therefore, we could say this idea of applying knowledge graph to find missing values in databases is reasonable and can be developed then adopted as missing values prediction technique.

CONCLUSION

The work with knowledge graph is one of the logical inspired idea to solve the problem of the missing values to be ready to process all the database. Starting with this idea and using link prediction techniques, we tried to build our project's parts.

In this dissertation we took database of our classmates and transformed it to a graph. The missing values was the exam marks that was represented with X nodes. Here, we embodied the previous idea by using the common neighbor concept which detect the missing links, based on these new links we select the new neighbors and calculated either all the probabilities of each two neighbors then the mean of these probabilities or the mean of neighbors.

After implementation of the model, we had to calculate its accuracy to unearth its effective. Finally, we discussed the results.

From results of test we discovered that implementation of knowledge graph to predict missing values is reasonable as a new technique. So, it is able to be developed and adopted.

In the near future, this study has to be taken into consideration as approach to follow up and use it in different domains to detect its defects and its difficulties then to develop and optimize it. Therefore, the knowledge graph deserves attention and follow-up, it could be a very interesting way to solve many problems of prediction.

Bibliography

[1] S. S. Rai, '3 Methods to Handle Missing Data'. https://blogs.oracle.com/datascience/3-methods-to-handle-missing-data (accessed Aug. 16, 2020).

[2] K. Maladkar, '5 Ways To Handle Missing Values In Machine Learning Datasets', *Analytics India Magazine*, Feb. 09, 2018. https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/ (accessed Aug. 16, 2020).

[3] 'About | DBpedia'. https://wiki.dbpedia.org/about (accessed Dec. 29, 2019).

[4] Stephanie, 'Absolute Error & Mean Absolute Error (MAE)', *Statistics How To*, Oct. 25, 2016. https://www.statisticshowto.com/absolute-error/ (accessed Aug. 16, 2020).

[5] Baijayanta Roy, 'All About Missing Data Handling Missing Data Imputation Techniques', Sep. 03, 2019. https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184 (accessed Feb. 27, 2020).

[6] C. Matuszek, J. Cabral, M. Witbrock, and J. DeOliveira, 'An Introduction to the Syntax and Content of Cyc.', Jan. 2006, pp. 44–49.

[7] Jarred McGinnis, 'Artificial Intelligence and the Knowledge Graph-Ontotext', Oct. 10, 2019. https://www.ontotext.com/blog/artificial-intelligence-and-the-knowledge-graph/ (accessed Jan. 02, 2020).

[8] Sebastien Dery, 'Challenges of Knowledge Graphs. From Strings to Things — An Introduction'. https://medium.com/@sderymail/challenges-of-knowledge-graph-part-1-d9ffe9e35214 (accessed Dec. 24, 2019).

[9] Nayantara Jeyaraj (Taro) and Nayantara Jeyaraj (Taro), 'Conceptualizing the Knowledge Graph Construction Pipeline', Mar. 12, 2019. https://towardsdatascience.com/conceptualizing-the-knowledge-graph-construction-pipeline-33edb25ab831 (accessed May 04, 2020).

[10] 'DATA -W3C'. https://www.w3.org/standards/semanticweb/data (accessed Dec. 24, 2019).

[11] Christian Bizera and and all, 'DBpedia - A crystallization point for the Web of Data', Sep. 2009.

[12] 'DBpedia - A crystallization point for the Web of Data - ScienceDirect'. https://www.sciencedirect.com/science/article/abs/pii/S1570826809000225 (accessed Aug. 16, 2020).

[13] Bill Slawski, 'Google Knowledge Graph Reconciliation-SEO by the Sea ⚓', May 08, 2019. http://www.seobythesea.com/2019/08/google-knowledge-graph-reconciliation/ (accessed Dec. 29, 2019).

[14] David LICOPPE, 'google konwledge graph: La method pour y etre presente', Sep. 24, 2019. https://premier-sur-google.be/algorithmes/google-knowledge-graph (accessed Dec. 29, 2019).

[15] 'graphgist', *Neo4j Graph Database Platform*. https://neo4j.com/graphgist/ (accessed Aug. 16, 2020).

[16] N. Sharma, 'How to build a Knowledge Graph : part-1 – ConfusedCoders'. https://confusedcoders.com/random/how-to-build-a-knowledge-graph-part-1 (accessed Aug. 16, 2020).

[17] Alvira Swalin, 'How to Handle Missing Data', Jan. 31, 2018. https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4 (accessed Feb. 28, 2020).

[18] 'How To Measure The Accuracy Of Predictive Models', *Acheron Analytics*. http://www.acheronanalytics.com/2/post/2018/05/how-to-measure-the-accuracy-of-predictive-models.html (accessed Aug. 18, 2020).

[19] Jose Manuel, Gomez-Perez, and Ronald Denaux, ‘Hybrid Techniques for Knowledge-Based NLP - Knowledge Graphs Meet Machine Learning and All Their Friends’, in *The Semantic Web – ISWC 2018*, Denny Vrandečić Google San Francisco, CA USA, and All., Monterey, CA, USA, 2018, p. 30.
[20] A. Neelakantan and M.-W. Chang, ‘Inferring Missing Entity Type Instances for Knowledge Base Completion: New Dataset and Methods’, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, May 2015, pp. 515–525, doi: 10.3115/v1/N15-1054.
[21] Stephanie, ‘Jaccard Index / Similarity Coefficient’, *Statistics How To*, Dec. 03, 2016. https://www.statisticshowto.com/jaccard-index/ (accessed Aug. 16, 2020).
[22] A. Chakure, ‘K-Nearest Neighbors (KNN) Algorithm’, *Medium*, Jun. 07, 2020. https://towardsdatascience.com/k-nearest-neighbors-knn-algorithm-bd375d14eec7 (accessed Aug. 16, 2020).
[23] Yuqing Gao, Jisheng Liang, Benjamin Han, Mohamed Yakout, and Ahmed Mohamed, ‘KDD-2018 Tutorial T39 Building a Large-scale, Accurate and Fresh Knowledge Graph’, ICC Capital Suite Room 10 (Level 3), ExCel London, Aug. 13, 2018, [Online]. Available: https://kdd2018tutorialt39.azurewebsites.net/KDD%20Tutorial%20T39.pdf.
[24] ‘Knowledge Graph - Technology and Its Applications in Business and Beyond’. https://softaria.com/knowledge-graph-technology-and-its-applications-in-business-and-beyond/ (accessed Aug. 16, 2020).
[25] S. Lenka, ‘Knowledge Graph — Algorithmic implementation of a thought process?🧐 — Part I’, *Medium*, May 30, 2019. https://blog.usejournal.com/knowledge-graph-algorithmic-implementation-of-a-thought-process-part-i-3c88b7588695 (accessed Aug. 16, 2020).
[26] ‘Knowledge Graph — Ryte Wiki - Wiki du marketing digital’. https://fr.ryte.com/wiki/Knowledge_Graph (accessed Aug. 16, 2020).
[27] Mariia Rizun, ‘Knowledge Graph Application in Education: a Literature Review’, *Wydawnictwo Uniwersytetu Łódzkiego*, Acta Universitatis Lodziensis Folia oeconomica, pp. 8–12, Aug. 2019.
[28] ‘Knowledge graph google : j’ai pas compris’, *Blog Axe-net.fr*. https://blog.axe-net.fr/knowledge-graph-google-pas-compris/ (accessed Aug. 16, 2020).
[29] ‘Knowledge Graph part-2 : Modelling tabular data as graph – ConfusedCoders’. https://confusedcoders.com/data-engineering/knowledge-graph-part-2-modelling-tabular-data-as-graph (accessed Aug. 16, 2020).
[30] N. Sharma, ‘Knowledge Graph part-3 :Building REST API over Knowledge Graph – ConfusedCoders’. https://confusedcoders.com/data-science/knowledge-graph-part-3-building-rest-api-over-knowledge-graph (accessed Aug. 16, 2020).
[31] Heiko Paulheim, ‘Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods | www.semantic-web-journal.net’, pp. 1–10, Sep. 15, 2015.
[32] ‘Knowledge Graph Tutorial’. http://knowledgegraph.info/ (accessed Aug. 16, 2020).
[33] ‘Knowledge Graph-WorldLife blog’. https://wordlift.io/blog/en/entity/knowledge-graph/ (accessed Aug. 10, 2020).
[34] ‘Knowledge Graph: what is it and why is it important?’, *ContentKing*. https://www.contentkingapp.com/blog/knowledge-graph/ (accessed Aug. 16, 2020).
[35] ‘Link predection Algorithms’. http://be.amazd.com/link-prediction/?fbclid=IwAR1LM55Wd60wmSsSavFdW-3IA8Pjg5uaX1mTpkXqQFPgF8aTSPfboEBPFIU (accessed Aug. 05, 2020).
[36] Mark Needham, ‘Link Prediction with Neo4j Part 1: An Introduction’, Mar. 08, 2019. https://medium.com/neo4j/link-prediction-with-neo4j-part-1-an-introduction-

713aa779fd9 (accessed Apr. 29, 2020).
[37] JJ, ‘MAE and RMSE — Which Metric is Better?’, *Medium*, Mar. 23, 2016. https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d (accessed Aug. 16, 2020).
[38] ‘Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)’. http://www.eumetrain.org/data/4/451/english/msg/ver_cont_var/uos3/uos3_ko1.htm (accessed Aug. 16, 2020).
[39] ‘Mean Absolute Error vs Root-Mean Square Error’, *BrainsToBytes*, Dec. 17, 2019. https://www.brainstobytes.com/mean-absolute-error-vs-root-mean-square-error/ (accessed Aug. 18, 2020).
[40] ‘Model: Relational to Graph- Neo4j Graph Database Platform’. https://neo4j.com/developer/relational-to-graph-modeling/ (accessed Apr. 25, 2020).
[41] ‘Neo4j Graph Platform- Neo4j Graph Database Platform’. https://neo4j.com/developer/graph-platform/ (accessed Jul. 22, 2020).
[42] C. Allen, I. Balazevic, and T. M. Hospedales, ‘On Understanding Knowledge Graph Representation’, *arXiv:1909.11611 [cs, stat]*, Sep. 2019, Accessed: Aug. 16, 2020. [Online]. Available: http://arxiv.org/abs/1909.11611.
[43] Shivani Rawal, Dr. Shuchi Gupta, and Shekhar Singh, ‘Predicting Missing Values in a Dataset : Challenges and Approaches’, *2017*, pp. 34–37, Sep. 2017.
[44] ‘Querying Large Knowledge Graphs over Triple Pattern Fragments: An Empirical Study’, in *The Semantic Web - ISWC 2018*, Denny Vrandečić Google San Francisco, CA USA, and All., 17th International Semantic Web Conference, Monterey, CA, USA, 2018, pp. 125–134.
[45] Michael Hunger, Ryan Boyd, and William Lyon, ‘Relational Database vs. Graph Database Model |Neo4j’, Feb. 29, 2016. https://neo4j.com/blog/rdbms-vs-graph-data-modeling/ (accessed Apr. 25, 2020).
[46] Neena A C, ‘SEO Efforts to Get Listed in Google Knowledge Graph’, Mar. 2015. https://www.techwyse.com/blog/search-engine-optimization/seo-efforts-to-get-listed-in-google-knowledge-graph/?b=1 (accessed Dec. 29, 2019).
[47] *SmartDataAnalytics/Knowledge-Graph-Analysis-Programming-Exercises*. Smart Data Analytics, 2020.
[48] Liyan Dong, Yongli Li, Yongli Li, Huang Le, and Mao Rui1, ‘The Algorithm of Link Prediction on Social Network’, pp. 1–6, Sep. 2013.
[49] ‘The prevention and handling of the missing data’, May 24, 2013. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/#B12 (accessed Feb. 23, 2020).
[50] Denny Vrandečić, Kalina Bontcheva, and and all, *The Semantic Web – ISWC 2018*, Denny Vrandečić Google San Francisco, CA USA, and All. 17th International Semantic Web Conference, Monterey, CA, USA, 2018.
[51] Lisa Ehrlinger and Wolfram Wöß, ‘Towards a Definition of Knowledge Graphs’, Johannes Kepler University Linz, Austria, pp. 1–3, 2016.
[52] B. Simard, ‘Tutoriel sur une introduction à Neo4j, une base de données orientée graphe’, *Developpez.com*. http://logisima.developpez.com/tutoriel/nosql/neo4j/introduction-neo4j/ (accessed Aug. 16, 2020).
[53] B. Slawski, ‘User-Specific Knowledge Graphs to Support Queries and Predictions’, *SEO by the Sea* ⚓, Nov. 25, 2019. https://www.seobythesea.com/2019/11/user-specific-knowledge-graphs/ (accessed Aug. 16, 2020).
[54] ‘What is a Graph Database? -Neo4j Graph Database Platform’. https://neo4j.com/developer/graph-database/ (accessed Jul. 22, 2020).
[55] ‘What is a Knowledge Graph? |Ontotext’.

https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/#:~:text=The%20knowledge%20graph%20(KG)%20represents,situations%20or%20abstract%20concepts%20%E2%80%93%20where%3A&text=Entity%20descriptions%20contribute%20to%20one,the%20entities%2C%20related%20to%20it. (accessed May 30, 2020).
[56] Maxim Kolchin, 'What is a Semantic Web Knowledge Graph? The main building blocks', Jul. 27, 2018. https://medium.com/datafabric/what-is-a-semantic-web-knowledge-graph-82078ea481bc (accessed Dec. 24, 2019).
[57] PRATEEK JOSHI, 'What is Knowldge graph | Build a Knowledge Graph from Text Data', Oct. 14, 2019. https://www.analyticsvidhya.com/blog/2019/10/how-to-build-knowledge-graph-text-using-spacy/ (accessed Apr. 30, 2020).
[58] T. Bishop, 'Will Google Knowledge Graph Steal Your Website Traffic?', *Ezoic*, Nov. 06, 2018. https://www.ezoic.com/will-google-knowledge-graph-steal-your-website-traffic/ (accessed Aug. 16, 2020).
[59] J. S. on, 'WTF is a knowledge graph? | Hacker Noon'. https://hackernoon.com/wtf-is-a-knowledge-graph-a16603a1a25f (accessed Aug. 16, 2020).
[60] https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e
[61] 'DBpedia version 2016-10 | DBpedia'. https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10 (accessed Sep. 02, 2020).
[62] J. Z. Pan, G. Vetere, J. M. Gomez-Perez, and H. Wu, Eds., *Exploiting Linked Data and Knowledge Graphs in Large Organisations*. Springer International Publishing, 2017.
[63] Singhal A, 'Introducing the Knowledge Graph: things, not strings', *Google*, May 16, 2012. https://blog.google/products/search/introducing-knowledge-graph-things-not/ (accessed Sep. 02, 2020).
[64] 'Wikidata'. https://www.wikidata.org/wiki/Wikidata:Main_Page (accessed Sep. 02, 2020).
[65] 'Wikimedia Commons'. https://commons.wikimedia.org/wiki/Main_Page (accessed Sep. 02, 2020).

الملخص:

بعد انتشار مصطلح الرسم البياني المعرفي في السنوات الأخيرة وتحديد ماهيته وكيفية تمثيله، تم استخدامه من طرف عدة شركات لأغراض مختلفة. كما ان KG له ارتباط ببعض المجالات وهو قيد التطوير.

الهدف من هذه المذكرة التحقق من إمكانية تطبيق الرسم البياني المعرفي في مجال التنقيب عن البيانات من أجل حل مشكلة القيم المفقودة في قاعدة البيانات ليصبح الرسم البياني المعرفي تقنية تنبؤ قابلة للتطوير ويتم استخدامها في هذا المجال.

الكلمات المفتاحية: الرسم البياني المعرفي، القيم المفقودة، تقنيات التنبؤ، قاعدة البيانات.

Abstract:

After the spread of the term knowledge graph in the recent years and the definition of its concept and how to represent it, it was used by many companies for different purposes. As well, KG has relation with some domains and it is under development.

The aim of this dissertation is to verify the possibility of applying the knowledge graph in the data mining domain in order to solve the problem of missing values in the database. Therefore, KG becomes a prediction technique which can be developed and used in this domain.

Key words: Knowledge graph, missing values, prediction techniques, Database.

Résumé :

Après la diffusion du terme graphe de connaissances ces dernières années et la définition de son concept et comment le représenter, il a été utilisé par nombreuses entreprises pour des différents objectifs. De plus, KG présente des liens étroits avec certains domaines et il est actuellement un champ de recherche et de développement.

Le but de ce mémoire est la vérification de la possibilité d'appliquer les graphes de connaissance dans le domaine de data mining afin de résoudre le problème des valeurs manquantes dans les bases de données. Par conséquent, il y a lieu de proposer une technique de prédiction des valeurs manquantes qui peut être développée et utilisée dans ce domaine.

Mots clés : graphe de connaissances, valeurs manquantes, techniques de prédiction, base de données.