Literature Review of Course Analysis Using Knowledge Graph

Author Name:

Date:

Contents

1. Abstract..... 3

2. Introduction..... 4

3. Knowledge Graphs (KG)..... 7

4. Text extraction tools..... 11

4.1 Textract..... 11

4.2 Apache Tika..... 12

4.3 PyPDF2..... 13

4.4 NLTK (Natural Language Toolkit)..... 14

4.5 OCRopus..... 15

4.6 Beautiful Soup..... 16

4.7 Challenges and Limitations of text extraction and possible solution..... 16

5. Literature review..... 18

6. Discussion..... 24

5.1 Methods of Gathering Text from Various Formats..... 25

5.2 Best Practices for Constructing and Maintaining a Knowledge Graph..... 26

7. Conclusion..... 28

7.1 Potential of Study..... 29

References..... 31

1. Abstract

Knowledge graphs have increased in popularity in different fields, including education. Applying knowledge graphs for analyzing and comparing different courses is a promising area of research. The paper provides a review of past literary works that are recent and relevant relating to the use of knowledge graphs in the analysis of and comparison of course. The study will discuss the main findings, techniques used in constructing knowledge graphs, types of data sources used in the graphs, and metrics used in the comparison. The review reveals most studies use structured or semi-structured data sources, including catalogs, syllabi, and learning outcomes, to create the knowledge graphs. Other studies have also integrated unstructured, for instance, student feedback and discussion forums, in their graphs.

Regarding techniques employed to construct knowledge graphs, most studies used rule-based or machine learning-based techniques, while other studies have combined the two techniques. The metrics used in course comparison differ from the studies. However, the most common measures are similarity, diversity, and coverage.

Several advantages have been identified to be associated with using knowledge graphs in the analysis and comparison of different courses; these include identifying associations and dependencies between different courses to determine gaps in course offerings and recommend courses to its students based on their interests and past performance. However, knowledge graphs pose several challenges, including data standardization and cleaning requirements, identifying relevant ontologies and vocabularies, and developing efficient querying and visualization tools.

In general, literature analysis provides useful information on the knowledge graphs in analyzing and comparing various courses and the need for further research on the best approach to extract the information required to form the knowledge graphs and the best approach to constructing the knowledge graphs for course analysis and comparison. The paper concedes by providing areas for future research, including integrating social network analysis and natural language processing methods into constructing the knowledge graphs and developing customized course recommendation systems according to basic graph approaches.

Keywords: Knowledge base, Course analysis, Knowledge graph, Content analysis, Text extraction, Natural language processing, Course recommendation system

2. Introduction

In recent years, the field of education has experienced a rapid transformation, especially with the introduction of online learning platforms and the tremendous increase in information. The huge volume of information and the need to customize learning experiences according to individuals' preferences has created an increasing challenge for students and instructors to navigate the huge volume and diverse educational content. The challenge is even worsened by the fact that the courses are differently designed and delivered, which increases the difficulty in effectively comparing them. Knowledge graphs (KG) provide a potential solution to such a challenge by providing a unified knowledge representation which is crucial in comparing the courses.

A knowledge Graph-based data structure used to represent information and the associations between different concepts in a particular domain, this tool has increased in

popularity across various sectors besides education, including finance, healthcare, and e-commerce. Education knowledge graphs represent the associations between concepts, courses, and learning outcomes. Knowledge graphs allow students and instructors to determine the relationships between different concepts quickly and easily, enabling them to compare and contrast the different courses effectively.

Recent research studies have shown the potential associated with knowledge graphs in the analysis and comparison of courses. Researchers such as Buchgeher et al. (2021) have conducted a study comparing computer science courses provided by different universities. The results of this study demonstrated that knowledge graphs could effectively capture a course's structure and content, enabling individuals to identify similarities and differences between different courses. Another researcher, Abu-Salih (2021), utilized a knowledge graph to analyze associations between concepts in different programming languages. The result of the study demonstrated that knowledge graphs could help identify conceptual gaps between courses; this enables instructors to improve the design and delivery of various courses.

The construction of knowledge graphs for carrying out analysis of different courses has the potential to affect the education sector in an appositive way greatly. Integrating data from multiple sources, including student enrollment data, learning outcomes, and syllabi, knowledge graphs offer a more wholesome and nuanced view of the education landscape, resulting in several benefits. The first use of knowledge graphs in the education sector can help achieve better student outcomes; this is made possible through the use of knowledge graphs to analyze the way the courses are related, identify patterns that exist between the course, and determine how students are performing is useful for identifying the courses which are most appropriate for

achieving specific learning outcomes, this, in turn, helps instructors to dessin curricula that is more effective improving the student outcome. For instance, they can modify the course content or structure, the course sequence, or include new methods of teaching that are better aligned with the required learning outcomes (Chen et al., 2021).

Also, the use of knowledge graphs has the potential to result in improved student engagement; through analysis of how the courses are associated, the information on student enrollment and interest of students, instructors can design customized learning experiences that are more engaging, they can change the course content or structure to make sure that its s better aligned with the interests of a specific group of students, or institutions can also create new course specially meant to satisfy the interests and needs of specific groups of students, hence improving student engagement, motivation and satisfaction with their learning experiences which also contributes to better learning outcomes (Chen et a., 2018).

Finally, knowledge graphs are also essential in the education sector as they help improve the allocation of resources. Institutions can incorporate details regarding how resources are allocated in a knowledge graph and identify the resources that are most effectively used and the ones that are being underutilized. By integrating information on student enrollment, how the course is associated with each other, and details on resource allocation, institutions can identify how to optimize how different resources, including classrooms, faculty, and equipment, are allocated (Chen et al., 2018). Hence helping institutions save costs and enhance efficiency through giving resources to courses and programs with the highest demand and ensuring that the available resources are utilized most effectively and efficiently.

However, despite the potential of knowledge graphs in information analysis which researchers have demonstrated, further research is required to explore the effectiveness of knowledge graphs compared to a different course. Particularly further research should be conducted to examine the most effective approach that can be used to develop the most practical knowledge graph. In addition, there is a need to look at how knowledge graphs can be utilized in the different courses. This research study aims to conduct a literature review using knowledge graphs to analyze and compare different courses. The study is built on the most relevant and current research studies related to the research domain. The study also analyzes the different tools used for gathering the texts from the various sources and their effectiveness and accuracy for the different types of resources. The study also aims to identify the best practices available for the organization and maintenance of the knowledge graph regarding comparing and analyzing different courses.

The paper begins with an overview of the knowledge graph, its properties, and applications across the domains. A comprehensive review of past literature on using knowledge graphs in education will be provided, mainly concentrating on the analysis and comparison of courses. The literature review will also include a critical analysis of the research methodologies, data sources, and data analysis techniques used.

3. Knowledge Graphs (KG)

A structured information representation known as a "knowledge graph" is used to model the associations and connections between various things within a given domain. This graph-based knowledge representation uses nodes and edges to represent knowledge (Fensel et al., 2020). The edges show how these entities are connected, while the node represents the

entities, which can be things, occasions, or concepts. The properties of a knowledge graph include expressiveness, flexibility, and scalability. Knowledge graphs can be expanded to accommodate large data volumes and express relationships between entities. A knowledge graph is flexible since it can be updated and modified easily in response to new knowledge. It is a structured knowledge base representing knowledge as a triple of the form (h,r,t), with the ***h*** representing the head entity, the ***t*** representing the tail entity, and the ***r*** representing the associations between ***h*** and ***t.*** An instance of a knowledge graph is illustrated in Figure 1 and Figure 2 with lectures and course-represented entities. The triple (Prof S.B, taught_2_2018, CS311) is interpreted as Prof. S.B teaching CS311 in the second semester of the 2018 session. In the same way (CS121, prerequisite, CS211), CS121 is interpreted as the prerequisite of CS211.

Importance of Knowledge Graphs in different domains

Different fields have implemented knowledge graphs to help them in several ways. For instance, the healthcare sector uses knowledge graphs to enhance patient care by providing customized treatment recommendations based on a patient's medical history and symptoms (Pereira et al., 2021); these knowledge graphs assist healthcare providers by enhancing patient care through integration and analysis of data from different sources including clinical trials, electronic records, and media literacy. Knowledge graphs enable more accurate patient diagnosis by combining information regarding patients and knowledge in the field of medicine; they can also have a better option for treatment and patient care plans that are more personalized. For instance, physicians can utilize knowledge graphs to identify how drugs interact with patients and their side effects and propose personalized healthcare plans based on an individual's medical history.

Second, the finance sector uses knowledge graphs to perform analysis of financial data and identify suspicious activities suspected of being fraudulent (Abu-Salih, 2021). Through the integration of data obtained from different sources, knowledge graphs are able to help fraud detection as these data, including records of transactions and customer profiles, are analyzed for patterns and hidden associations between them that reveal suspicious activities indicating possible fraudulent activities, helping prevent financial losses. For instance, banking institutions utilize knowledge graphs and connect them to several other connected customer profiles, transactions, and external databases that have details of identified fraudsters so as to help prevent possible fraud.

Third, the manufacturing sector also uses knowledge graphs to enhance efficiency and productivity through integration and analysis of data obtained from different sources, including sensor data, production records, and logs of maintenance; knowledge graphs can help in connecting data points and identification of hidden trends to help in optimizing the production process, minimize downtime and avoid failures in equipment. For instance, the company can use knowledge graphs to determine the root cause that results in the breakdown of machines by establishing a connection between maintenance records, sensor records, and production records (Xu et al., 2020).

Knowledge graphs are also applied in the agriculture sector; they help in crop yield optimizations and waste reduction through the integration and analysis of data obtained from various sources, including weather data, soil data, and crop models. Knowledge graphs can be used to help in pattern identification and associations between different factors to provide the optimum number of pesticides and other chemicals to use to enhance efficiency and resource allocation within the farm (Drury et al., 2019). Currently, several farms utilize knowledge graphs to help them determine the most suitable time and location for plating different types of crops based on the conditions of the soil, crop models, weather, and soil conditions.

In summary, knowledge graphs have the potential for application in a wide range of industries and domains as it assists in inferring meaning in complex data, extracting useful insights, and improvement of different processes, products, and services. The ability to support complex reasoning tasks is one of the key advantages of knowledge graphs. They can be used to derive new knowledge from current data and offer solutions to challenging questions that are challenging for conventional databases to handle. Also, people have the ability to combine data

from several sources, which aids in the discovery of novel relationships and insights that would otherwise be challenging to find.

4. Text extraction tools

4.1 Textract

The document enables ingestion of source documents in multiple formats, including PDF and Word documents (doc). The document model provides an abstraction layer between the character-based document stream and annotation-based document components. The individual plugin components provide linguistic analysis capabilities; these share annotation repository, Lexi cache, and vocabulary and interact with each other by posing the results and reading prior analyses from them. Plugins share a common interface and are controlled by a plugin manager according to declared dependencies among plugins; the resources manager controls shared resources such as lexicons, glossaries, or gazetteers (Hegghammer, 2022); this tool is illustrated in Figure 1.

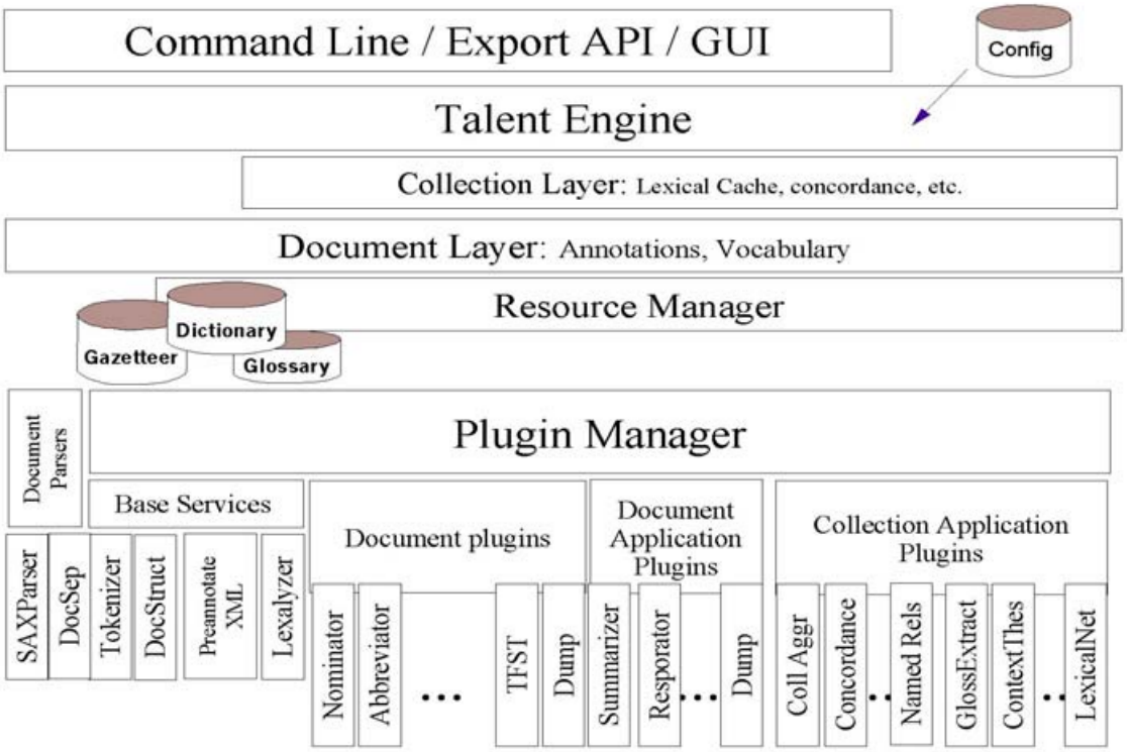

Figure 1: The talent system: Textract Architecture and data model by M. Neff, 2017

4.2 Apache Tika

The parser allows for the intake of documents in multiple formats. It is responsible for analyzing the document and producing a stream of content and metadata for the file. The detector element identifies the file format of the input file, and this element determines the appropriate parser to be used depending on the input file. The language identifier element is used for language detection of the input file; the Named Entity Recognition module identifies named entities, including places, people, and organizations contained in the input file. The Metadata Extractor element is responsible for metadata extraction from the input file, including date, author, and file formats. Finally, the content extractor element is responsible for content extraction from the input file, including images and texts, among other media (Méndez et al., 2021). Figure 2 gives an illustration of the tool.

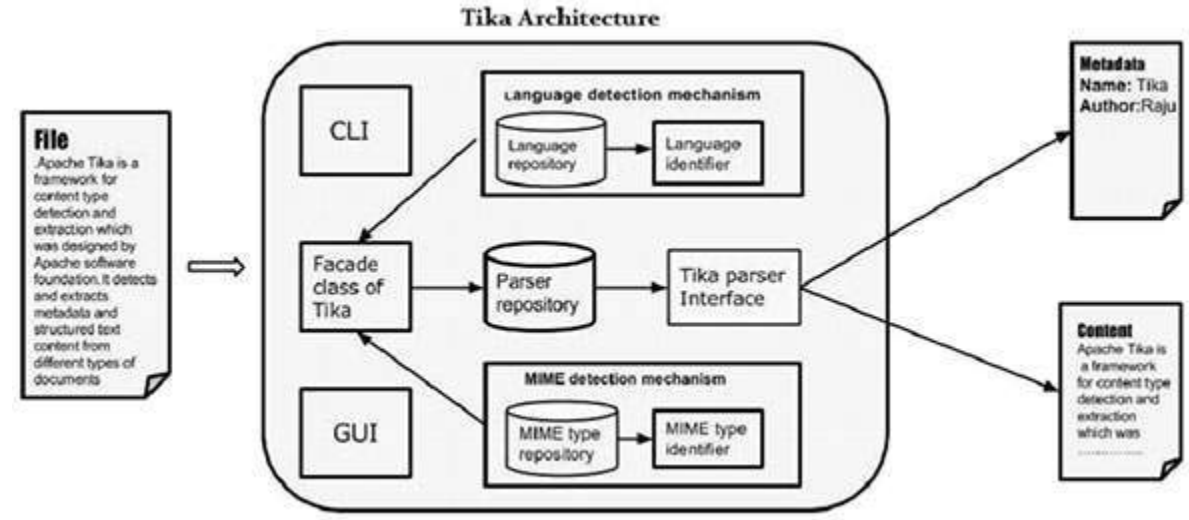

Figure 2: Apache Tika architecture by Tutorialspoint, 2023

4.3 PyPDF2

The tool begins by first analyzing the PDF file using the PdfFileReader class, enabling reading and text attraction from the files. The class takes in the file as input and produces a PdfFileReader object which details the file, including its page count, bookmarks, and metadata. The PageObject is a representation of a single page of a PDF file. It contains methods that facilitate the extraction of images and texts, among other data, from a single page. Finally, the PdfFileWriter enables the creation of new PDF files or modifications, and this contains methods that support the addition of pages, metadata, and bookmarks, among other elements (Hitha & Kiran, 2021); this is illustrated in Figure 3.

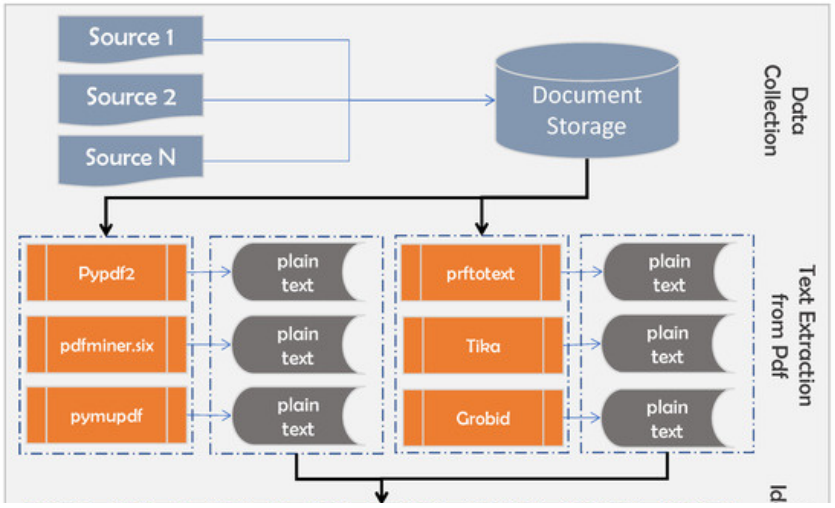

Figure 3: PyPDF2 tutorial architecture by Dhanashree 2023

4.4 NLTK (Natural Language Toolkit)

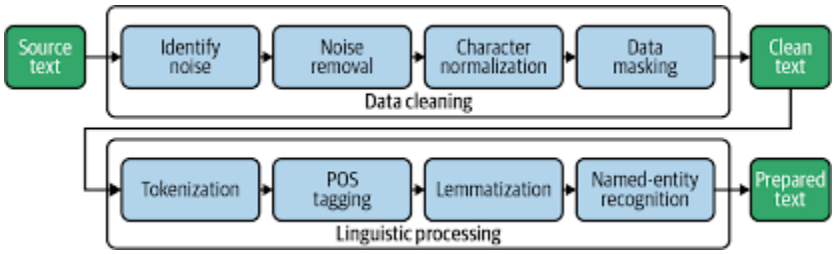

Figure 4: NLTK architecture by Singh, 2023

This is a Python library that is popular for tasks involving natural language processing, which also includes extracting texts from documents, as illustrated in Figure 4. It provides different tools that assist in different tasks, which include stemming, tokenization, and part of speech tagging, just to mention a few. The tool operates by first loading the data, either from a file or by storing it in a string variable. The NLTK tokenization module then tokenizes the data into words or sentences, which involves breaking down the text into smaller units and allowing for further analysis (Yogish et al., 2019). The next stage is part-of-speech tagging (POS), which is carried out by the POS tagging module, and this involves the identification of the grammatical parts of words in the file and assigning a tag based on the context. The next step is named entity

recognition (NER), which involves the identification and classification of named entities in the text; this process is performed by a NER module that extracts the named entities from the file (Yogish et al., 2019).

4.5 OCRopus

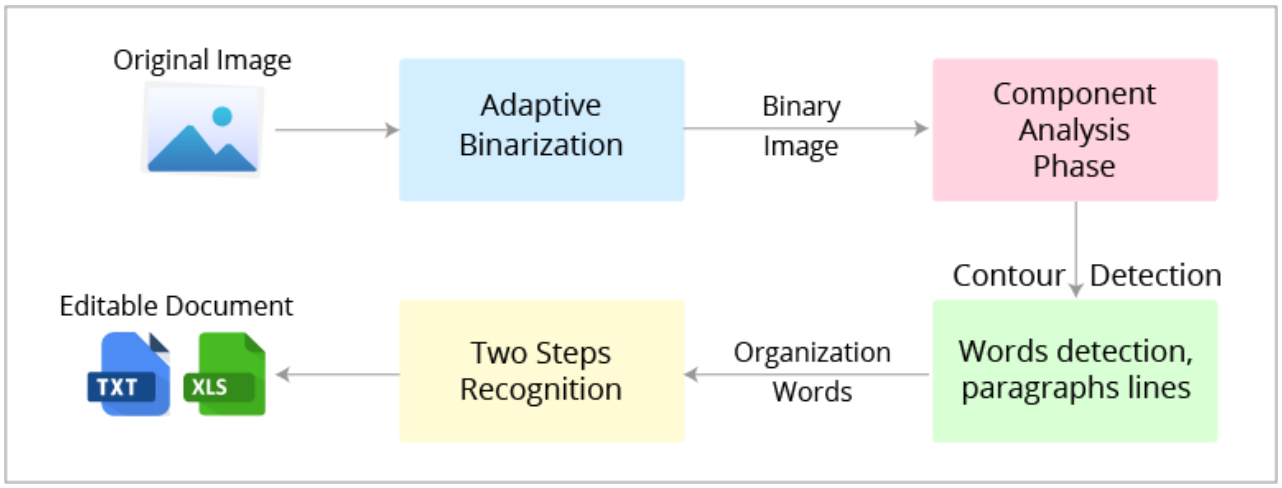

Figure 5: OCRopus architecture by Singh, 2023

This tool leverages optical character recognition capabilities and machine learning algorithms for text recognition from scanned documents, as illustrated in Figure 5. It is an open-source system consisting of several modules designed for different tasks, including layout analysis, image processing, and character recognition (Jain et al., 2021). The tool accepts the scanned image data and has OCRopus tool preprocessing capabilities that improve the image quality, after which the OCRopus binarization module binarizes the preprocessed image, transforming it into black and white pixels (Jain et al., 2021). The next step is layout analysis, which involves segmenting the image into regions that might include text lines, words, and characters. This is done by the layout analysis module. Finally, the text recognition module recognizes the characters contained in each region and outputs the corresponding text.

4.6 Beautiful Soup

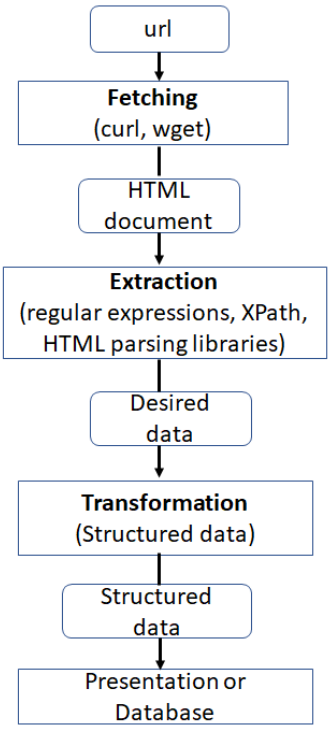

Figure 6: Beautiful soup architecture by Khder, 2021

This is a Python library tool designed for scraping and data extraction from different files, especially HTML and XML files, as illustrated in Figure 6. Through various methods, it can read and identify specific texts within HTML and XML files. The tool works by first loading the HTML file from which the data needs to be extracted, and then, with the aid of the parser module, it conducts data extraction by parsing the HTML data (Khder, 2021)). After extraction, the tool makes use of its navigational techniques to enable browsing through the HTML and locating particular terms.

4.7 Challenges and Limitations of text extraction and possible solution

Unstructured data is one of the challenges in data extraction; this refers to data that lacks organization in a predefined structured format; the main issue with such kind of data is that they

lack a fixed structure making extraction of useful insights a huge challenge (Acheampong et al., 2020). The extraction process requires a significantly huge amount of processing capabilities and resources, and conventional extraction techniques become ineffective. However, tools that integrate natural language processing(NLP) and machine learning capabilities can be used to solve this challenge; these include the use of a tool such as Apache Tika, which is useful in extracting structured information from unstructured data.

The other challenge is the inability to process large data volumes. The lack of ability to process big data volumes is the other problem. Due to the alarming rate at which data are being produced, their existence in various formats, and their generation from numerous sources, traditional methods for processing data are ineffective when managing such complex data, making data extraction challenging (Machado & Pardo, 2022) since they possess the capability to perform analysis on enormous data volumes or cloud-based alternatives, including Amazon Web Services (AWS).

Data quality issues are another key challenge in data extraction. Data quality should be considered an important component in the extraction process since incorrect, inaccuracies, and inconsistencies in data can result in inaccurate results, compromising the effectiveness of the process (Kobayashi et al., 2018). These issues can result from errors during data entry, issues in data integration, or human errors. However, this challenge can be solved through data cleaning and validation procedures. This technique involves error identification and correction as well as fixing data inconsistencies, while validation involves ensuring precision in data and consistency in formats and values.

5. Literature review

Recent years have seen a growth in interest concerning knowledge graphs for analysis and comparisons of courses, and this entails modeling the course as entities and presenting their associations with other entities, including learning objectives, perquisites, and assessment criteria. The use of knowledge graphs in course analysis and comparisons enable instructors and learners to better comprehend the content and structure of the different courses and their suitability for the different learners based on their past performance and interest. Knowledge graphs are simply graph databases containing a collection of nodes referred to as entities and edges referred to as relationships that reveal how the different entities are related. Each node represents a piece of information. On the other hand, the relationship is represented using the edges, making the knowledge graphs a tool suitable for course analysis and compassion.

The ability of knowledge graphs to present complex associations between entities in a manner that is easy to comprehend and analyze is considered the key benefit of knowledge graphs. Knowledge graphs can represent associations between courses, students, professors, and coursebooks, among other relevant information (Yu et al., 2021). The organization of this information in structured graphs enables researchers to easily visualize how different entities are associated with each other, which is useful in the identification of patterns and insights which might otherwise not be immediately identified.

One of the most crucial parts of constructing a knowledge graph is the extraction of text from different file formats. According to Hegghammer (2022), one tool that is suitable to text extraction is Textract which is a Python library that has been gaining popularity over the years, and this is because it is able to extract text from various file formats, these include Word

documents, PDF documents, and JPEG images. Hegghammer (2022) analyzed OCR-based tools to extract text from PDF files, and the study involved sentiment analysis on customer reviews. According to the study, the tool was able to efficiently extract text from PDF files in regards to processing time. However, the accuracy of the tool was affected by complex file structures.

Another tool proposed by Oussous & Benjelloun(2022) is Apache Tika which is an open-source toolkit for analyzing and detecting content that is capable of extracting text and metadata from different document formats; these formats include PDF, word documents (doc), PowerPoint files (ppt) and Hyper Text Markup Language documents (HTML). According to Couscous & Benjelloun(2022), the tool was able to accurately extract texts from PDF files despite their complex structure. However, the processing time was longer compared to other tools, such as PyDF2.

PyPDF2 is another proposed by Hitha & Kiran (2021); this is a Python library for handling PDF files which also includes extraction of text and metadata elements. Hitha & Kiran (2021) used the tool to extract text from PDF files in their research study that involved analysis of content recognition techniques and content analysis of articles in computer science. According to the study, the tool accurately extracted text from the PDF files and had an efficient processing time. However, the accuracy was affected by complex file structures resulting in errors occurring in the extracted text.

Issues of data integration are one of the main challenges in knowledge graph construction, as these data have different formats and structures. According to Zhang et al. (2020), the best practice for overcoming this challenge is the identification of relevant day sources, standardization of the data formats and metadata, and resolving semantic heterogeneity

through ontological and controlled vocabularies. In their study, Zhang et al. (2020) developed a domain-specific ontology to help them interrogate the data obtained from multiple sources in the healthcare sector.

Another important aspect of knowledge graphs is the entities and their associations. According to Yue et al. (2021), the best practices involve the identification of entities that are relevant and identifying their associations, extracting them, and representing them using standardized formats such as the Resources Description Framework (RDF). Yue et al. (2021) successfully used the named entity recognition and relation extraction approach to construct knowledge graphs for the Chinese Forestry sector. The model framework for the construction of the forestry knowledge graph is provided in Appendix A.

In order to ensure the accuracy and reliability of the knowledge graph, the quality of data used in the construction of the graph should be carefully considered. According to Filgueira et al. (2019), the data should be cleaned, validated, and normalized to remove errors and inconsistencies. He proposed a quality control framework for the construction of knowledge graphs, and this framework includes validation rules and error reporting mechanisms.

A study was carried out by Chen et al. (2018) to perform a comparison of an online course using knowledge graphs. The research utilized knowledge graphs to analyze and compare the content and structure of online courses hosted on multiple platforms. The study's findings revealed that knowledge graphs effectively identify the similarities and differences between the courses. It is helpful in pinpointing the areas where the course content needs to be changed. The following details about the course were provided: the course's name, its description, the name of the course instructor, and the material covered in the course. The research involved the building

of a knowledge graph. Then, based on the courses' structure and content, a clustering technique was applied to group related courses. Also, according to Dang et al. (2019), the knowledge graphs enable the identification of similar courses across different platforms, even where the courses have different names and descriptions. In regard to research studies, the knowledge graphs identified areas that require improvement concerning course content. The researchers discovered that some courses were significantly overlapping regarding content. In contrast, others had gaps in their coverage of specific topics; such information is crucial for improving course design and delivery, resulting in better student outcomes.

Another study carried out by Guo et al. (2020) used a knowledge graph to evaluate the effectiveness of various instructional strategies in online courses. The study's findings revealed that knowledge graphs are essential in identifying patterns in students' behavior and performance. The study involved creating knowledge graphs that included course information. The researchers also included the student's information, including their demographic information performance in the course. The study was successful in identifying patterns in the behavior of students and their performance. For instance, they were able to identify specific instructional strategies for the engagement and retention of students. Similar to the first study, these research findings revealed that the knowledge graphs could be utilized to identify areas that required course design and delivery improvement. The study discovered that specific instructional strategies impacted specific types of students more. The information is crucial in customizing courses to meet the needs of different learners more effectively.

Zou (2021) conducted a study comparing an online course using the Knowledge Graph approach. The study mainly used a knowledge graph to compare online courses provided by

different instructions. The study's findings revealed that the knowledge graph was essential in identifying courses offered by different intuitions that discussed similar topics. According to the research, the knowledge graph also helped the researchers to determine the areas where certain institutions had unique offerings or strengths. As such, the researchers discovered that some institutions offered courses that concentrated mostly on practical applications, while others mainly concentrated on theoretical concepts. The study's findings enabled the researchers to conclude that knowledge graphs are important in comparing online courses provided by different institutions. Despite the success of the topic, the study also revealed some challenges that need to be addressed, including a lack of standardized data formats and the possibility of bias in user ratings.

A study by Hogan et al. (2021) focused on using knowledge graphs to give students useful recommendations concerning the different courses. Like the other studies, the researchers developed a knowledge graph that included information on the course, including the course's name, description, learning outcomes, and perquisites. The study also included the students' academic history, such as previous courses and performance. According to the study's findings, the knowledge graph was useful in generating customized course recommendations for the students considering their academic history and interests. According to the research, the knowledge graphs provided more accurate and precise course recommendations than conventional recommendation systems.

The research findings also revealed that the knowledge graph was useful in determining the gaps in the academic history of students and providing suggestions on the courses to mitigate such gaps; this is especially essential for students who want to change majors or pursue

interdisciplinary studies. The study's findings enabled the researchers to conclude that the knowledge gaps were particularly important in improving the effectiveness of course recommendations and providing students with a better-customized guide. Similar to the last research, this study covered challenges faced in constructing a knowledge gap. The challenges include the lack of standardized data and possible bias recommendations. The findings also revealed that the knowledge graph could be used in the identification of trends in course content over some time, as such the researchers discovered that course associated with artificial intelligence and machine learning has increased in popularity over the years; thus, information was a useful for institutions as they try to ensure that their course offerings remain up to date and relevant.

There are several metrics that can be utilized in assessing the effectiveness of knowledge graphs, and these include precision, recall, F1-score, and mean reciprocal rank (MRR). First, precision assesses the accuracy of the knowledge graph by analyzing the proportion of relevant results among all the retrieved results. The recall metrics are used to analyze the completeness of the data provided, and they do this by measuring the fraction of relevant results retrieved by the knowledge graph out of all the possible relevant results (Alagarsamy et al., 2022). F1-score is a metric that is a combination of recall and a single score, and it evaluates the overall performance of the knowledge graph. The Mean Reciprocal Rank (MRR), which is a statistic, calculates the average of the reciprocal ranks of the first relevant outcomes during the process to assess the quality of the ranking system (Alagarsamy et al., 2022).

However, due to the nature of the data, the measurements do have some restrictions. Precision may be effective where the aim is to retrieve all the relevant results since it also

considers the results that have been retrieved. On the other hand, recall might be ineffective where the aim is the identification of the most relevant results since it can also assess the relevant results, and finally, the F1-score might be ineffective where the importance of precision and recall varies for different data.

6. Discussion

The various literary works that have been reviewed have highlighted the potential associated with knowledge graphs to enhance course design and delivery while also facilitating the comparison of courses provided by different institutions. In summary, all the past literature works that have been reviewed have demonstrated that knowledge graphs are crucial in analyzing courses to identify strengths and areas that can be improved. However, despite the usefulness of knowledge graphs, some studies identified challenges involved in constructing effective knowledge graphs and further emphasized that these challenges should be addressed. The lack of standardized data format and ontologies for educational information is one of the key challenges faced by the researchers in the construction of the knowledge graph, thus making it difficult to develop knowledge graphs that can be shared with ease and compared across multiple platforms and institutions. The other key challenge is the lack of sufficient research on the effectiveness of knowledge graphs in education. Despite the reviewed literature providing information regarding knowledge graphs in education, there is still a need for further research to examine the effectiveness of graphs in enhancing course design and privacy.

Also, as demonstrated by the review of the various past literary works, it is clear that knowledge graphs are increasing in popularity as the ideal tool for performing analysis and comparison of courses; this is because they are able to integrate a wide variety of data sources

and develop meaningful associations between different entities. From the literary works, we can deduce the methods involved in gathering texts from various areas used as inputs for the knowledge graph, the current best practices involved in the construction and maintenance of knowledge graphs as well as the entire process involved in constructing knowledge graphs.

5.1 Methods of Gathering Text from Various Formats

It is challenging to gather texts from different file formats, particularly when dealing with complex file structures. There are several tools that can be used to extract text from different file formats. The literature reviews revealed three of the most commonly used tools, these include Textract, Apache Tika, and PyPDF2.

First, Textract is a Python package that is capable of text extraction from various file formats, including Word documents (doc), pdf, and PowerPoint files (ppt). The tool has optical character recognition (OCR) technology, which helps it recognize text that appears in images within documents. This tool's main benefit is that it is straightforward and simple to use. Nonetheless, the intricacy of a file may have an impact on the tool's accuracy. The tool might struggle in situations that have files with complex layouts, fonts, or images.

Second, the Apache Tika tool is a have-based toolkit that is capable of text extraction from various file formats similar to the extract, and these file formats include Word documents (doc), pdf, and PowerPoint documents (ppt). It is equipped with machine learning algorithms that are capable of identifying and extracting texts from files. The Apache Tika tool is considered the most accurate tool for text extraction from different file formats as it has a high success rate and can so handle complex file structures, including files that have several languages and encodings.

Finally, PyPDF2 is a Python library that is capable of text education from only pdf files, and this tool works by parsing the files' structure to identify and extract the texts. It is a reliable tool for the extraction of text from pdf files, and it is relatively easy to use. PyPDF2, however, may struggle with complex pdf files having complex layouts or containing images which are also the same as with the extract tool.

After a comparison of the three tools, we can conclude that Apache Tika is the most accurate tool for text attraction from various file formats; this is because it makes use of machine learning algorithms that enable it to handle files with complex structures and accurately identify texts as shown in Table 2. Even though Textract and PyPDF2 are tools that are reliable, they may struggle when dealing with more complex files. The Textracat tool may have challenges with files that have complex layouts, fonts, or images. On the other hand, PyPDF2 may have challenges with pdf files that have complex structures with multiple pages, images, or non-standard fonts. In summary, the selection of the best tool for text extraction from different file formats depends on the project's specific needs, however as seen from the discussion Apache Tika stands out as the most suitable tool in regard to accuracy when dealing with different file formats, and this means that it is a reliable tool during the construction process of a knowledge graph for analysis and comparison of courses as shown in Figure 6.

5.2 Best Practices for Constructing and Maintaining a Knowledge Graph

Construction and maintenance of knowledge graphs require careful planning, organization, and execution. The literature review of past literature works has enabled us to deduce some of the best practices in regard to the construction and maintenance of knowledge graphs that can be used in the analysis and comparison of different courses.

The description of relationships between these things, which entails determining the connections between the entities and the kinds of links that exist between them, is the following stage after all the relevant entities have been successfully identified. For instance, the course may be connected to the project domain in terms of content, learning objectives, and the institution that offers the course. The definition of these associations is critical for the creation of a knowledge graph that is comprehensive and accurate.

After defining the entities and their relationships, the next step is the creation of a data model that guides the process of constructing the knowledge graphs, and this model should include the attributes and properties relating to each entity and their associations. The data model is crucial as it provides a framework for data structuring and organization.

After completing the data model, the next step is gathering data from various sources, which might include catalogs, syllabi, and textbooks; this collected data should be cleaned and structured in a manner that fits the data model; this cleaning process might involve the identification and removal of duplicate, irrelevant or data that is inaccurate as well as standardization of the data formats to fit the data model.

The next step is the construction of the knowledge graph with the gathered and structured adapt; this process will involve the population of the graph with entities, their properties, and associations based on the adapt model. The accuracy and completeness of the knowledge graphs depend on the quality and completeness of the gathered and structured data. After construction of the knowledge graph, it is important to continuously maintain it, and this process involves the continuous update of the graph with new data, removal of outdated data, and making sure that the associations between the entities remain accurate, maintenance of the knowledge graph is a

continuous process that requires constants monitoring and updates to make sure it is accurate and relevant.

In summary, the construction and maintenance of a knowledge graph used in the analysis and compassion, of course, require careful planning, organization, and execution. The construction process involves first identification of all the necessary entities, the definition of the relationships, the development of a data model, data gathering and structuring, construction of the knowledge graphs, and finally, maintenance of the graph over time. The adherence to the best practices will result in a knowledge graph that is both accurate and comprehensive and effective in the analysis and comparison, of course.

7. Conclusion

In conclusion, knowledge graphs are constantly increasing in popularity in regard to analysis and comparison of course; this is attributed to their ability to provide a comprehensive understanding of the associations between different entities. With the constant increase in the availability of educational data, knowledge graphs offer a promising approach that enables institutions to take advantage of their adapt for more informed decision making. The literature review reveals that several research studies have been carried out to explore the use of knowledge graphs in the analysis and comparison of different courses. According to the literature review, the knowledge graphs are capable of improving course recommendation systems, helping in the development of curriculum, and facilitating competency-based education.

Construction of effective knowledge graphs for analyzing and comparing courses requires careful planning and organization. Hence individuals should make sure that they follow

these practices, which include entity identification, the definition of relationships, the creation of data models, data gathering, the construction of the knowledge graphs, and then the maintenance of the graphs over time. In addition, different tools are available for the extraction of text from different file formats; the research study carried out the compassion of the three most commonly used tools, thee include extract, Apache Tika, and PyPDF2. The study's findings lead us to the conclusion that the Apache Tika tool is the most accurate option because other tools have trouble processing files with intricate structures. The best tool will, however, rely on the particular requirements of the project.

The potential of the study

There are several potential contributions that can be attributed to the findings of this research study. The first is improved course selection. Students can gain useful insights from comparing courses across various institutions or even within their own institutions to enable them to make better-informed decisions concerning the most appropriate courses to select according to their interests and goals. The second potential application of the knowledge gained from this study is the development of curricula. The members of the faculty can utilize this knowledge to identify gaps or overlaps in the courses offered in their institutions, enabling them to arrive at an informed decision concerning the course to include in their program. This information could also help accrediting bodies in assessing the quality and rigor of the various courses and programs, ensuring that they arrive at informed decisions concerning accreditation and entirely contributing to increased equity and access. Career counselors could also find the findings of this study beneficial, as they can gain insights that help them determine the courses

that are aligned with certain career paths and advise students on the areas in which they need to strengthen their skills.

The findings of this research study could have vast implications for the education sector. The first implication is an improvement in the quality of courses since institutions can determine gaps and overlaps in the programs offered and make necessary modifications. This can result in better student satisfaction and an improved reputation for the institutions, as this information can help them enhance the quality of their programs. The findings could also result in greater equity in the education sector by identifying courses that are more beneficial for students who are underrepresented.

Overall, knowledge graphs provide a valuable approach for educational institutions to gain insights from their data and improve their operations. Further research is required for the exploration of the knowledge graph potential for other educationalist cases and to develop best practices for knowledge graph construction and maintenance.

References

Abu-Salih, B. (2021). Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications*, *185*, 103076.

Acheampong, F. A., Wenyu, C., & Nunoo‐Mensah, H. (2020). Text‐based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*(7), e12189.

Alagarsamy, S., James, V., & Raj, R. S. P. (2022). An experimental analysis of optimal hybrid word embedding methods for text classification using a movie review dataset. *Brazilian Archives of Biology and Technology*, *65*.

Aliyu, I., Kana, A. F. D., & Aliyu, S. (2020). Development of knowledge graph for university courses management. International Journal of Education and Management Engineering, 10(2), 1.

Buchgeher, G., Gabauer, D., Martinez-Gil, J., & Ehrlinger, L. (2021). Knowledge graphs in manufacturing and production: A systematic literature review. *IEEE Access*, *9*, 55537-55554.

Chen, Z., Tang, W., Ma, L., & Qian, D. (2021). Research on knowledge graphs in the education field from the perspective of knowledge graphs. *Communications in Computer and Information Science*, 347–360. https://doi.org/10.1007/978-981-16-1160-5_27

Chen, P., Lu, Y., Zheng, V. W., Chen, X., & Yang, B. (2018). Knowedu: A system to construct knowledge graphs for education. *Ieee Access*, *6*, 31553-31563.

Chen, Z., Wang, Y., Zhao, B., Cheng, J., Zhao, X., & Duan, Z. (2020). Knowledge graph completion: A review. *Ieee Access*, *8*, 192435-192456.

Dang, F., Tang, J., & Li, S. (2019, July). MOOC-KG: a MOOC knowledge graph for cross-platform online learning resources. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)* (pp. 1-8). IEEE.

Dhanashree. (2023, March 16). PYPDF2 tutorial - working with PDF in python: Nanonets. Nanonets AI & Machine Learning Blog. Retrieved April 2, 2023, from https://nanonets.com/blog/pypdf2-library-working-with-pdf-files-in-python/

Drury, B., Fernandes, R., Moura, M. F., & de Andrade Lopes, A. (2019). A survey of semantic web technology for agriculture. *Information Processing in Agriculture*, *6*(4), 487-501.

Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., ... & Wahler, A. (2020). Introduction: what is a knowledge graph?. *Knowledge graphs: Methodology, tools and selected use cases*, 1-10.

Filgueira, R., & Garijo, D. (2022, May). Inspect4py: a knowledge extraction framework for python code repositories. In *Proceedings of the 19th International Conference on Mining Software Repositories* (pp. 232-236).

Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., & He, Q. (2020). A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, *34*(8), 3549-3568.

Hegghammer, T. (2022). OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. *Journal of Computational Social Science*, *5*(1), 861-882.

Hitha, K. C., & Kiran, V. K. (2021, November). Topic Recognition and Correlation Analysis of Articles in Computer Science. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 1115-1118). IEEE.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys (CSUR)*, *54*(4), 1-37.

Jain, P., Taneja, K., & Taneja, H. (2021). Which OCR toolset is good and why: A comparative study. *Kuwait Journal of Science*, *48*(2).

Khder, M. A. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing & Its Applications*, *13*(3).

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational research methods*, *21*(3), 733-765.

Machado, M., & Pardo, T. A. S. (2022, June). Evaluating methods for extraction of aspect terms in opinion texts in portuguese-the challenges of implicit aspects. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3819-3828).

Méndez, S. J. R., Omran, P. G., Haller, A., & Taylor, K. (2021). MEL: Metadata Extractor & Loader. In ISWC (Posters/Demos/Industry).

Neff, M. S., Byrd, R. J., & Boguraev, B. (2019). The talent system: Textract Architecture and data model. ACL Anthology. Retrieved April 2, 2023, from https://aclanthology.org/W03-0801/

Ouassous, A., & Benjelloun, F. Z. (2022). A COMPARATIVE STUDY OF DIFFERENT SEARCH AND INDEXING TOOLS FOR BIG DATA. *Jordanian Journal of Computers and Information Technology*, *8*(1).

Pereira de Souza, V., Baroni, R., Choo, C. W., Castro, J. M. D., & Barbosa, R. R. (2021). Knowledge management in health care: an integrative and result-driven clinical staff management model. *Journal of knowledge management*, *25*(5), 1241-1262.

Singh, Y. (2023). *Extracting text from images with tesseract OCR, opencv, and python*. Opcito. Retrieved April 6, 2023, from https://www.opcito.com/blogs/extracting-text-from-images-with-tesseract-ocr-opencv-and-python

Tika - Architecture. Tutorials Point. (n.d.). Retrieved April 2, 2023, from https://www.tutorialspoint.com/tika/tika_architecture.htm

Xu, D., Ruan, C., Korpeoglu, E., Kumar, S., & Achan, K. (2020, January). Product knowledge graph embedding for e-commerce. In *Proceedings of the 13th international conference on web search and data mining* (pp. 672-680).

Yogish, D., Manjunath, T. N., & Hegadi, R. S. (2019). Review on natural language processing trends and techniques using NLTK. In *Recent Trends in Image Processing and Pattern*

Recognition: Second International Conference, RTIP2R 2018, Solapur, India, December 21–22, 2018, Revised Selected Papers, Part III 2 (pp. 589-606). Springer Singapore.

Yue, Q., Li, X., & Li, D. (2021). Chinese relation extraction on forestry knowledge graph construction. Computer Systems Science and Engineering, 37(3), 423-442.

Yu, X., Stahr, M., Chen, H., & Yan, R. (2021, January). Design and implementation of curriculum systems based on knowledge graphs. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)* (pp. 767-770). IEEE.

Zhang, H., Guo, Y., Prosperi, M., & Bian, J. (2020). An ontology-based documentation of data discovery and integration process in cancer outcomes research. *BMC Medical Informatics and Decision Making*, *20*(4), 1-22.

Zou, X. (2020, March). A survey on application of knowledge graph. In *Journal of Physics: Conference Series* (Vol. 1487, No. 1, p. 012016). IOP Publishing.

Appendix A

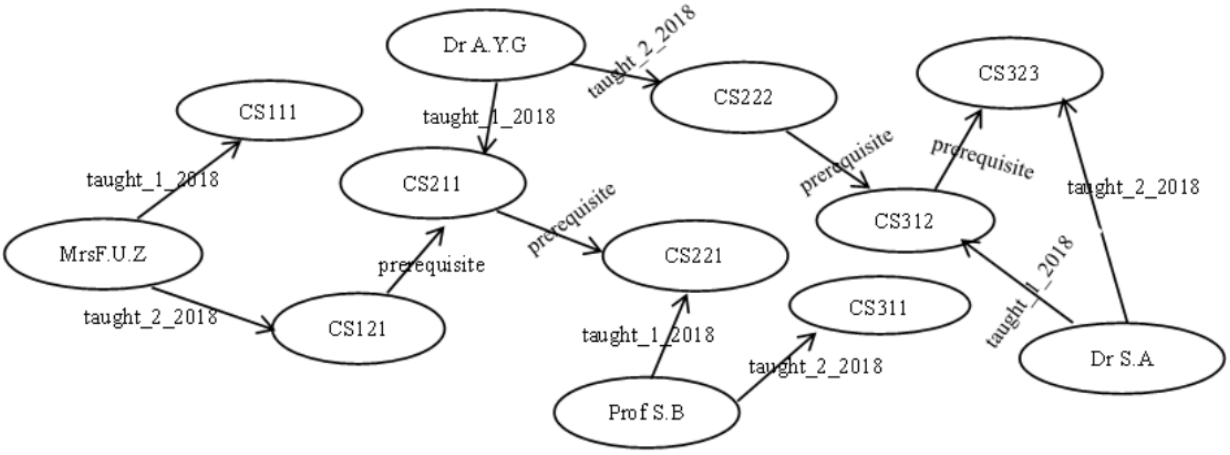

Figure 8: Sample Knowledge graph for courses by Aliyu & Kana, 2020

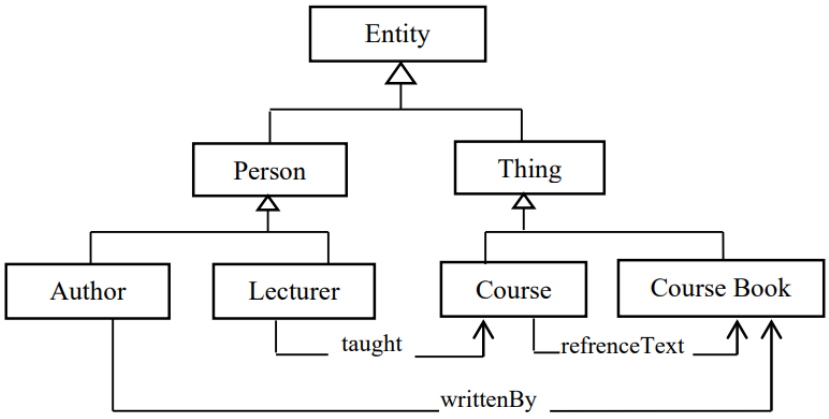

Figure 9: Sample data schema for the courses management model by Aliyu & Kana, 2020

Table 1: List of relation of courses by Aliyu & Kana, 2020

Relation	Description
Prerequisite	relation between *course* and another *course* entities
taught_in_semester_year	relation between *course* entity and *lecturer* entity
writtenBy	relation between *author* entity and *coursebook* entity
referenceText	relation between *course* entity and *coursebook* entity

Table 2: Comparison of extraction tools

Tool	Language	Supported Formats	Extracts Metadata	OCR Capabilities	Accuracy	Speed
Textract	Python	pdf, doc, ppt, docx	Yes	No	Accurate	Fast
Apache Tik	Java	pdf, doc, ppt, docx, xls, xlsx	Yes	Yes	Most accurate	Medium
PyPDF2	Python	pdf	No	No	Acurate	Fast

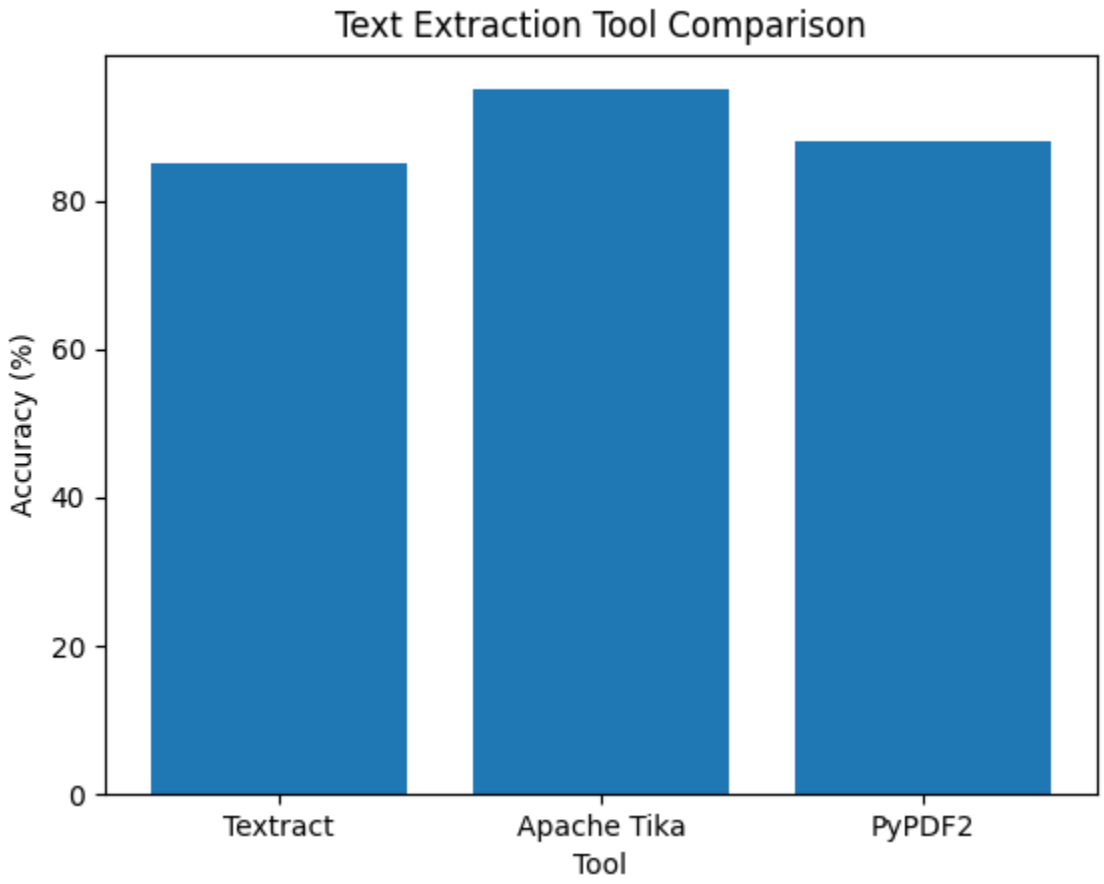

Figure 10: Graphical comparison of extraction tools

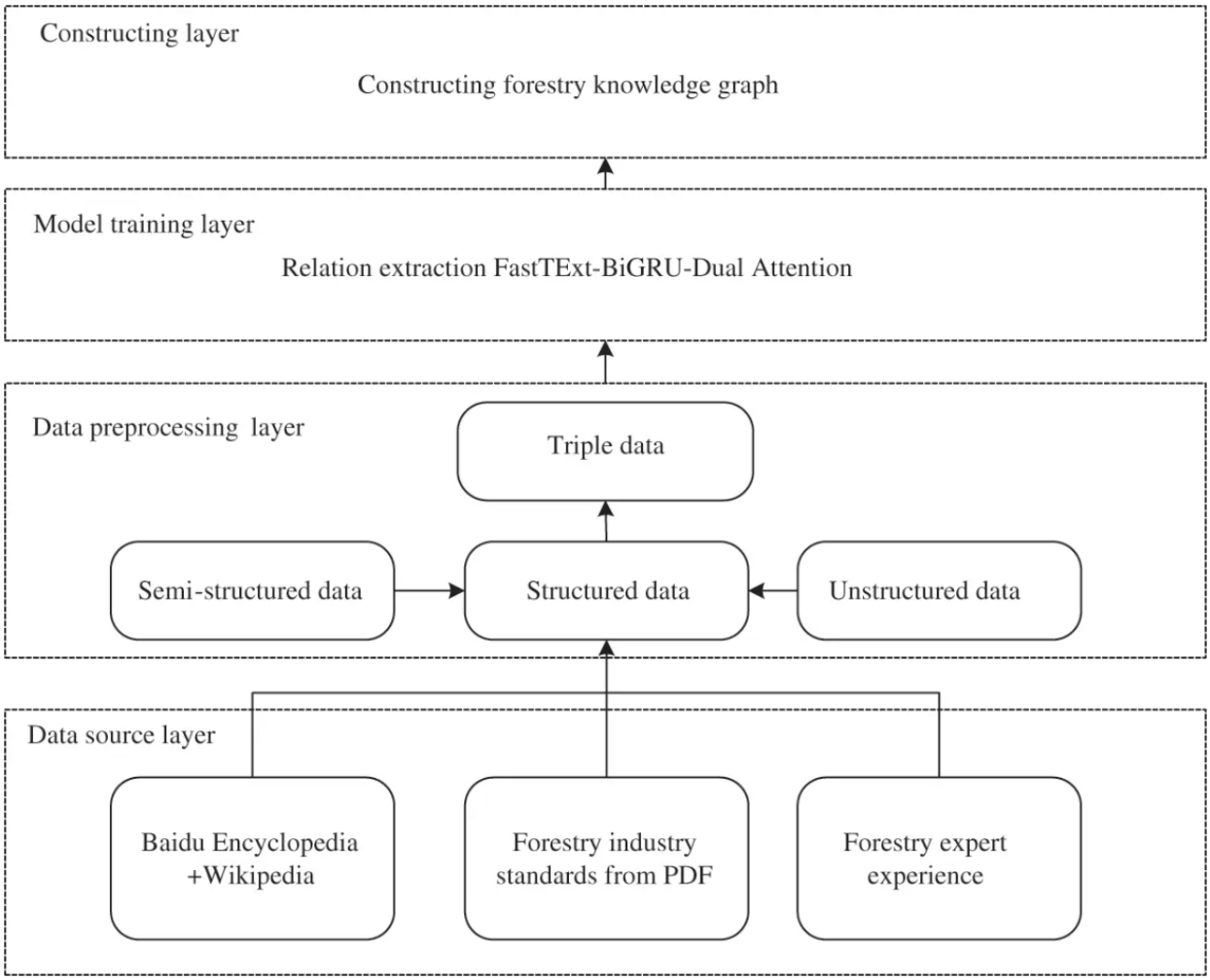

Figure 11: Model framework for the forestry KG by Yue & Li, 2021

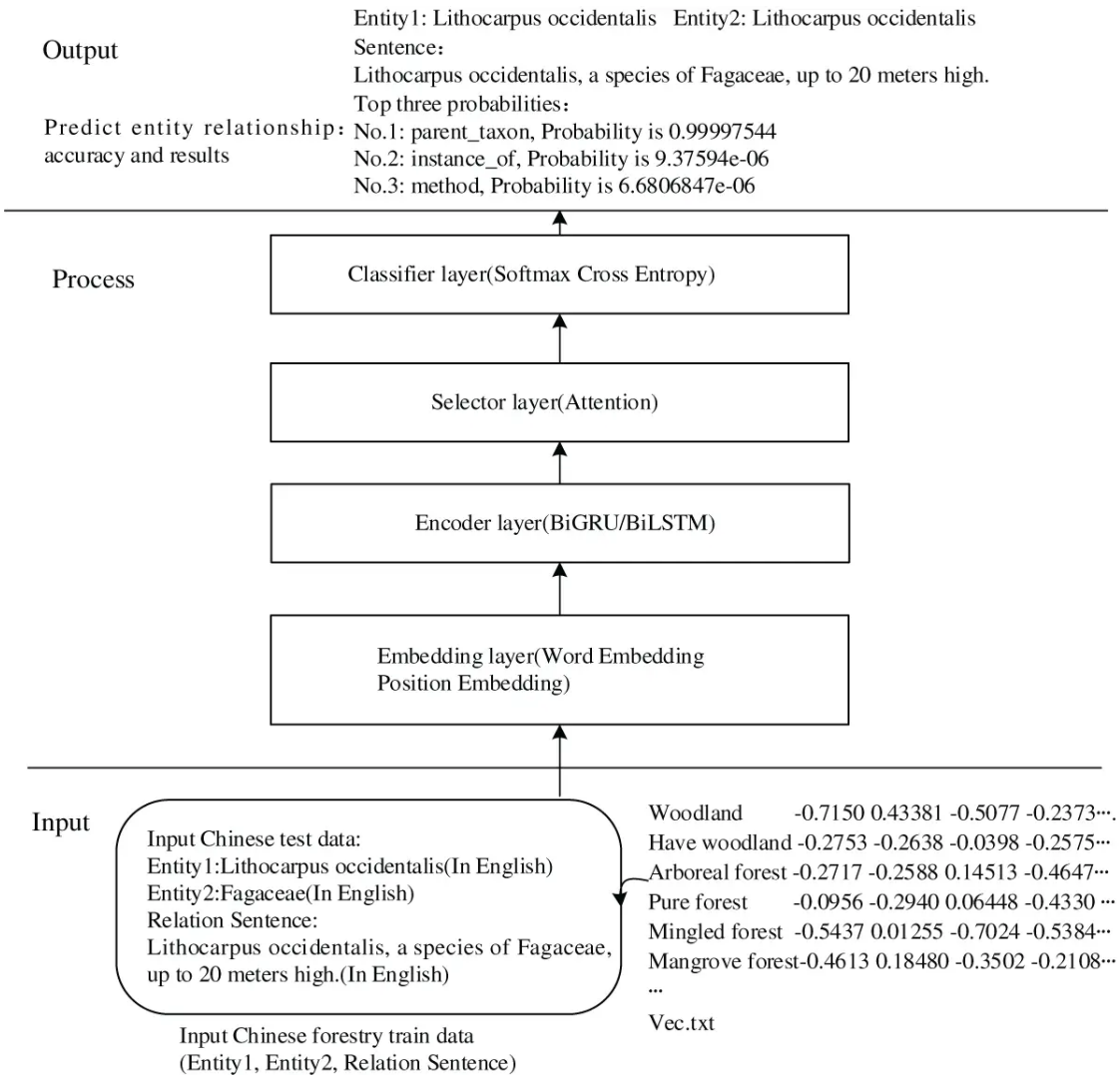

Figure 12: The pipeline of forestry entity relation extraction by Yue & Li, 2021