吉林大学

JILIN UNIVERSITY

Construction of a question-answering system for stock knowledge graph.

College of Computer Science and Technology, Jilin University

–zmgGroup –

Document date:2023 11.07

Group member:Tianxu Zhang,Zengyao Man,Jun Gao

目录

1 Knowledge Graph Codebook

1.1 Knowledge Graph general description

. The Problem the KG aims to solve:

The Stock based Financial Knowledge Q&A system aims to provide a platform for users to query stock related information, which can:

Explain terminology: Explain technical terms and concepts in the stock market and financial transactions.

Analyzing trends: Analyzing stock market trends, including technical analysis and fundamental analysis.

Forecasting the market: Using historical data and algorithms to predict the future performance of a stock.

Personalized advice: Provide investment advice based on the user's investment preferences and risk tolerance.

Educate investors: Help investors understand the operation mechanism and investment strategy of the stock market.

Enhance decision making: Provide data support to help investors make more informed investment decisions.

Such a system is beneficial for individual investors, financial advisors, market analysts, and financial learners as it provides a way to quickly access, analyze, and understand stock market information.

How the KG can solve The Problem

Data Integration and Association: A knowledge graph integrates information from diverse data sources such as stock market data, news reports, financial statements, etc., to establish connections between data points. This helps users understand the complex relationships between stocks and related entities.

Understanding the Context of Queries: Through natural language processing of user queries, a knowledge graph can comprehend user intent and provide relevant information, such as term definitions or stock performance analysis.

Automated Q&A: Systems can automatically answer user questions using predefined graph structures and algorithms, reducing reliance on professional analysts.

1.2 Data level

1.2.1 Datasets general details

The QABasedOnKnowledgeGraph system uses multiple datasets to build and train the knowledge graph, supporting question-answering functionality. These datasets contain knowledge and information from various domains, enabling the system to provide accurate and comprehensive answers.

The overview of the datasets is as follows:

Entity dataset: It contains information about various entities such as people, places, organizations, events, etc. Each entity has a unique identifier and related attributes like name, description, category, etc.

Relationship dataset: This dataset describes the relationships and connections between entities. These relationships can be hierarchical, associative, or attribute-based. Each relationship in the dataset includes the relationship type, starting entity, and target entity.

Attribute dataset: It includes attribute information of entities such as age, gender, date of birth, address, etc. Each attribute in the dataset has a corresponding data type and value range.

External dataset: This dataset comprises data from public databases, knowledge repositories, and other data sources to supplement and enrich the content of the knowledge graph.

The combination and integration of these datasets form the knowledge graph in the QABasedOnKnowledgeGraph system. Through analysis, processing, and reasoning on these datasets, the system can understand user queries and provide accurate answers. The quality and completeness of the datasets are crucial for the system's performance and accuracy, making data collection and maintenance essential aspects of system development.

1.2.2 Datasets metadata documentation

In this section, we provide detailed metadata documentation for each variable in the datasets. We describe the variable types, meanings, and the possible values they can take. Additionally, we include any other relevant information about the variables that can help in understanding the data and its relationship to the knowledge graph.

For each dataset variable, the metadata documentation includes the following information:

Variable name: The name or identifier of the variable in the dataset.

Variable type: The data type of the variable, such as string, integer, date, etc.

Variable meaning: A description of what the variable represents or signifies in the context of the knowledge graph.

Possible values: The range or set of values that the variable can take. This may include specific values, ranges, or categories.

Relationship to the knowledge graph: An explanation of how the variable is related to the entities, relationships, or attributes in the knowledge graph.

Additional information: Any other relevant details about the variable, such as units of measurement, data format, or any special considerations.

By providing this metadata documentation, users can have a comprehensive understanding of the variables in the datasets and their significance in the context of the knowledge graph. This information is crucial for data analysis, interpretation, and utilization within the QABasedOnKnowledgeGraph system.

1.3 Ontology level

The ontology level section aims to describe the underlying KG ontology, through the description of its elements at each level, reporting so the language, conceptual and schema resources used within it.

1.3.1 Ontology general details

Sources: The data used to construct the knowledge graph in the QABasedOnKnowledgeGraph system is sourced from various reputable and reliable sources. These sources include public databases, knowledge repositories, academic publications, and other relevant data sources. The specific sources used for each dataset are documented in the respective dataset metadata.

External Ontology: The QABasedOnKnowledgeGraph system may adopt external ontologies to enhance the final knowledge graph. These external ontologies provide additional domain-specific knowledge and help in organizing and structuring the information within the knowledge graph. The selection and adoption of external ontologies are based on their compatibility and relevance to the system's domain and objectives. The details of any external ontologies adopted are documented in the ontology level description.

Description of External Ontology: If an external ontology is adopted, its description includes information about its purpose, scope, structure, and any specific concepts or relationships it introduces to the knowledge graph. The external ontology's contribution to the final knowledge graph is explained, highlighting how it enriches the system's understanding and answering capabilities.

By including this information, users can understand the origins of the data used in the system, the team responsible for its development, and any external ontologies that have been incorporated to enhance the knowledge graph. This transparency allows for better evaluation and utilization of the QABasedOnKnowledgeGraph system.

1.3.2 Ontology metadata documentation

Ontology Description: QABasedOnKnowledgeGraph

1. Terms:
   - Term 1: Question
     - Description: A statement or inquiry seeking information or clarification.
     - Properties: Text, Category, Difficulty Level, Answer

   - Term 2: Answer
     - Description: A response or solution to a question.
     - Properties: Text, Source, Confidence Level

2. Concepts:
   - Concept 1: Knowledge Graph
     - Description: A structured representation of knowledge that captures entities, relationships, and attributes.
     - Properties: Name, Description, Sources, Entities, Relationships

   - Concept 2: Natural Language Processing
     - Description: A field of artificial intelligence that focuses on the interaction between computers and human language.
     - Properties: Name, Description, Techniques, Applications

3. ETypes (Entity Types):
   - EType 1: Entity
     - Description: A specific object, person, place, or concept.
     - Properties: Name, Description, Attributes, Relationships
     - Relationships: Related Entities, Associated Attributes

   - EType 2: Relationship
     - Description: A connection or association between two or more entities.
     - Properties: Name, Description, Type, Strength
     - Relationships: Connected Entities, Associated Attributes

4. Relations:
   - Relation 1: Related Entities
     - Description: Indicates a relationship between two entities that are related or connected.
     - Properties: Relationship Type, Strength of Connection
     - Related Entities: Entity, Entity

   - Relation 2: Associated Attributes
     - Description: Indicates the attributes or properties associated with an entity or relationship.
     - Properties: Attribute Name, Attribute Value
     - Related Entities: Entity, Attribute

This ontology focuses on the domain of the QABasedOnKnowledgeGraph system and provides a structured representation of questions, answers, knowledge graphs, and related concepts. It captures the relationships between these elements, enabling the system to effectively process and provide accurate answers to user queries. The ontology also incorporates the concepts of natural language processing to facilitate the understanding and interpretation of questions in natural language form.

1.4 Knowledge Graph Evaluation

Strengths:

Comprehensive Coverage: The knowledge graph encompasses various aspects of the stock market, including key nodes such as stocks, market indices, top managers, and financial movement, providing a holistic view.

Sentiment Analysis: The incorporation of sentiment analysis adds a layer of understanding regarding market perception, providing users with insights into the market sentiment surrounding specific stocks.

Weakness:

Dynamic Nature of the Market: Given the dynamic nature of the stock market, it's important to continuously update the knowledge graph to reflect changing market trends, investor sentiments, and other relevant factors.

Data Accuracy and Integrity: Ensuring the accuracy and integrity of both historical and real-time data remains crucial to maintain the credibility of the knowledge graph. Any inaccuracies could lead to misguided decisions by users relying on this data.

2 Knowledge Graph Development Process

2.1 Scope Definition

Stock information has always been A hot topic. The project collected the basic information data of A-shares through the stock information on the websites of Oriental Finance, Flush and Wen cai, and then carried out data cleaning. In this stage, duplicate items are deleted and missing values are completed and processed. Then, py2neo library is used to create a stock-centered knowledge graph to realize knowledge import, and knowledge fusion is carried out on entities, relationships and attributes of stock companies to improve the quality of knowledge graph

2.2 Inception

2.2.1 Scenarios of usages

Actor: Alice,25

Scenario: Alice is a young investor who is very interested in the stock market. She hopes to increase her wealth by investing in stocks. She was looking for a query solution that would provide basic stock information, historical data and market trend analysis. She hopes to be able to easily understand the potential investment value of stocks through this solution and make informed investment decisions.

Actor: Bob,40

Scenario: Bob is a stock trader with many years of investing experience. He needed a query solution that could provide real-time stock quotes, trading data, and technical indicator analysis. He hopes that through this solution, he can quickly obtain the dynamics of stocks in the market and make timely trading decisions based on the analysis results.

Actor: Charlie, 60

Scenario: Charlie is a retired investor who hopes to preserve and increase his value by investing in stocks and provide a stable source of income for his retirement life. He was looking for a query solution that would provide basic information about the stock, the company's financials and dividend payouts. Through this solution, he hopes to be able to learn about potential high-dividend stocks and make long-term investment decisions.

2.2.2 CQs definition

Table 1: Query Description

Actor	Query
Alice	As a newly graduated undergraduate,Alice is a young investor who is interested in the stock market. She has little experience of investing a stock.
Alice	She was looking for a query solution that would provide basic stock information, historical data and market trend analysis. We should contain some concept and history data.
Alice	She hopes to be able to easily understand the potential investment value of stocks through this solution and make informed investment decisions.wo should give detailed information.
Bob	Bob is a stock trader with many years of investment experience. He needed a query solution that could provide real-time stock quotes, trading data, and technical indicator analysis. We should share the latest data such as buy signal sell signal.
Bob	He hopes that through this solution, he can quickly obtain the dynamics of stocks in the market and make timely trading decisions based on the analysis results.we should provide the movement of the stock.
Charlie	Charlie is a retired investor who wants to preserve and increase his value by investing in stocks and provide a stable source of income for his retirement life.so wo should obtain some stable stocks.
Charlie	He was looking for a query solution that would provide basic information about the stock, the company's financial and dividend payouts.we should get the basic information of stocks and select the promising stocks.

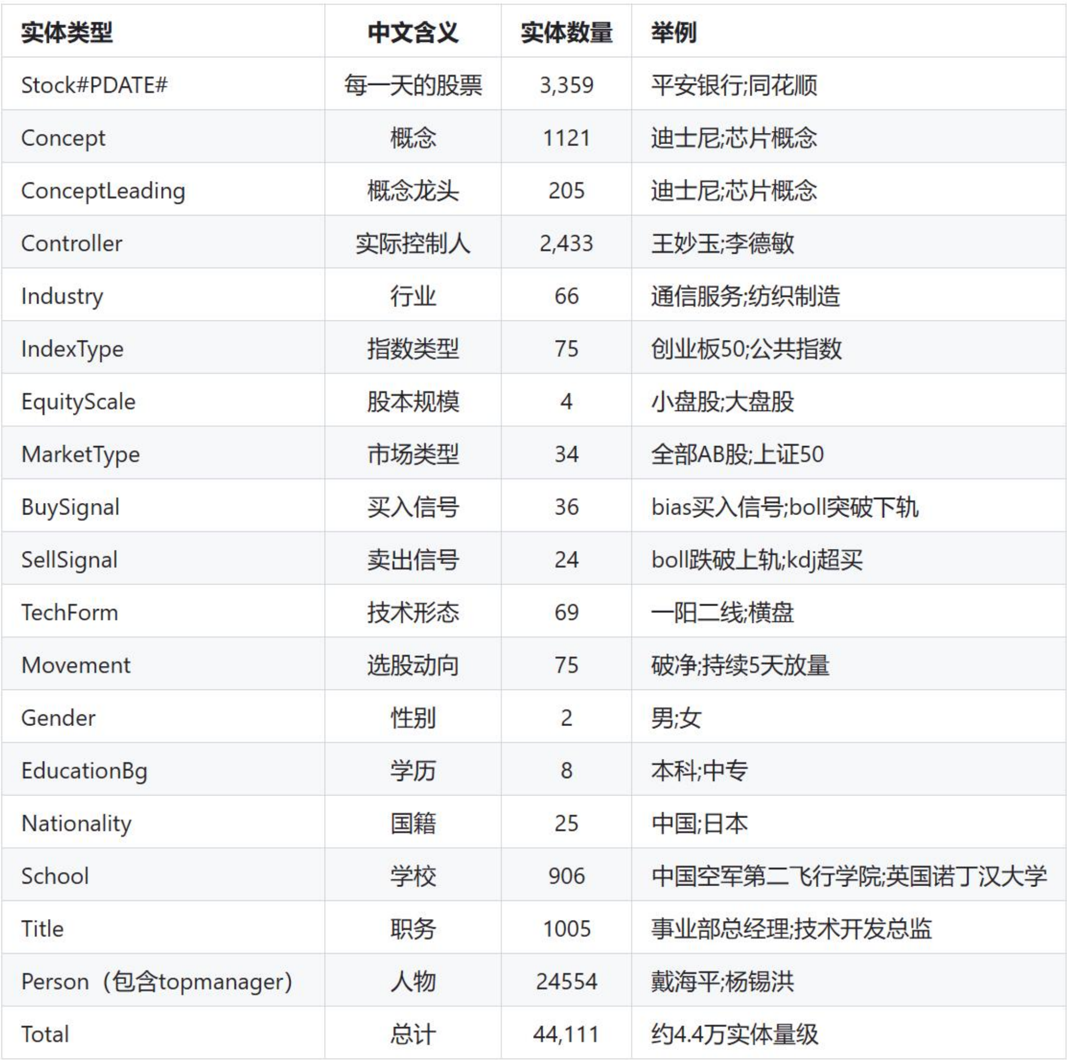
2.2.3 Initial Datasets description

Since we focus on providing detailed information to stock buyers, in order to get enough stock data, we visited several websites and finally chose https://iwencai.com, one of the largest stock information websites that lists buy signals, sell signals, movements, stock names, concepts and leaders. We also use tushare interface to obtain stock trading information to improve the opening price, closing price, volume ratio and other attributes of stock entities, update stock information in real time, and obtain the senior executives of stock companies to improve the Q&A system.

2.2.4 Datasets metadata documentation

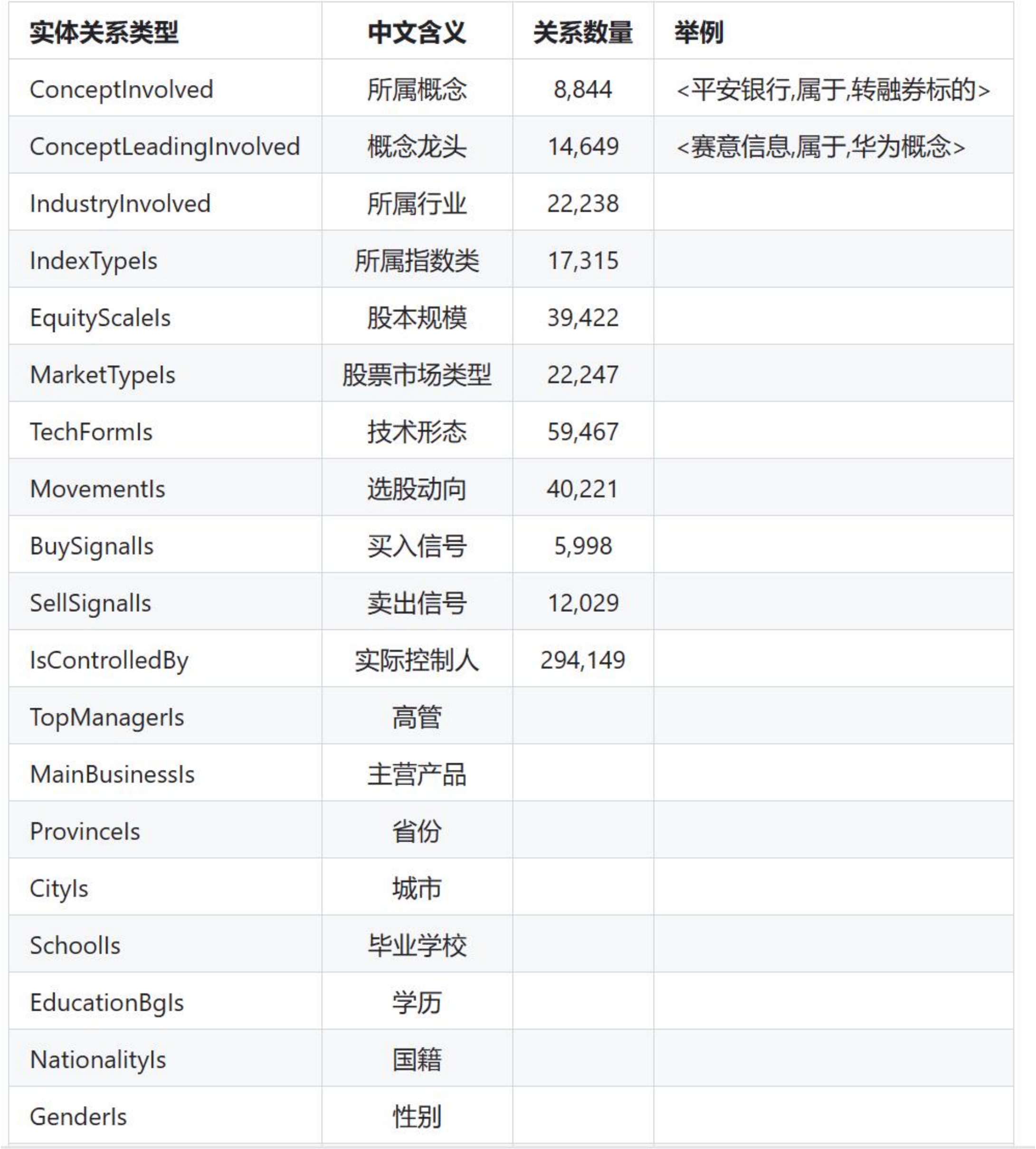
We have collected three datasets and are overall information of the stocks,additional information of the stocks and the the top manager of stock company. here are the metadata tables about them respectively.

1.Entity Data



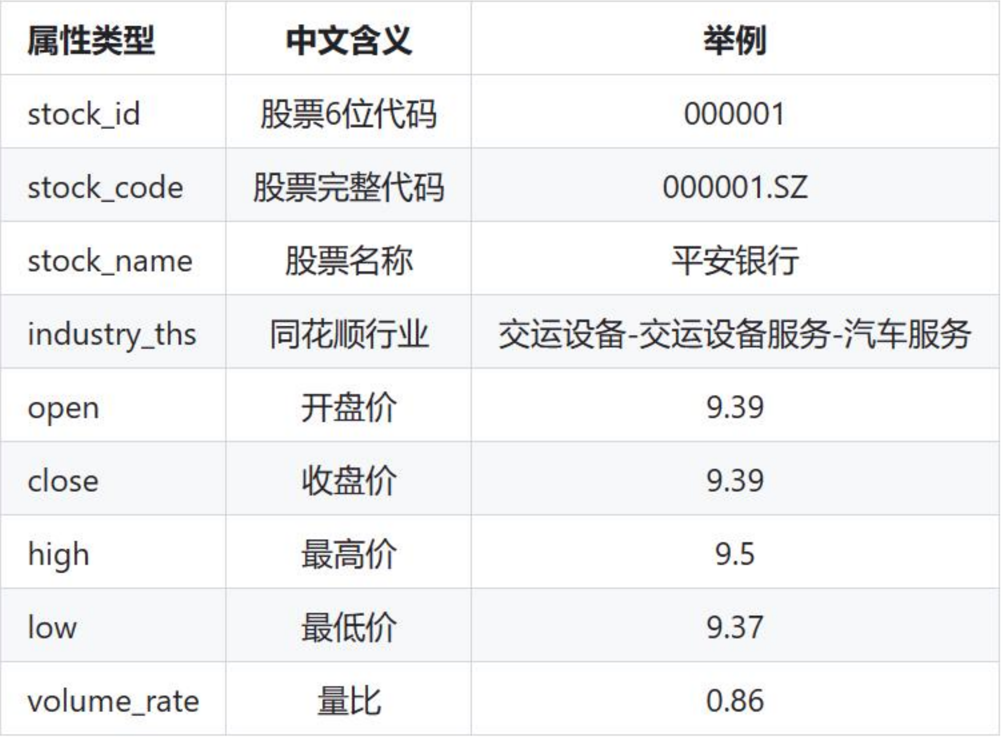
实体类型	中文含义	实体数量	举例
Stock#PDATE#	每一天的股票	3,359	平安银行;同花顺
Concept	概念	1121	迪士尼;芯片概念
ConceptLeading	概念龙头	205	迪士尼;芯片概念
Controller	实际控制人	2,433	王妙玉;李德敏
Industry	行业	66	通信服务;纺织制造
IndexType	指数类型	75	创业板50;公共指数
EquityScale	股本规模	4	小盘股;大盘股
MarketType	市场类型	34	全部AB股;上证50
BuySignal	买入信号	36	bias买入信号;boll突破下轨
SellSignal	卖出信号	24	boll跌破上轨;kdj超买
TechForm	技术形态	69	一阳二线;横盘
Movement	选股动向	75	破净;持续5天放量
Gender	性别	2	男;女
EducationBg	学历	8	本科;中专
Nationality	国籍	25	中国;日本
School	学校	906	中国空军第二飞行学院;英国诺丁汉大学
Title	职务	1005	事业部总经理;技术开发总监
Person（包含topmanager）	人物	24554	戴海平;杨锡洪
Total	总计	44,111	约4.4万实体量级

2.Relation Data



实体关系类型	中文含义	关系数量	举例
ConceptInvolved	所属概念	8,844	<平安银行,属于,转融券标的>
ConceptLeadingInvolved	概念龙头	14,649	<赛意信息,属于,华为概念>
IndustryInvolved	所属行业	22,238	
IndexTypeIs	所属指数类	17,315	
EquityScaleIs	股本规模	39,422	
MarketTypeIs	股票市场类型	22,247	
TechFormIs	技术形态	59,467	
MovementIs	选股动向	40,221	
BuySignalIs	买入信号	5,998	
SellSignalIs	卖出信号	12,029	
IsControlledBy	实际控制人	294,149	
TopManagerIs	高管		
MainBusinessIs	主营产品		
ProvinceIs	省份		
CityIs	城市		
SchoolIs	毕业学校		
EducationBgIs	学历		
NationalityIs	国籍		
GenderIs	性别		

3.Attribute Data



属性类型	中文含义	举例
stock_id	股票6位代码	000001
stock_code	股票完整代码	000001.SZ
stock_name	股票名称	平安银行
industry_ths	同花顺行业	交运设备-交运设备服务-汽车服务
open	开盘价	9.39
close	收盘价	9.39
high	最高价	9.5
low	最低价	9.37
volume_rate	量比	0.86

2.2.5 Datasets collection process

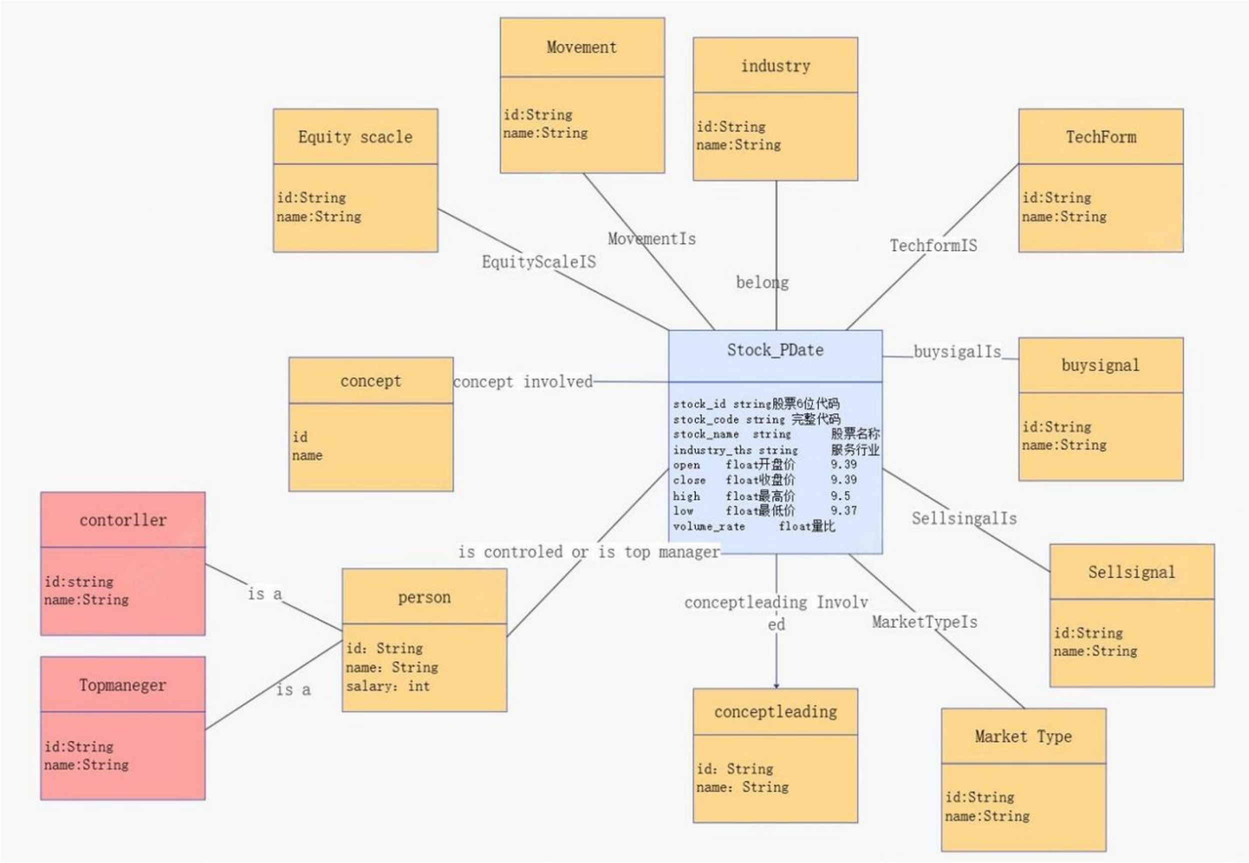
We try to use web scraping tools to obtain data from websites such as Oriental Finance and Flush and some open source data to create a stock based financial knowledge Q&A system. Given Python's convenience in web scraping, we use Python crawlers to get the data we need. In the Python script, we extract the important features of each stock, as mentioned above. To get real-time data, we used tushare's API to get stock trading information and managers. Finally, we have all the raw data sets we need.

2.3 Informal Modeling

This section is dedicated to the Informal Modeling phase description. The Section is divided in Schema and Data level in order to report the details of the elements involved in the generation of the schema, as well as the description of the datasets evolution in this phase. Moreover a specif section, one for each level, reports the difference between the elements defined in this phase and the definitions in the previous phase, analyzing in this way the variance in the different phases.

2.3.1 Schema level

2.3.1.1 ETypes and EER Model definition



To complete the equity-based financial Q&A, we built the EER model graph above. We use the stock as the core entity because we will focus on the properties of the stock to evaluate the status of a stock. As the most basic information of the event, the stock information should be as detailed as possible, so we collected the stock code, name, opening price, closing price, service industry and other information, in addition, we also collected some major data, such as the share capital size of the stock, buy signal, sell signal and trend information.

2.3.1.2 Variance respect CQs definition

Variance respect CQs refers to the consideration of question variations in the QABasedOnKnowledgeGraph system when answering complex questions. Complex questions may have multiple expressions or variants, but they involve the same topic or information. To provide accurate and comprehensive answers, the system needs to identify and understand the variance of the questions and provide corresponding answers for different question variants.

The definition of Variance respect CQs includes the following aspects:

Question Variants: Question variants refer to the variations in expression or syntactic structure of questions that involve the same topic. For example, for the question "Who was the first president of the United States?", there may be variants like "Who was the first president of America?" or "Who was the person to first hold the presidency in the United States?"

Variance Recognition: The system needs to be able to recognize the variance of questions, which means understanding the relatedness and similarity between different question variants. This can be achieved through natural language processing techniques and models that analyze the semantics and structure of questions to determine the similarity and differences between question variants.

Variance Modeling: The system needs to establish appropriate models to handle the variance of questions. This includes establishing the relationships and mappings between question variants and identifying shared information and features among question variants. By modeling the variance of questions, the system can better understand the meaning and requirements of the questions and provide accurate and consistent answers.

Variance-aware Answering: The system needs to provide corresponding answers based on the variance of questions. This means that the system should be able to retrieve relevant information from the knowledge graph and generate appropriate answers based on the specific expression and requirements of the question. By considering the variance of questions, the system can provide more personalized and tailored answers to meet the users' needs.

The concept and implementation of Variance respect CQs are crucial for the performance and user experience of the QABasedOnKnowledgeGraph system. It enables the system to better understand and handle different variants of complex questions and provide accurate and consistent answers.

2.3.2 Data level

The data level section in this phase reports the evolution of the datasets collected previously, reporting the metadata information for each new data, or new version of data, obtained.

2.3.2.1 Datasets management process

The dataset management process for the stock knowledge graph involves handling and organizing datasets used for training, testing, and evaluating the system. This process ensures the quality, completeness, and accessibility of the datasets, enabling effective development and improvement of the stock question-answering system.

The datasets management process includes the following steps:

1. Data Collection: Relevant data is collected from various sources such as websites like Shenzhen Stock Exchange and East Money. This data includes stock prices, company financial data, market indices, and more. The diversity of data ensures the system's ability to answer stock-related questions from different domains.

2. Data Cleaning: The collected data may contain errors, missing values, or inconsistencies. Data cleaning involves removing duplicates, handling missing values, correcting errors, and standardizing data formats. This step ensures the quality and reliability of the datasets.

3. Data Annotation: Data annotation involves labeling or tagging the collected data with relevant information such as stock company names, industry classifications, market values, and relationships between stock prices and financial indicators. Annotation ensures accurate answers to stock-related questions.

4. Dataset Versioning: Maintaining different versions of the dataset is important as the stock market and company situations change over time. Different versions allow tracking changes, comparing results, and ensuring reproducibility.

5. Dataset Documentation: Proper documentation of the dataset is essential for understanding its structure, annotation guidelines, limitations, and biases. Documentation includes information about data sources, collection methods, annotation guidelines, and any specific considerations.

6. Dataset Updates: Regular updates are necessary to incorporate the latest stock prices, financial data, and user feedback. Updates ensure the system's ability to handle new information and improve performance.

7. Dataset Security and Privacy: Ensuring the security and privacy of sensitive data is crucial. Measures should be taken to protect personal information and comply with data protection regulations.

The dataset management process plays a critical role in the development and performance of the QABasedOnKnowledgeGraph system. It ensures the availability of high-quality and relevant data for training and evaluation, enabling the system to provide accurate and reliable answers to stock-related questions.

2.3.2.2 Datasets metadata documentation

Dataset Metadata Documentation During the data cleaning process, some changes were made to the dataset attributes, but the metadata remains the same as provided in section 2.2.4. For example, we added attributes such as industry classification and market value for stock companies, corrected errors in financial indicators, and standardized the dataset format.

2.3.2.3 Variance respect Inception datasets

In the stock knowledge graph, the dataset primarily consists of quantitative indicators such as stock prices, market values, and financial indicators. Due to the wide range of variations in these indicators, it is important to normalize the numerical values in the dataset for better comparison and analysis of different stocks.

2.3.3 Informal Modeling Evaluation

In the development of the stock knowledge graph, the reliability of the informal model is evaluated based on the accuracy of query results. For example, when a user queries the market value of a specific stock, the system should provide the corresponding market value from the dataset rather than incorrect information. By evaluating the accuracy of query results, we can assess the reliability of the informal model in answering stock-related questions.

2.4 Formal Modeling

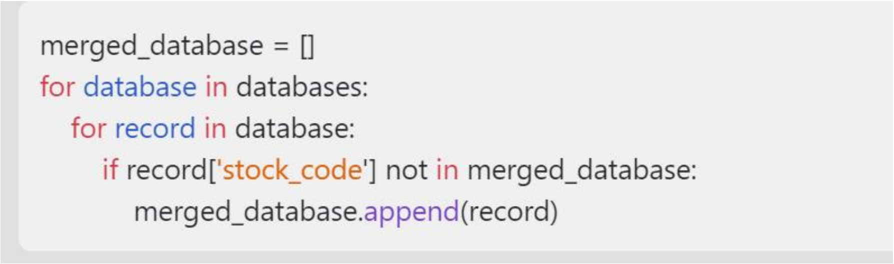
This section is dedicated to the Formal Modeling phase description. The Section is divided in Schema and Data level in order to report the details regarding both the ontology generated and the datasets version in the current phase.

2.4.1 Schema level

The schema level section in the current phase, reports the detailed description of the ontology generation.

2.4.1.1 Ontology definition

In the process of constructing a stock knowledge graph, the first step is to search for other reference ontologies. Although there are not many reference examples in this project, based on prior knowledge, we know that stock data comes from different data sources, including attributes such as stock code, stock name, trading date, opening price, closing price, and so on. There is a one-to-one correspondence between stock codes and stock entities, and stock codes are provided by the stock entities themselves. The data sources are responsible for collecting and organizing relevant information without subjective processing. Therefore, the stock code attribute fields in the heterogeneous databases completely overlap without ambiguity, which facilitates data-level fusion. In this project, we adopt a simple approach, which is to merge multiple data sources based on the restriction of stock codes to create a fused database. The following is an example pseudo code for this algorithm:



```
merged_database = []
for database in databases:
    for record in database:
        if record['stock_code'] not in merged_database:
            merged_database.append(record)
```

2.5 Data integration

2.5.1 Data integration operations and tool

In this project, the main strategies for data fusion are matching and concatenation of identical domain attributes based on heterogeneous databases, as well as filtering based on query conditions. The stock code attribute serves as a bridge between different databases, and its uniqueness and identifiability facilitate the implementation of the fusion strategy. In addition, filtering based on personalized query statements supports the export of correct results.

2.5.2 Variance respect Formal Modeling datasets

The final part of the data integration stage aims to describe the differences between the data set integrated with the ontology and the data set collected in the previous stage, and analyze it in the data integration platform containing the knowledge graph. This analysis can highlight the operational results of the final stage of the data integration process. In the process of constructing a stock knowledge graph, we will analyze the differences between the data set integrated into the ontology and the previously collected data set, in order to highlight the operational results of the final stage of the data integration process.

Data Source Tonghuashun:

- Stock Code: AAPL
- Stock Name: Apple Inc.
- Trade Date: 2021-01-01
- Opening Price: 130.84
- Closing Price: 132.05

Data Source Dongfang Fortune:

- Stock Code: AAPL
- Stock Name: Apple Corporation
- Trade Date: 2021-01-01
- Opening Price: 131.50
- Closing Price: 133.00

During the data integration process, we apply matching and splicing based on the same domain attributes, such as the stock code. By matching the stock codes, we merge the stock information from both sources and obtain the following integrated result:

Integrated Result:

- Stock Code: AAPL
- Stock Name: Apple Inc.
- Trade Date: 2021-01-01
- Opening Price: 130.84
- Closing Price: 132.05

As we can see, the integrated result retains the stock name and price information from Data Source A, while ignoring the duplicate information from Data Source B. Through data integration, we successfully merge the stock information from both sources into a unified record, eliminating duplicate and redundant data.

This example demonstrates the use of matching and splicing based on the same domain attributes, as well as the filtering of duplicate information, during the data integration process. The integrated result provides more accurate and consistent stock information, serving as a reliable data foundation for constructing a stock knowledge graph.