IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received March 31, 2021, accepted April 26, 2021, date of publication May 6, 2021, date of current version May 17, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3077916*

# Financial Knowledge Graph Based Financial Report Query System

**SAMREEN ZEHRA[1], SYED FARHAN MOHSIN[1], SHAUKAT WASI[1], SYED IMRAN JAMI[1], MUHAMMAD SHOAIB SIDDIQUI[2], (Member, IEEE), AND MUHAMMAD KHALIQ-UR-RAHMAN RAAZI SYED[1], (Member, IEEE)**

[1]Department of Computer Science, Mohammad Ali Jinnah University, Karachi 75400, Pakistan
[2]Faculty of Computer and Information Systems, Islamic University of Madinah, Medina 42351, Saudi Arabia

Corresponding authors: Syed Farhan Mohsin (farhan.mohsin@duhs.edu.pk) and Shaukat Wasi (shaukat.wasi@jinnah.edu)

This work was supported in part by the Mohammed Ali Jinnah University, Karachi, and in part by the Islamic University of Madinah.

**ABSTRACT** Annual Financial Reports are the core in the Banking Sector to publish its financial statistics. Extracting useful information from these complex and lengthy reports involves manual process to resolve the financial queries, resulting in delays and ambiguity in investment decisions. One of the major reasons is the lack of any standardization in the format and vocabulary used in the reports. An automated system for resolution of intelligent financial queries is therefore difficult to design. Several works have been proposed to overcome these problems using Information Extraction; however, they do not address the semantic interoperability of the reports across different institutions. This work proposed an automated querying engine to answer the financial queries using Ontology based Information Extraction. For Semantic modeling of financial reports, a Financial Knowledge Graph, assisted by Financial Ontology, has been proposed. The nodes are populated with entities, while links are populated with relationships using Information Extraction applied on annual reports. Two benefits have been provided by this system to stakeholders through automation: decision making through queries and generation of custom financial stories. The work can further be extended to other domains including healthcare and academia where physical reports are used for communication.

**INDEX TERMS** Ontology, financial knowledge graph, information extraction.

## I. INTRODUCTION

Stakeholders seek information regarding company profile and its general financial standing before taking any decision though various channels. They usually restrict their research to the financial indicators, such as, revenues, net profit, earnings per share (EPS) and price to earnings ratio (PE ratio) and credit ratings mentioned in its financial filings. The required information is widely dispersed in the quarterly and annual reports declared by companies; therefore, it is difficult for investors to read and interpret the financial implications mentioned therein. The other problem is almost each and every company produce a bulky report and it is quite cumbersome for the stakeholders to go through it. Yet experts argue that investors should study management discussion and analysis along with directors' report to get a clear understanding of current state of affairs of a business.

The associate editor coordinating the review of this manuscript and approving it for publication was Wajahat Ali Khan.

Automatic information extraction from these financial disclosures is hard, owing to the lack of boundaries between the items to be extracted, context dependence of the targets entities, language pattern variations, and statistical methods limitations [1]. Another problem in information extraction from these financial datasets is that these are usually available as non-structured texts or in PDF that involves meticulous manual preprocessing or application of sophisticated ETL (Extract, transform, load) tools in order to ingest data automatically [2], [3]. This step will be done manually for our research work and resulting data will be stored in separate text files for each entity. The system scope is defined after analyzing the dataset and competency questions.

The integration of information extraction and the semantic web help in extracting the related information from the heterogeneous data formats and from multiple sources in the desired format. The addition of the new node or relations or deletion of the previous node has made no effect in the consistency and the information schema. It has the flexibility

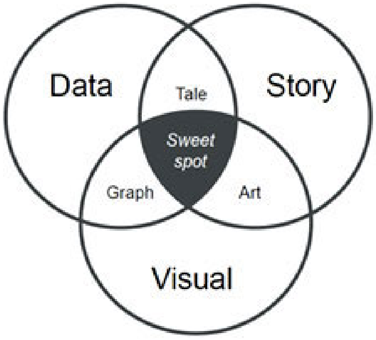


**FIGURE 1.** Sweet spot of data, story and visual [51].

to gather, exchange, and update the information from different sources; the new nodes and the extracted information is easily adjustable in the existing format without disturbing the structure of the ontology [47].

We have employed knowledge graph (KG) for several reasons in this research. Firstly, it may not necessarily have some semantic layer to describe the entity model, as it is required in relational databases [4]. Secondly, the schema is flexible and easily adjustable that it is always easier to add new properties to an existing record or modify schema without affecting other graph entities [5]. Highly variable, incomplete, or dynamic data can be represented by Property Graph stores that consume less space and supports attribute/link discovery. Finally, graph databases can answer queries that span over multiple entities by graph traversal. Only those nodes which are accessible because of the query, are explored by the graph database engine. Because every record is handled individually, it drastically boosts the query performance and helps in reducing resource cost of the query results [6], [7].

Our knowledge graph will assist an investor by generating financial stories that can aid decision making. Further plan is to generate text based financial stories and further extend it to visualizations/graphics as proposed by [8], the whole concept is shown in Fig. 1.

Following are the major contributions of this work: (A) Integration of Information Extraction with Semantic Web (B) Proposed Financial Knowledge Graph to model the domain of Financial Systems (C) Mapping Financial Reports in the domain using Ontology (D) Extending manual financial reports as machine readable using Information Extraction.

## II. RELATED WORK

### A. INFORMATION EXTRACTION

In recent years, most of the Financial Information extraction work has been done on a specific reporting standard known as XBRL (eXtensible Business Reporting Language) which is a freely available and global framework for exchanging business information. In 2011, the Securities and Exchange Commission (SEC) mandated XBRL as the filing standard for all US public companies. A rule-based information extraction methodology was introduced in [1] for the extraction of highly accurate financial information to aid investment decisions. They trained two different rule-based symbolic learning models using Tabu Search algorithm and Greedy Search algorithm and evaluated their performances using financial filings. In another research [13], authors implemented a software agent which extracts fundamental company data from the Electronic Data Gathering, Analysis and Retrieval (EDGAR) database of the United States Securities and Exchange Commission (SEC) and outputs this data in a format which is useful to support stock market trading decisions. EDGAR is a specialized database which stores information as provided by companies in the 10-k Format or XBRL formats [14]. A two-step approach was proposed in [15] to perform rule-based text extraction and acquisition of structured data from unstructured text. In this work, we are working with annual financial disclosures which require data extraction from PDF files that includes tables, graphics, structured and unstructured text [2], [16], [17] [18].

### B. ONTOLOGY BASED INFORMATION EXTRACTION

In Ontology-Based Information Extraction (OBIE), information extraction process is assisted by Ontologies [19]. Ontology is defined as a formal and explicit specification of a shared conceptualization and is usually knowledge domain specific [20]. As the task of information extraction also deals with retrieving information for a particular domain, ontology is one of the candidate solutions in information extraction [21]. Researchers have been using ontology-based mechanisms for extracting required information from unstructured or semi-structured natural language text [21], [22]. An Ontology model is developed for mobile payment data risk control domain in [43]. The model takes the user as entity and operation/transaction as relationship and gathers the data on separate timestamp to fulfill the requirements of the financial risk control domain.

To overcome the problem of the heterogeneous data, an Ontology related to poverty alleviation domain is constructed in [46]. This ontology is further used to create the nodes and edges of the knowledge graph. The visualization techniques are applied on the knowledge graph which helps in providing the results of the different queries related to the poverty alleviation. Bankruptcy Prediction Computational Model (BPCM) is presented in [47], which is used to perform the bankruptcy predictions of the financial institutes or the companies. Ontology of the Bankruptcy Prediction (OBP) is constructed to uniformly extract the data from different data sources and to utilize the financial data of the companies. Semantic Analysis Graph Database (SAGRADA) is created which consumes the OBP ontology, while the graph database is used for storing and visualization of the data.

### C. KNOWLEDGE GRAPH CONSTRUCTION

A knowledge graph (KG) is a semantic graph consisting of vertices (or nodes) and edges. The vertices represent concepts or entities. The edges represent the semantic relationships between concepts or entities [6]. By exploiting KG, partially observed entities and concepts can be connected

**TABLE 1. Graph database data structures [9].**

| | Graph Type | | | | Nodes | | Edges | | |
|---|---|---|---|---|---|---|---|---|---|
| *Graph Database* | *Simple* | *Hypergraph* | *Nested* | *Attributed* | *Node labeled* | *Node attribution* | *Directed* | *Edge labeled* | *Edge attribution* |
| *AllegroGraph* | ✓ | | | | ✓ | | ✓ | ✓ | |
| *DEX* | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *Filament* | ✓ | | | | ✓ | | ✓ | ✓ | |
| *G-Store* | ✓ | | | | ✓ | | ✓ | ✓ | |
| *HyperGraphDB* | | ✓ | | | ✓ | | ✓ | ✓ | |
| *InfiniteGraph* | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *Neo4j* | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *Sones* | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *vertexDB* | ✓ | | | | ✓ | | ✓ | ✓ | |

together to form a complete and structured knowledge repository [4].

For a knowledge graph, usually an ontology defines the architecture and constraints for the data residing in it. Ontology assisted Financial knowledge graph (KG) population deals with attaching the detected named entity with the correct label/category. When a named entity is detected from the unstructured text, which has no ontological mapping defined, right node category is sought in KG to attach the entity; this task is known as fine-grained named entity classification [23]. Otherwise, if the desired entity mapping exists in the ontology, the aim of this task is to link this detected entity mention with its corresponding real-world entity in the knowledge graph (KG), which is known as the entity linking task.

An extensive comparison survey of well-known graph databases was conducted by the authors in [9]. The results are summarized in Table 1. Owing to the powerful features and flexibility of Neo4j,[1] we selected it for our knowledge graph implementation.

Knowledge graph identification (KGI) is a technique for knowledge graph construction that jointly reasons about entities, attributes, and relations in the presence of uncertain inputs and ontological constraints [24]. Candidate facts from an information extraction system can be represented as an extraction graph; where entities are nodes, categories are labels associated with each node, and relations are directed edges between the nodes [25]. In another work [22], the authors studied financial documents for knowledge graphs population with financial entities and their interrelationships. They presented experimental results and discussed knowledge graph (KG) construction techniques on financial filings along with its challenges and possible solutions. Reference [43] utilizes the knowledge graph to represent the transactions data visually which helps in reducing financial frauds in the domain of financial risk control. Reference [44] claims that an Anti-TrustRank algorithm based on the knowledge graph data of the financial institutions can be used for Anti Money Laundering purpose also. The algorithm considers the web as a graph, pages and the link between the pages, as node and edges, respectively and it assists the financial institutions in finding the money launderers and helps to protect the financial institute from money laundering. Reference [45] proposed a financial news recommendation framework, based on NNR and INNR models, which uses the knowledge graph for the financial news recommendations. The edges of the graph are updated during the stock market trading through INNR and after the stock market closing through NNR. Both the models are then combined in the end to attain better, accurate, and efficient financial news recommendations. The idea of extracting financials news specifically related to the Chinese stock market from different Chinese encyclopedias and financial news websites is introduced in [49]. According to the author, the news will set the stock market sentiments. The ontology is created, and the financial knowledge graph is used to construct the relationship between the entities of the stocks from the financial news, which is further used to analyze and identify the impact of the news on different stock prices and the different possibilities of the stock risks involved in timely manner.

### D. QUESTION ANSWERING

Question answering system based on knowledge graph of Chinese classic poetry is proposed in [48]. The Chinese poetry related information is extracted from the classical Chinese poetry website and the knowledge graph is constructed on the basis of this data and stored in the database. The Rasa framework, as a natural language processing, is adapted to answer the queries of the user. A graph data-driven framework is proposed in [50], which provides the answers of the natural language questions using RDF graph repository. In the first step the authors have translated the natural language questions into SPARQL and then in the next step all the translated SPARQL's are evaluated by the system, which provides the answer of the question. IBM based researches proposed Question Answering system [26] to sequentially perform linguistic analysis of query, do named entity extraction, entity / graph search, fusion and ranking of possible answers. Our research is also following the similar approach.

### E. GAP ANALYSIS

Previous subsections identified research background in different related areas, which show that related works have been performed in parts by communities of Semantic Web and Information Extraction and Visualization. However, the existing research literatures have not provided any complete system that extends flat financial reports to machine readability for making it query-able. With the complex nature of these reports, the manual process of finding relevant information for investment is quite tedious and cumbersome task. This

[1] https://neo4j.com/product/

**TABLE 2.** Gap analysis.

| Paper | Year | Information extraction from Unstructured / Structured Financial Text | Entity Resolution /Entity linking/Relation extraction using Ontology/Rule-based/LOD | Construction of Financial KG | Ontology enrichment and KG identification | Story telling /Knowledge discovery using KG | Q/A using KG |
|---|---|---|---|---|---|---|---|
| [12] | 2018 | H | H | H | H | NH | H |
| [42] | 2018 | H | H | H | H | NH | NH |
| [22] | 2017 | H | NH | H | H | NH | NH |
| [28] | 2018 | NH | Ontology for top down+ LOD bottom-up | H | H | H | H |
| [24] | 2013 | H | Ontology | H | H | NH | NH |
| [1] | 2012 | H | NH | NH | NH | NH | NH |
| [21] | 2012 | H | Rule based for ER / Ontology for event detection | H | NH | NH | NH |
| [37] | 2013 | NH | Ontology | H | H | NH | NH |
| [26] | 2016 | NH | LOD (WikiData) | NH | H | H | H |
| [32] | 2018 | NH | NH | NH | NH | H | NH |
| [3] | 2017 | NH | NH | NH | NH | NH | NH |
| [23] | 2012 | H | Ontology | H | NH | NH | NH |
| [7] | 2017 | NH | NH | H | H | NH | H |

work is an attempt to integrate information extraction work with semantic web on Financial Reports using Knowledge Graph.

Table 2 shows the limitations of existing research on financial knowledge graph based financial query system. It depicts that very limited number of research have been done in this domain. In Table 2 , the H stands for Handled and NH stands for Not Handled. In most of the research papers, the information extraction part is missing because extracting the related information from the curated, heterogeneous, and complex data is still a challenging task. The construction of financial knowledge graph for the purpose of querying, visualization, and producing results and stories is also not commonly used by the researchers. The facility to get the answers of the users' questions and queries in the form of natural language is also an important factor. This factor is also not commonly provided by the researchers, as shown in Table 2.

By finding the above deficiencies, a novel approach in this domain is proposed for extracting financial information from different banks annual disclosures using ontology and store that in an efficient Financial Knowledge Graph (FKG) for future refinements, agility, and fact discovery. The graph can be queried for getting answers to user queries and will be able to generate user stories according to the needs of different users.

## III. METHODOLOGY

### A. PROJECT PHASES BREAKDOWN

We have divided the research work broadly in three phases. All the phases are discussed separately in this paper.

#### 1) DEFINING COMPETENCY QUESTIONS

For our knowledge graph (KG) to answer the competency questions, following steps were needed

- Find Information resources that can provide valuable information
- Gather Information
- Translate information in machine readable form
- Extract Director's report
- Study and analyze data set

#### 2) ONTOLOGY ENGINEERING

Keeping competency questions in mind from previous step, enumerate all the terms that should be available in ontology:

- Define Concepts/Classes
- Define Object properties and Data Properties
- Populate static instances into ontology like bank
- Names that will help in information extraction phase
- Define axioms/constraints

The sample competency questions are shown in Fig. 2.

#### 3) ONTOLOGY-BASED INFORMATION EXTRACTION

The steps involved in ontology-based information extraction are mentioned below:

- Study and analyze data set
- Define stop words
- Document preprocessing
- Extract all the terms from ontology that are known instances/entities and can be directly mentioned in the text along with its direct and indirect super classes. This will serve as a gazetteer list.
- Extract relationship names between two entities using object property of an ontological concept.
- Apply rule-based information extraction techniques with supporting information found in previous two steps. A set of rules were manually crafted and implemented to extract each target.

**SAMPLE COMPETENCY QUESTIONS**

- Name the bank having most after- tax profitability in year YYYY?
- Name the bank having most before-tax profitability in YYYY?
- Which bank paid the highest tax amount?
- Name the banks who's profitability has declined in YYYY compared from the previous year.
- Which bank had the highest dividend payout ratio in XXXX?
- Name the bank having highest capital reserve ratio in XXXX?
- Name the banks those are satisfying Capital Reserve requirement as set by SBP.
- Rate the bank in terms of the profitability and sustainability factors.
- Which banks have long term rating AAB or higher?
- Which banks have short term rating A1+ or higher?
- Who did Bank X audit in YYYY?

**FIGURE 2.** Sample competency questions.

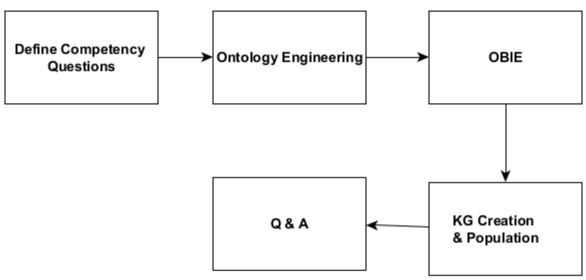


**FIGURE 3.** Phases of Construction of Financial Knowledge Graph from Banks' Annual Disclosures.

#### 4) KNOWLEDGE GRAPH CREATION

This phase is overlapped with the previous phase as the information extracted from text and ontology will help in knowledge graph (KG) creation that is Knowledge graph population with appropriate nodes, relationships, and labels.

#### 5) QUESTION/ANSWERING AND STORY TELLING

This phase involves validating if the knowledge graph can answer the queries well. It involves extracting target entities from the query, convert it into a Graph querying statement, extract results from graph and display the results.

The different phases involved in the Construction of Financial Knowledge Graph from Banks' Annual Disclosures are shown in Fig. 3.

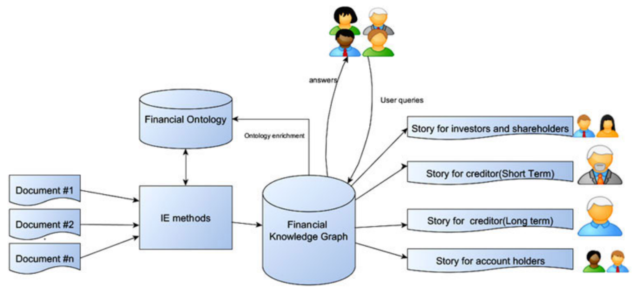


**FIGURE 4.** Proposed system architecture.

#### 6) KNOWLEDGE GRAPH ENRICHMENT

This occurs when stakeholder asks for a piece of information and it is not mapped into entity. Ontology will be updated in that case and documents will be rescanned to enrich knowledge graph (KG) with newer nodes and relationships. In case the information is new, its attribute will be analyzed and it will be added in existing financial knowledge graph (KG).

### B. DATA SOURCES

Primarily, we ingest information from financial filings; however, further data sources can add value to knowledge graph (KG) enrichment. We have identified following three types of potential data sources [52].

#### 1) SEMI-STRUCTURED

- Annual Report.
- Longer company profiles.
- Imprint information on company web pages.
- Running tickers on company information.

#### 2) STRUCTURED SOURCES

- Publicly available balance sheets in structured format.
- Short company profiles (e.g. from Business Registers, Stock Exchange, web pages, etc.)n
- Wikipedia Infoboxes.

#### 3) UNSTRUCTURED

- Annexes to balance sheets in annual reports of companies.
- Newspapers.
- Specialized web pages etc.

### C. PROPOSED SYSTEM ARCHITECTURE

The proposed System architecture, which presents how the user queries are processed and how the system will generate the results, is shown in Fig. 4.

### D. POTENTIAL USERS

The proposed system will benefit investors, creditors, external agencies, regulators and account holders for decision making as shown in Fig. 5.

## IV. ONTOLOGY DEVELOPMENT

### A. INTRODUCTION

Ontology Based Information Extraction (OBIE) is exploited in this research. Ontology develops a shared understanding

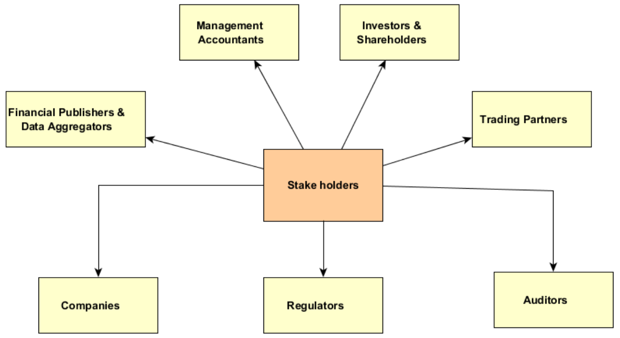


**FIGURE 5.** Stakeholders of the proposed system.

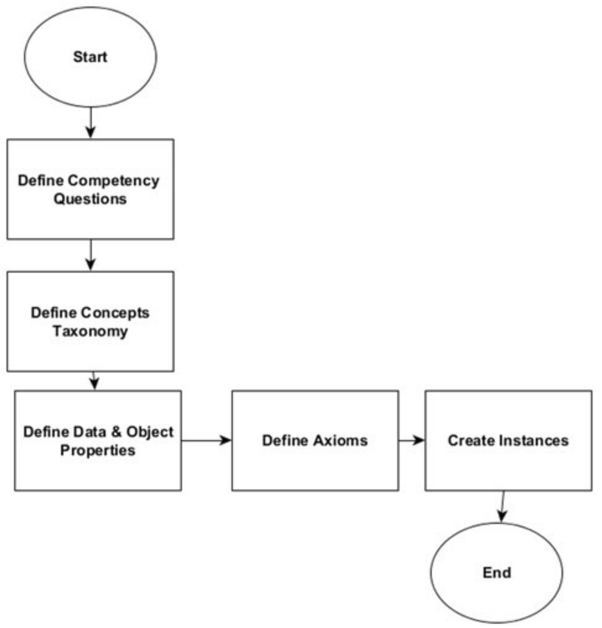


**FIGURE 6.** Ontology engineering process.

of domain by building common vocabulary. Ontology comprises of concepts, instances, object/data properties, and sometimes other existing ontologies [20].

There are several tools available for ontology engineering. In this work, we have used Protégé,[2] being user friendly and widespread tool for editing and developing ontologies. It hides the underlying complexity of domain modeling and enables users to focus on the domain knowledge in terms of real-world entities, inter-entity relationships, and constraints. Protégé ontologies can be effortlessly exported into multiple formats, such as, Resource Description Framework (RDF), Turtle, and Web Ontology Language (OWL). In addition to this, we have used VOWL and Onto Graf plugins to visualize ontology, and SPARQL query language for retrieving data from the ontology. We have used OWLAPI to import and manipulate our ontology in Eclipse[3] (Java IDE). The different steps involved in the ontology engineering process are shown in Fig. 6.

MCB[4] Bank Limited reported Profit Before Tax (PBT) of Rs.31.01 billion and Profit After Tax (PAT) of Rs. 22.46 billion. In comparison with the last year, Profit Before Tax has decreased by 14.03% whereas Profit After Tax has increased by 2.59% on account of reversal of prior year tax charges. Net markup income of the Bank was reported at Rs. 42.41 billion, down by 3.21% over last year owing to the maturity of high yielding bonds and comparative low-interest rate environment. On the gross markup income side, the Bank reported an increase of Rs. 6.69 billion whereas on the interest expense side, the Bank registered an increase of Rs. 8.09 billion over last year. To supplement its net interest margins, the Bank remained focused on increasing its low- cost deposit base and venture in higher-yielding assets. The Board of Directors declared a final cash dividend of Rs.4 per share for the year ended December 31, 2017, which is in addition to Rs. 12.0 per share interim dividends already paid to shareholders, taking the dividend payout ratio to 83.14%. The effect of the recommendation is not reflected in the above appropriations.

**FIGURE 7.** Financial information extracted from MCB bank limited annual report [53].

### B. DATA SET

For domain modeling, annual reports are collected from online repository[5] published by banks across the globe in English Language. The concepts generated from these reports are modeled as tuples to help in guided information extraction of meaningful patterns. The ontological dataset is then applied on two major commercial banks of Pakistan as proof of concept towards adaptability.

The related financial information is extracted from MCB Bank Limited Annual report 2017, which contains the brief information about the bank's financial information, shown in Fig. 7.

The related financial information is extracted from United Bank Limited Annual report 2017, which contains the brief information about the bank's financial information, shown in Fig. 8.

### C. DEFINE DOMAIN AND SCOPE

Keeping in view the data available in the Annual report, competency questions from dataset are used to define ontological concepts/properties and limit system scope.

- Name the bank having most after- tax profitability in year YYYY?
- Banks whose Market Price per Share @Year Start > @YearEnd OR whose stock price has increased this year
- Which bank pays the highest dividend to his stockholders?
- List banks whose stock/share market price increased over last three years?

[2]Protégé https://protege.stanford.edu/
[3]Eclipse https://www.eclipse.org/
[4]MCB https://www.mcb.com.pk/
[5]https://www.annualreports.com/

UBL[6] posted profit after tax (PAT) amounting to Rs. 25.4 billion during the year ended December 31, 2017 compared to Rs. 27.7 billion in 2016. Earnings per share were reported at Rs. 20.77 in 2017 against Rs. 22.65 per share last year. Profit before tax (PBT) closed at Rs. 40.2 billion in 2017 compared to Rs. 46.0 billion in 2016. The consolidated PAT stood at Rs. 26.2 billion in 2017 (2016: Rs. 28.0 billion) with earnings per share recorded at Rs. 21.39 (2016: Rs. 22.70). Gross revenues stood at Rs. 78.6 billion (Dec'16: Rs. 80.7 billion). Despite the low interest rate regime, growth in the balance sheet maintained net markup income in line with the 2016 level to close at Rs. 56.4 billion. Non-markup Income decreased by 6% year on year to reach Rs. 22.2 billion in 2017 but mainly due to lower capital gains and dividends. The cost to income ratio increased from 39.6% in 2016 to 45.0%.

**FIGURE 8. Financial information extracted from united bank limited annual report [54].**

- Name the bank having most before-tax profitability in YYYY?
- Which bank paid the highest tax amount in year YYYY?
- Which bank had the highest dividend payout ratio?
- Name the bank having highest capital reserve ratio in YYYY.
- What is Profitability After Tax of XXXX(bank name) in YYYY? What was the long term rating of ABL in YYYY?
- How many banks have rating "Extremely Strong" in a long term?
- Will UBL default in next year?
- XXX audited which banks this year?
- How many banks were given Extremely Strong long term rating this year by PACRA?
- Which banks have stable profitability ratios over a period of time?
- Who won Best Local bank award in YYYY?
- Who gave Best Investment Bank Award in YYYY?
- Which banks' stock are safer to invest in?
- EPS > 0 in last year
- Long term rating Extremely Strong or Strong
- Short term rating High
- Current year Stock price > Last year stock price
- Which banks stocks are risky but yield is higher
- Long term rating NOT(Extremely Strong or Strong)
- Short term rating NOT High
- (Current year Stock price - Last year stock price)/ Last year stock price > some threshold percentage
- Which bank Net Interest Income has increased this year?
- Does MCB has Islamic Window?
- What is the currency of USD bond?
- Which bonds are foreign government bonds?
- What type of dividend did bank XXX gave to its stock holders in YYYY?
- Did UBL invested in Manufacturing Sector in YYYY?

[6]UBL https://www.ubldigital.com/

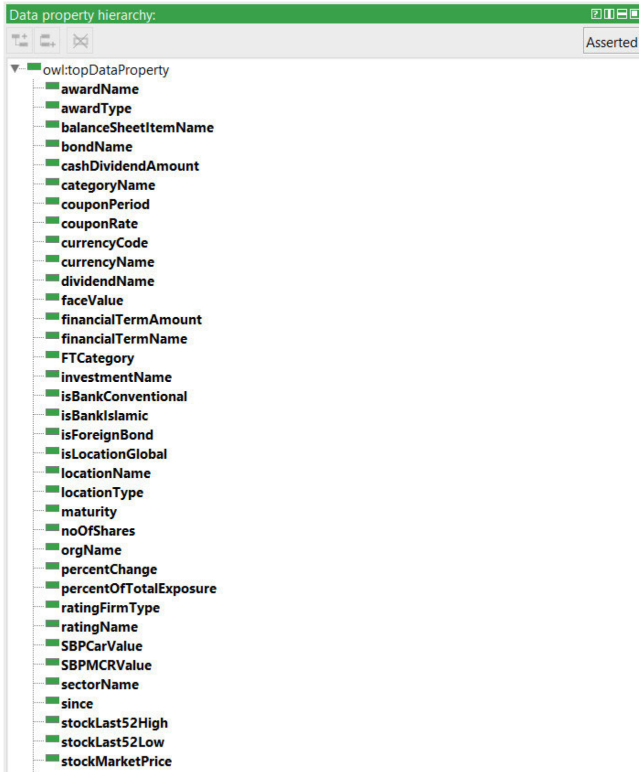


**FIGURE 9. Data properties.**

- MCB invested in which Government bonds this year?
- What was UBL total investment in Bonds in YYYY?

### D. CONCEPTS/CLASS HIERARCHY

The detailed information regarding classes, description regarding classes, parent class and instances are shown in Table 3.

### E. DATA AND OBJECT PROPERTIES

Object Properties will serve as relationships in the KG and data properties will serve as Entity attributes. The details are shown in Fig. 9.

### F. OWL API FOR ONTOLOGY IMPORT AND QUERYING

SPARQL is used for querying Ontology. All the queries were first validated in Protégé then were used in Java for information extraction from Ontology. Fig. 10 shows the SPARQL Query for retrieving valid Bank names which are instances of Concept "Bank" and Fig. 11 shows Ontology.

## V. INFORMATION EXTRACTION FROM ANNUAL REPORTS

Information Extraction (IE) deals with automatic retrieval of certain types of information from natural language text. It aims to retrieve occurrences of a particular class of objects and identify relationships among them [19]. Once the text corpora is developed manually, next phase is entity extraction/recognition and relationship extraction/prediction [27]. We are using ontology for entity recognition, entity alignment

**TABLE 3.** Class hierarchy.

| Class | Parent Class | Description | Instances |
|---|---|---|---|
| Award | Award | Award given to a bank by an award giving organization | PBA |
| BalanceSheetItem | BalanceSheetItem | Items present in a Balance Sheet | Cash, Interest payable, Property, Tax payable |
| Bond | Bond | Types of Bonds | ASKARI BANK LTD. - TFC, ASPIN PHARMA (PVT) LTD - SUKUK, BANK AL-HABIB LTD. - TFC, JS BANK LTD. - TFC, K-ELECTRIC LTD. - SUKUK, MASOOD TEXTILE MILLS LTD. - SUKUK, National Saving Bonds, Wapda Bond |
| Category | Category | Categories of Banks | Bank Islami, Best Bank Award, DIB, HBL, Khushali Microfinance Bank, Pakistan Banking Awards |
| Currency | Currency | Currency of different countries | AUD, GBP, PKR, USD |
| Dividend | Dividend | Profit distribution of the companies to its shareholders | Bi Yearly Dividend, Bonus Shares, Quarterly Dividend, Right Shares, Yearly Dividend |
| Financial Term | Financial Term | Different vocabularies and terminologies commonly used in the financial market | |
| Investment | Investment | Spending money for generating income | Amazon, Apple, Gold, MCB-DCF Income Fund, PIB, Pakistan Income Fund, Real Estate, Wapda Bonds |
| Location | Location | Locations of the Financial Institutions / Banks | France, Germany, Pakistan, UK, USA |

**TABLE 3.** *(Continued.)* Class hierarchy.

| | | | |
|---|---|---|---|
| Rating | Rating | Ratings awarded by the Rating Firms | A, A-, A-1, A-1, A-2, A-3, A1, A2, A3, AA, AA-, AAA, Aa1, Aa2, Aa3, Aaa, BBB, BBB-, Baa1, Baa2, Baa3, F1, F2, F3, P-1, P-2, P-3 |
| Organization | Organization | Financial Institutions / Organizations | ABL, BankIslami, DIB, Deloitte, Dubai_Islamic_Bank, EY, Fitch, HBL, IBP, JCR-VIS, KPMG, MCB, Moody's, PACRA, PwC, S&P, UBL |
| Sector | Sector | Group of companies that operate in the same segment of the economy or share a similar business type | Abbott Laboratories Pak Ltd., Allied Bank Ltd., Buxly Paints Ltd., Fauji Cement Co Ltd., Fauji Fertilizer Co. Ltd., Habib Bank Limited., Highnoon Laboratories Ltd., KAPCO, PIA, Sanofi-Aventis Pakistan Ltd. |
| Stock | Stock | Ownership certificates of any company | |
| Asset | BalanceSheetItem | Resources owned by the company | Cash, Property |
| Liabilities | BalanceSheetItem | Debt or obligations of the company | Interest payable, Tax payable |
| CorporateBond | Bond | Corporate Bond is a debt security which is issued by company and sold to investors to meet its financial requirements. | ASPIN PHARMA (PVT) LTD - SUKUK, K-ELECTRIC LTD. - SUKUK, MASOOD TEXTILE MILLS LTD. - SUKUK |
| GovtBond | Bond | Government bond is a debt security loaned by a government to assist government spending, most often issued in the country's local interest. | National Saving Bonds, Wapda Bond |

**TABLE 3.** *(Continued.)* Class hierarchy.

| | | | |
|---|---|---|---|
| TermFinanceCertificate | Bond | Certificates issued by the companies for the generation of short and medium term funds. | ASKARI BANK LTD. - TFC, BANK AL-HABIB LTD. - TFC, JS BANK LTD. - TFC |
| AwardCategory | Category | The different categories awarded on the basis of the performance and efforts in different sectors | Best Bank Award, Pakistan Banking Awards |
| BankCategory | Category | The different categories of the bank | Bank Islami, DIB, HBL, Khushali Microfinance Bank |
| Cash_Dividend | Dividend | Funds paid to shareholders in the form of cash | Bi Yearly Dividend, Quarterly Dividend, Yearly Dividend |
| Stock_Dividend | Dividend | Funds paid to shareholders in the form of stock | Bonus Shares, Right Shares |
| Earnings_Per_Share | Financial Term | Earnings per share indicates the company's financial position in the market | |
| EPS | Financial Term | Earnings per share indicates the company's financial position in the market | |
| Gross_Markup_Income | Financial Term | The revenue generated after eliminating the cost | |
| Interest_Expense | Financial Term | The cost of borrowing money from financial institutions | |
| InterestIncomeEarned | Financial Term | Amount earned by an investor's money that he places in an investment or project. | |
| Net_Interest_Income | Financial Term | Difference between the interest incomes a bank earns from its lending activities and the interest it pays to depositors. | |
| NII | Financial Term | Difference between the interest incomes a bank earns from its lending activities and the interest it pays to depositors. | |
| PAT | Financial Term | Profits after payment of tax | |
| PBT | Financial Term | Profits before payment of tax | |
| Profit_Before_Tax | Financial Term | Profits before payment of tax | |

**TABLE 3.** *(Continued.)* Class hierarchy.

| | | | |
|---|---|---|---|
| ProfitAfterTax | Financial Term | Profits after payment of tax | |
| BondInvestment | Investment | Investment in the form of Bonds | PIB, Wapda Bonds |
| MutualFundInvestment | Investment | Investment in the Mutual Funds | MCB-DCF Income Fund, Pakistan Income Fund |
| SectorInvestment | Investment | Investment in the different sectors of the economy | Gold, Real Estate |
| StockInvestment | Investment | Investment in the different stocks of the companies | Amazon, Apple |
| AuditFirm | Organization | Audit firm investigates frauds, deficiencies in the organization | Deloitte, KPMG, PwC |
| AwardFirm | Organization | Award Firm reward and recognizes the company's efforts | IBP |
| Bank | Organization | Financial institute which accepts deposits and provide loans to the customers | ABL, BankIslami, DIB, HBL, MCB, UBL |
| RatingFirm | Organization | Credit rating agency is an independent enterprise that evaluates the financial standing of issuers of debt instrument and then assigns a rating that exhibits its assessment of the issuer's aptitude to make the debt payments. | Fitch, JCR-VIS, Moody's, PACRA, S&P |
| LongTerm | Rating | Long Term rating have the maturity of one year or more | A, A-, A1, A2, A3, AA, AA, AA-, AAA, Aa1, Aa2, Aa3, Aaa, BBB, BBB-, Baa1, Baa2, Baa3 |
| ShortTerm | Rating | Short Term rating have the maturity of one year or less | A-1, A-2, A-3, F1, F2, F3, P-1, P-2, P-3 |
| Banking | Sector | | Allied Bank Ltd., Habib Bank Limited. |
| Infrastructure | Sector | | KAPCO, PIA |
| Manufacturing | Sector | | Buxly Paints Ltd., Fauji Cement Co Ltd., Fauji Fertilizer Co. Ltd. |
| Pharmaceutical | Sector | | Abbott Laboratories Pak Ltd., Highnoon Laboratories Ltd., Sanofi-Aventis Pakistan Ltd. |

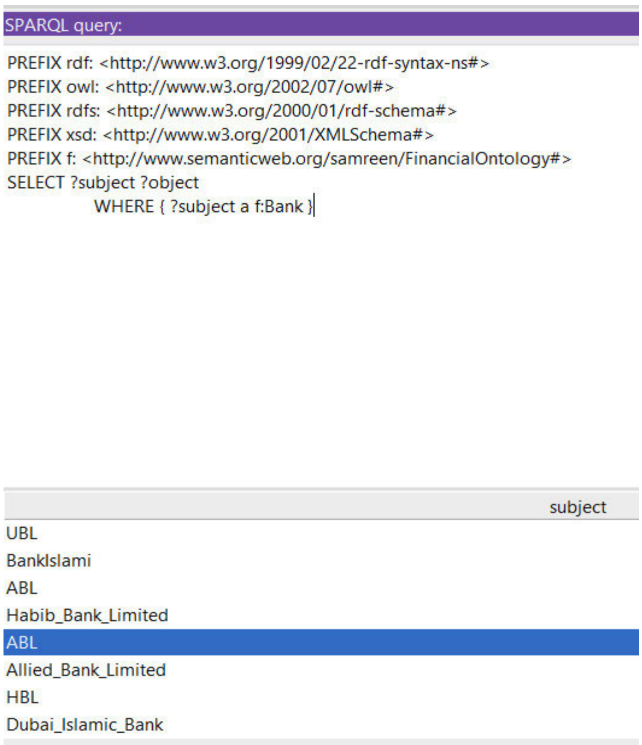


**FIGURE 10.** **SPARQL for retrieving bank names in Ontology.**

and relation extraction. Entities will be stored into RDF/OWL based knowledge graph for the sake of efficiency, scalability and flexibility [22].

In order to construct a knowledge graph, information extraction is critical for its correctness. Relevant information from unstructured text corpora is extracted and mapped to some pre-defined knowledge graph concept, consequently, the structural relationships are defined between extracted entities. Following sections discuss the detailed approach used for information extraction.

### A. BASIC PROCESSING

In this phase, some basic processing will be required to convert this character stream into a sequence of lexical items (words, phrases, and syntactic markers) to further consume information. Each document is passed through sentence splitter and tokenizer. Sentence splitter splits the sentence using some delimiter and tokenizer chops input character streams into tokens that can be words, numbers, identifiers or punctuation (depending on the problem).

### B. TEXT PRE-PROCESSING

The second phase includes Named Entity Recognition and co-reference resolution. The rule-based processing is employed to recognize the named entities alongside a gazetteer to hunt the overall sorts of entities (relation terms, stop words etc.) and ontology to seek out domain-specific entities like bank names, Rating Agencies, Award Names and Financial terms [21].

### C. ENTITY RESOLUTION

We have created different Classes with "Same as" attribute for detecting bank name and term variations in the text as United Bank Limited may be written as UBL or United Bank Ltd. Therefore, in our knowledge graph all will be considered the same.

### D. RELATION EXTRACTION

It deals with extracting relation between two entities detected by NE Recognizers with the help of additional annotation.

The core of Entity and relation extraction is a hand-crafted collection of rules. The patterns are generated through text analysis and represent the unique language constructions which are used to describe a particular Entity/Relation. These patterns are then matched with processed text to discover and extract required pieces of information [21].

Our system knows the bank names and financial terms from the ontology, therefore, they are defined as entities of the type Bank and financial term in the named entity recognition step. The currency amount "Rs. 25.4 billion" is classified on the same step and is given the Money annotation. The rule "Bank_ profit after tax _Amount" is triggered as the relation extraction phase begins, as its pattern is a perfect match for the input sentence, as shown in Fig. 12. The time of the mentioned sentence will be stored in knowledge graph (KG) on relation attribute [21]. The output of this phase are semantic triplets with additional information of direct and indirect super classes of the extracted entities and attributes related to each triplets (like YEAR in above example) to be added in the knowledge graph (KG). This project is different in terms of Entity types, therefore, standard IE libraries may not be applied directly. Below are category-wise items and the respective technique employed.

## VI. INFORMATION EXTRACTION FROM ANNUAL REPORTS

### A. OVERVIEW

Relational databases are so powerful and well understood yet still carry many limitations when it comes to efficient storage, scalability and efficient query processing where several joins are needed to get a specific piece of information. These limitations pushed researchers to develop alternative database technologies known as NOSQL databases [9]. These databases can be categorized on the basis of underlying data models like 1)Wide-column stores uses Google's BigTable model (e.g., Cassandra) 2) Document stores are designed to store semi-structured data (e.g., MongoDB) 3) Key-value stores maintains a key to value persistent map for data indexing and retrieval (e.g. BerkeleyDB); and 4) Graph Databases that store information in a graph-like data structure.

### B. KNOWLEDGE GRAPH CONSTRUCTION

Knowledge graph can be constructed using 1) top-down approach based on some knowledge base/ schema such as

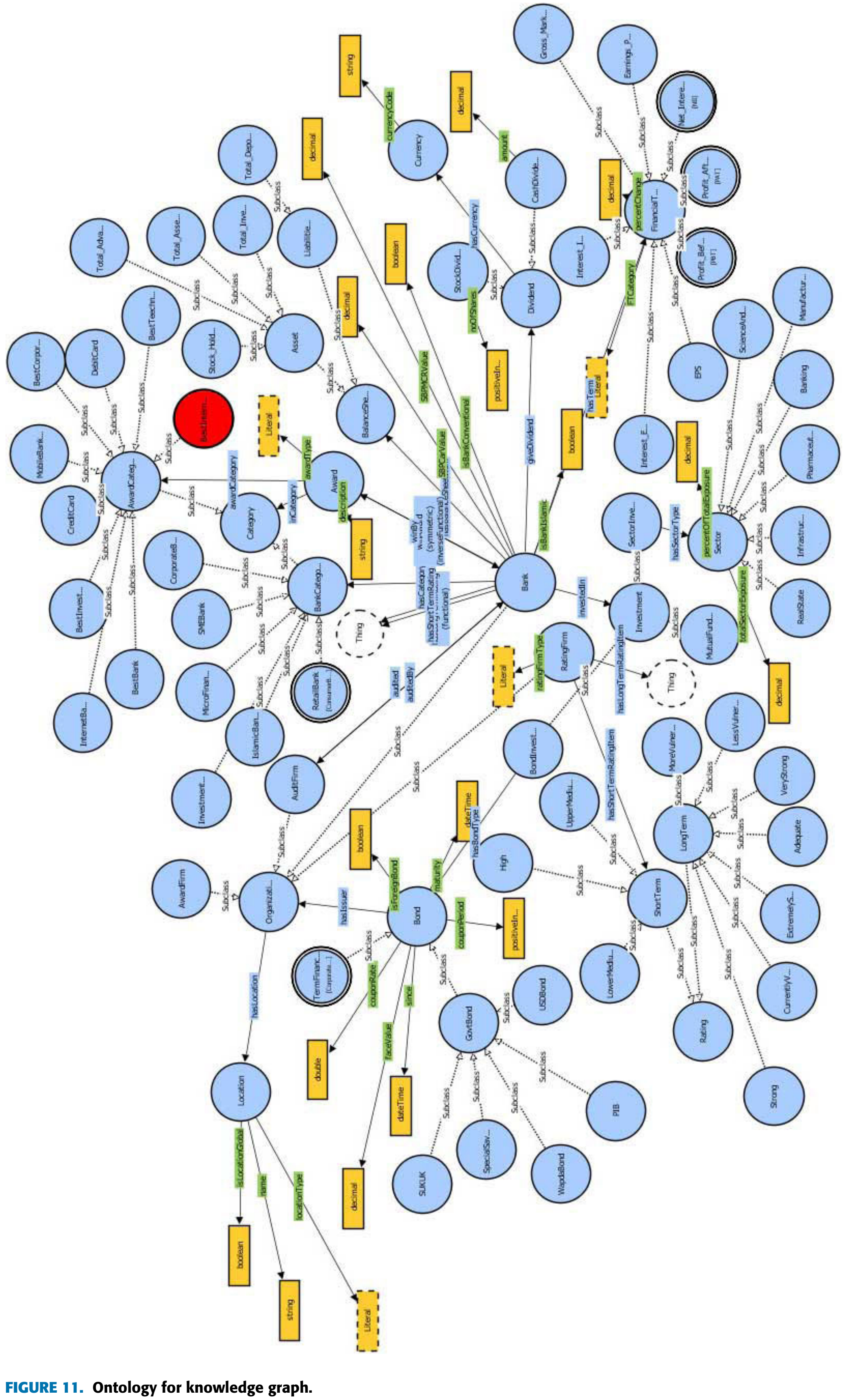


**FIGURE 11.** Ontology for knowledge graph.

```
"UBL posted profit after tax amounting
to Rs. 25.4 billion during the year ended
December 31, 2017"
Rule: Bank_ profit after tax _Amount
({Organization}): Bank
({Token, !Split})*
{Token.root == "profit after tax"}
({Token, !Split, ! Money })*
 ({Money }):Amount
```

**FIGURE 12.** SPARQL for retrieving bank names in Ontology.

**TABLE 4.** Information extraction techniques used.

| Extraction Type | Information Type | Technique applied |
|---|---|---|
| Entity Extraction | Amount | Regular expression |
| Entity Extraction | Percentages | Regular expression |
| Entity Extraction | Dates (Year only) | Regular expression |
| Entity Extraction | Bank | Ontology based extraction |
| Entity Extraction | Financial Term | Ontology based extraction |
| Entity Extraction | Ratings | Ontology based extraction |
| Entity Extraction | Audit Firm | Ontology based extraction |
| Entity Extraction | Award | Ontology based extraction |
| Entity Extraction | Balance Sheet Terms | Ontology based extraction |
| Entity Extraction | Currency | Ontology based extraction |
| Entity Extraction | Location | Ontology based extraction |
| Relation Extraction | Verbs stored in ontology as relationships | Hand Crafted Rules/Rule based |

the domain ontologies or 2) bottom-up approach focusing on knowledge instances such as Linked Open Data (LOD) datasets [28]. As we are using top-down approach, we developed the ontology in advance. Information extraction and Knowledge graph population is overlapped in our case. As the information is extracted from the text it is added into knowledge graph (KG) along with additional annotations like super classes of extracted entities.

We have used Neo4j for implementing our Graph Database, although it is not a pure knowledge graph (KG)

```
Location
CREATE
(l:Location{locationName:'Pakistan',locationType:"Country"
,isLocationGlobal:false})
CREATE CONSTRAINT ON (l:Location) ASSERT
l.locationName IS UNIQUE
Rating
CREATE(r:Rating:LongTerm:ExtremelyStrong{ratingName:'
AAA'})
Bank
CREATE(r:Organization:Bank{orgName:'HBL',isBankIslamic
:true,isBankConventional:true})
CREATE CONSTRAINT ON (a:Organization) ASSERT
a.orgName IS UNIQUE
FinancialTerm
CREATE(r:FinancialTerm:Profit_Before_Tax{financialTermN
ame:'ProfitAfterTax',financialTermAmount:1015000000,pe
rcentChange:4}) return id(r)
```

**FIGURE 13.** Sample script for node creation based on Ontology.

in a real sense but it provides the structure and API for the proof of our concept.

### C. INSTALLING AND CONFIGURING NEO4J

Neo4j is a browser based Graph Database which has API support for many languages. We have used Neo4j API for manipulating the database in Eclipse – JAVA IDE. The Ne4j server needs to be started before running any commands either on browser or from a program like any Database Server.

### D. CREATING NODES

For node creation, Ontological information will be extracted for creating Node labels (Class/Category). Similarly, data properties of a Concept in Ontology will become property of the Nodes. For Example, HBL is a Bank that is also an Organization, therefore, HBL will have two Node Labels; Organization and Bank. A bank may be situated on multiple locations therefore, hasLocation relationship will connect the Bank to multiple locations which are entities as well. The nodes creation examples are mentioned in Fig. 13.

### E. CREATING EDGES

Object Property names from ontology will become relationship names in the Financial knowledge graph (KG) with domains and ranges of the concept as target entities types. The link attributes will be populated using the information extracted from the text. Fig. 14 shows queries to establish the relationship between entities created in the last step.

```
hasLocation
MATCH (b:Bank{orgName:'HBL'}) ,
(l:Location{locationName:'Pakistan'}) CREATE (b)-
[h:hasLocation]->(l) return h
hasLonTermRating
CREATE
(r:Rating:LongTerm:ExtremelyStrong{ratingName:'AAA'})
CREATE CONSTRAINT ON (l:Rating) ASSERT l.ratingName IS
UNIQUE
hasTerm
MATCH(i:FinancialTerm)
where ID(i)=3224 with(i) match(b:Bank{orgName:'HBL'})
MERGE (b)-[:hasTerm{year:2017}]->(i)
```

**FIGURE 14. Sample script for relationship creation based on Ontology.**

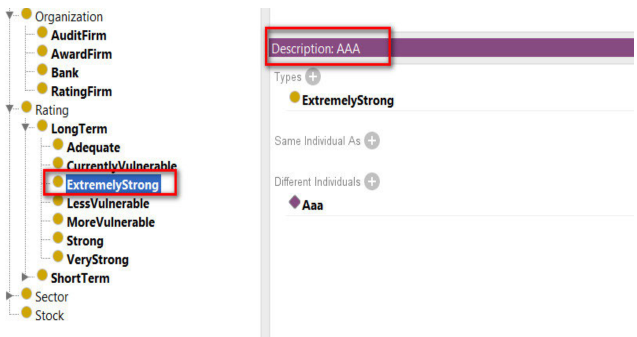


**FIGURE 15. Generating labels from the concept taxonomy.**


**FIGURE 16. Enumeration defined as slots for KG attribute values.**

Fig. 15 shows the labels generated from the concept taxonomy.

Fig. 16 shows the Enumeration defined as Slots for Knowledge graph attribute vales.

Fig. 17 shows the Node with attributes having multi-labels. The highlighted area with arrow sign in the Fig. 18 shows the Relationship of Node with the Attribute. The relationship name is mentioned within the arrow sign.

### F. ABBREVIATIONS AND ACRONYMS

Knowledge graph can never be complete as real world's formalized knowledge cannot reasonably reach full coverage, it contain information about each and every entity in the universe. Furthermore, it is nearly impossible to construct a knowledge graph which is fully correct, especially when

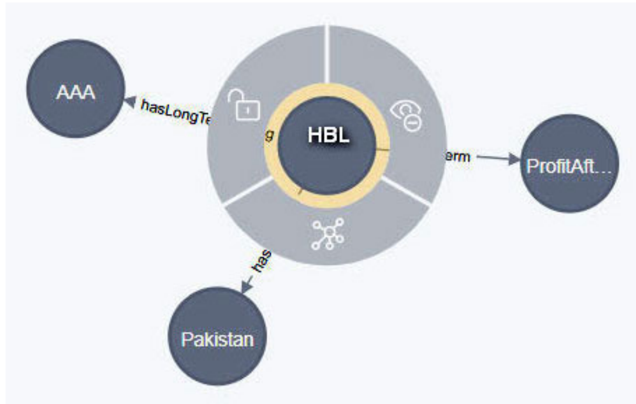


**FIGURE 17. Node with attributes having multi-labels.**

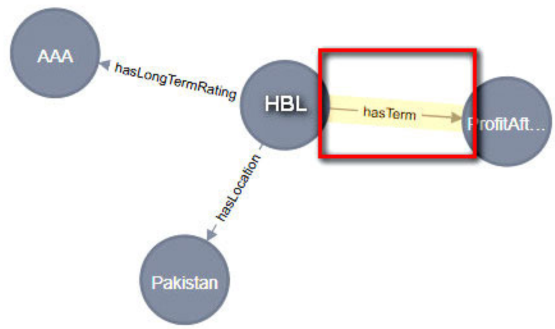


**FIGURE 18. Relationship with attribute.**

heuristic methods are applied. The trade-off between coverage and correctness is handled differently in each knowledge graph [29]. Knowledge graph refinement improves an existing knowledge graph like adding missing knowledge or identifying/removing errors. Logical reasoning is applied on some knowledge graphs for validating the consistency of statements in the graph, and removing the inconsistent statements.

In this research work, when new resources are added for Information Extraction in future, existing ontology works with an extension, if new taxonomy/property/relationships values are defined. If only data is of newer format/type then it can be directly ingested into knowledge graph (KG).

We need to apply rule-based extraction or statistical methods to get to know the information pattern based on relationship between target information with some other entity. If we can infer the entity type through its pattern (Information categorization/Entity Type recognition), the information will be appended in the Financial ontology. All the uploaded documents will be rescanned, and knowledge graph (KG) will be populated with the term related data values. In case of some new information category or entity property values, for now, ontology will be updated manually.

## VII. QUERYING OVER KNOWLEDGE GRAPH

It is difficult to extract who won the BEST INVESTMENT BANK AWARD in 2018 from ontology as winAward relationship does not specify the year in which an award was won

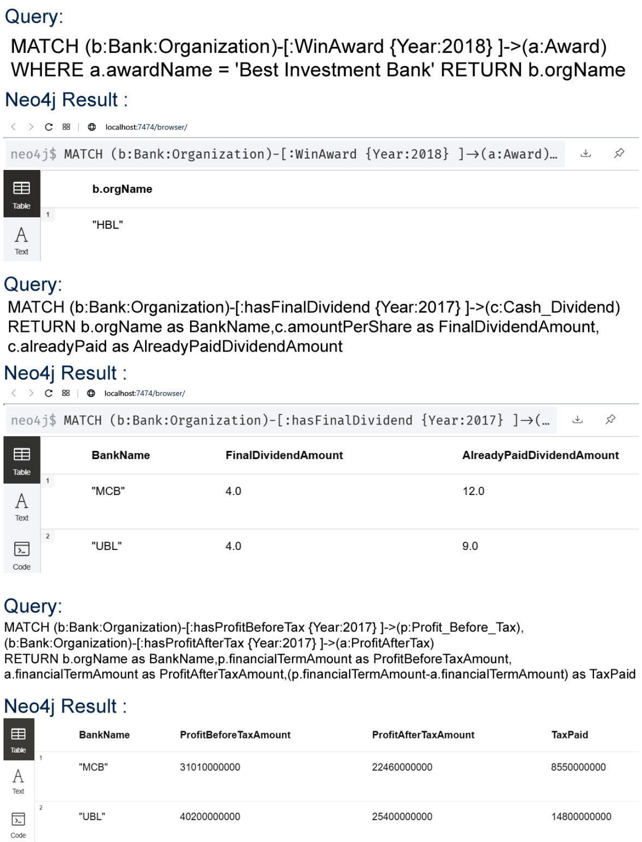


**FIGURE 19.** Example cypher queries.

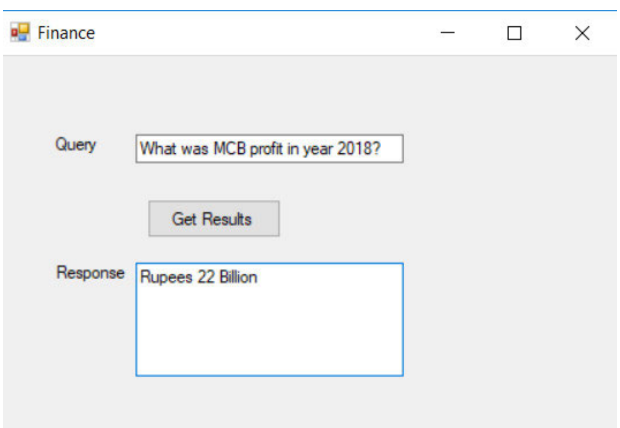


**FIGURE 20.** Sample interface for QA.

by a bank or we need to define a separate class in Protégé to maintain the relationship between AWARD and BANK for keeping track of the year as well. Whereas, it is very simple in knowledge graph to directly query. Similarly, the information about the final dividend and already paid dividend amount of all the banks or a particular bank or the list of highest dividends paid banks can also be available through a simple knowledge graph query. The information related to the profit before tax and after tax or list of the banks, who paid the highest tax during a particular year can also be obtained easily. The examples of different cypher queries with their results are shown in Fig. 19.

The multi-labeling and Edge Property is utilized in the above queries. In this work, similar approach is followed as followed in [26]. Presently, we have developed the interface in Visual Studio 2012 for querying Financial knowledge graph (KG) as displayed in Fig. 20.

**TABLE 5.** Stakeholder type and questions.

| S.No | Stakeholder Type | Question |
|---|---|---|
| 1 | Investor/Prospective Stock holder | Which banks' stock are safer to invest in? |
| 2 | Investor/Prospective Stock holder | Which bank received Best Investment Bank Award in YYYY? |
| 3 | Investor/Prospective Stock holder | Which bank pays the highest dividend to his stockholders? |
| 4 | Investor/Prospective Stock holder | What is EarningsPerShare of Bank XXX? |
| 5 | Creditors/Rating Firms | Name the bank having most after-tax profitability in year YYYY? |
| 6 | Investor/StockHolders | Banks whose Market Price per Share @Year Start > @YearEnd OR whose stock price has increased this year |
| 7 | Investor/StockHolders | Which bank pays the highest dividend to his stockholders? |
| 8 | Investor/StockHolders | List banks whose stock/share market price increased over last three years? |
| 9 | Creditors/Rating Firms | Name the bank having most before-tax profitability in YYYY? |
| 10 | Creditors/Rating Firms | Which bank paid the highest tax amount in year YYYY? |
| 11 | Creditors/Rating Firms | Name the banks who's profitability has declined in YYYY compared from the previous year. |
| 12 | Investor/StockHolders | Which bank had the highest dividend payout ratio? |
| 13 | Creditors/Rating Firms | Name the bank having highest capital reserve ratio in YYYY. |
| 14 | Creditors/Rating Firms | What is Profitability After Tax of XXXX(bank name) in YYYY? |
| 15 | Creditors/Rating Firms/Investor/StockHolders | What was the long term rating of ABL in YYYY? |
| 16 | Creditors/Rating Firms/Investor/StockHolders | How many banks have rating "Extremely Strong" in a long term? |
| 17 | Creditors/Rating Firms/General Public | Will UBL default in next year? |
| 18 | Creditors/Rating Firms/Investor/StockHolders | Which banks have stable profitability ratios over a period of time? |
| 19 | Creditors/Borrowers | Which bank Net Interest Income has increased this year? |
| 20 | Depositor | Does MCB has Islamic Window? |
| 21 | General public | Which bonds are foreign government bonds? |
| 22 | Investor/StockHolders/ Creditors | Did UBL invested in Manufacturing Sector in YYYY? |
| 23 | Investor/StockHolders | MCB invested in which Govt bonds this year? |
| 24 | Investor/StockHolders | What was UBL total investment in Bonds in YYYY? |
| 25 | General public | HBL is located in which countries? |

1) NAMED ENTITY (NE) AND RELATION EXTRACTION

We follow this step for user queries as done for financial knowledge graph (KG) generation.

2) ANCHORING

User vocabulary may differ from terminologies used in the KGs, this gap is filled by semantic expansion. Here, we do this by using Ontology – "Equivalent Classes".

3) GRAPH PATTERN SEARCH

This deals with mapping the extracted information from query to a Graph pattern search query for getting results. We have written rules to translate user requirement into a formal query that can be executed against the knowledge graphs.

4) MERGING OF RESULTS

Some queries are complex enough and need merging the results of multiple Cypher queries to give an answer to the user.

### A. CATEGORIZATION OF QUERIES WITH STAKEHOLDERS' PERSPECTIVE

The summary table contains the stakeholder type, and the related questions, which are mentioned in Table 5.

## VIII. CONCLUSION AND FUTURE WORK

This research proposed a novel approach for data extraction from Bank's Annual reports for the population of Financial Knowledge graph. We discussed the techniques used for information extraction, Ontology engineering procedure, Financial Knowledge graph creation, and Question Answering mechanism for the graph.

In most of the countries, financial regulatory body enforces companies to publish their annual reports online. In our research, in spite of availability of the required dataset in this report, the format required extensive efforts due to the PDF format. The report had multiple sections and automatic extraction needed much effort and time. Additionally, entity identifiers and formats for the same type of data some-times differ between organizations. Rule-based extraction is powerful but to handle multiple unforeseen patterns, our model should use some other approach like statistical learning, FIBO, etc. Also, to limit the research scope and generate a working prototype, we have focused on director's statement.

The result shows that our proposed system based on financial knowledge graph, successfully provides the desired information against the queries of the general public or investors related to the investment in different banks or for the particular bank, efficiently and smoothly.

In this research only banks are considered; however, this model can be extended to cater information of other companies in future. This research can be applicable in variety of domains like, healthcare, travelling, fuel companies, production companies, etc. where the annual reports are published but their customers and end users are unable to query or gather the desired information easily.

Additionally, for making graph more powerful and useful, our aim is to design a web crawler to get several other financial factors like, stock listings from authentic web pages, generalization of the model for companies other than banks, incorporate automatic entity linking and disambiguation mechanism for knowledge graph (KG) enrichment, financial Story generation for different type of stakeholders that can aid decision making, exploit the flexibility of open Information Extraction systems using unsupervised learning or semi-supervised learning, and use of readily available financial ontologies.

## ACKNOWLEDGMENT

The authors would like to thank the financial banks of local and foreign regions for providing access to their annual reports.

## REFERENCES

[1] M. Sheikh and S. Conlon, "A rule-based system to extract financial information," *J. Comput. Inf. Syst.*, vol. 52, no. 4, pp. 10–19, 2012.

[2] A. S. Corráa and P.-O. Zander, "Unleashing tabular content to open data: A survey on PDF table extraction methods and tools," in *Proc. 18th Annu. Int. Conf. Digit. Government Res.*, Jun. 2017, pp. 54–63.

[3] R. Rastan, "Automatic tabular data extraction and understanding," Ph.D. dissertation, Univ. New South Wales, Sydney, NSW, Australia, 2017.

[4] J. Yan, C. Wang, W. Cheng, M. Gao, and A. Zhou, "A retrospective of knowledge graphs," *Frontiers Comput. Sci.*, vol. 12, no. 1, pp. 55–74, Feb. 2018.

[5] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, Dec. 2017, doi: 10.1109/TKDE.2017.2754499.

[6] L. Ehrlinger and W. Wi, "Towards a definition of knowledge graphs," in *Proc. SEMANTiCS*, 2016, pp. 1–5.

[7] S. Choudhury, K. Agarwal, S. Purohit, B. Zhang, M. Pirrung, W. Smith, and M. Thomas, "NOUS: Construction and querying of dynamic knowledge graphs," in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, Apr. 2017, pp. 1563–1565.

[8] K. Veel, "Make data sing: The automation of storytelling," *Big Data Soc.*, vol. 5, no. 1, 2018, Art. no. 2053951718756686.

[9] R. Angles, "A comparison of current graph database models," in *Proc. IEEE 28th Int. Conf. Data Eng. Workshops*, Apr. 2012, pp. 171–177.

[10] U. Bhatt, "Addressing key challenges in financial services with Neo4j," White Paper, Accessed: Apr. 16, 2019. [Online]. Available: https://neo4j.com/resources-old/neo4j-financial-services-white-paper/

[11] J. Barrasa. (2017). *RDF Triple Stores vs. Labeled Property Graphs: What's the Difference*. [Online]. Available: https://neo4j.com/blog/rdf-triple-store-vs-labeled-property-graph-difference/.

[12] M. P. Muíoz, A. Llaves, and T. Kume, "Populating the FLE financial knowledge graph," in *Proc. Int. Semantic Web Conf.*, 2018, pp. 1–5.

[13] C. Leinemann, F. Schlottmann, D. Seese, and T. Stuempert, "Automatic extraction and analysis of financial data from the EDGAR database," *South Afr. J. Inf. Manage.*, vol. 3, no. 2, 2001, doi: 10.4102/sajim.v3i2.127.

[14] T. Stümpert, "Extracting financial data from SEC filings for US GAAP accountants," in *Handbook on Information Technology in Finance* (International Handbooks Information System), D. Seese, C. Weinhardt, and F. Schlottmann, Eds. Berlin, Germany: Springer, 2008, pp. 357–375, doi: 10.1007/978-3-540-49487-4_16.

[15] M. Atzmueller, P. Kluegl, and F. Puppe, "Rule-based information extraction for structured data acquisition using TextMarker," in *Proc. LWA*, vol. 8, 2008, pp. 1–7.

[16] M. Göbel, T. Hassan, E. Oro, and G. Orsi, "A methodology for evaluating algorithms for table understanding in PDF documents," in *Proc. ACM Symp. Document Eng.*, 2012, pp. 1–8.

[17] T. Loughran and B. Mcdonald, "Textual analysis in accounting and finance: A survey," *J. Accounting Res.*, vol. 54, no. 4, pp. 1187–1230, Sep. 2016.

[18] A. C. E. Silva, A. Jorge, and L. Torgo, "Automatic selection of table areas in documents for information extraction," in *Proc. Portuguese Conf. Artif. Intell.*, 2003, pp. 460–465.

[19] D. C. Wimalasuriya and D. Dou, *Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches*. London, U.K.: Sage, 2018.
[20] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," Stanford Univ., Stanford, CA, USA, 2001.
[21] E. Arendarenko and T. Kakkonen, "Ontology-based information and event extraction for business intelligence," in *Proc. Int. Conf. Artif. Intell., Methodol., Syst., Appl.*, 2012, pp. 89–107.
[22] J. Pujara, "Extracting knowledge graphs from financial filings," in *Proc. Int. Workshop Data Sci. Macro–Modeling Financial Econ. Datasets*, vol. 2, 2017, pp. 1–2.
[23] W. Shen, J. Wang, P. Luo, and M. Wang, "A graph-based approach for ontology population with named entities," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2012, pp. 345–354.
[24] J. Pujara, H. Miao, L. Getoor, and W. W. Cohen, "Ontology-aware partitioning for knowledge graph identification," in *Proc. Workshop Automated Knowl. Base Construct.*, 2013, pp. 19–24.
[25] J. Pujara, H. Miao, L. Getoor, and W. W. Cohen, "Large-scale knowledge graph identification using PSL," in *Proc. AAAI Fall Symp. Series*, 2013, p. 50.
[26] V. Lopez, P. Tommasi, S. Kotoulas, and J. Wu, "QuerioDALI: Question answering over dynamic and linked knowledge graphs," in *Proc. Int. Semantic Web Conf.*, 2016, pp. 362–382.
[27] P. Nelson. *Natural Language Processing (NLP) Techniques for Extracting Information*. Accessed: May 7, 2021. [Online]. Available: https://www.accenture.com/us-en/blogs/search-and-content-analytics-blog/natural-language-processing-techniques
[28] Z. Zhao, S. K. Han, and I. M. So, "Architecture of knowledge graph construction techniques," *Int. J. Pure Appl. Math.*, vol. 118, no. 19, pp. 1869–1883, 2018.
[29] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic web*, vol. 8, no. 3, pp. 489–508, 2017.
[30] J. R. Trigueros, "Extracting earnings information from financial statements via genetic algorithms," in *Proc. IEEE/IAFE Conf. Comput. Intell. for Financial Eng. (CIFEr)*, Apr. 1999, pp. 281–296.
[31] J. Stichbury. (2017). *WTF is a Knowledge Graph*. [Online]. Available: https://hackernoon.com/wtf-is-a-knowledge-graph-a16603a1a25f
[32] K. Veel. *Make Data Sing: The Automation of Storytelling*. Accessed: May 7, 2021, doi: 10.1177/2053951718756686.
[33] B. Yildiz, K. Kaiser, and S. Miksch, "Pdf2table: A method to extract table information from pdf files," in *Proc. IICAI*, 2005, pp. 1773–1785.
[34] A. Shigarov, A. Mikhailov, and A. Altaev, "Configurable table structure recognition in untagged PDF documents," in *Proc. ACM Symp. Document Eng.*, New York, NY, USA, Sep. 2016, pp. 119–122.
[35] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger, and T. Bogaard, "Building event-centric knowledge graphs from news," *J. Web Semantics*, vols. 37–38, pp. 132–151, Mar. 2016.
[36] R. Rastan, H.-Y. Paik, and J. Shepherd, "TEXUS: A task-based approach for table extraction and understanding," in *Proc. ACM Symp. Document Eng.*, Sep. 2015, pp. 25–34.
[37] J. Pujara, H. Miao, L. Getoor, and W. Cohen, "Knowledge graph identification," in *Proc. Int. Semantic Web Conf.*, 2013, pp. 542–557.
[38] W. Liu, J. Liu, M. Wu, S. Abbas, W. Hu, B. Wei, and Q. Zheng, "Representation learning over multiple knowledge graphs for knowledge graphs alignment," *Neurocomputing*, vol. 320, pp. 12–24, Dec. 2018.
[39] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embedding via dynamic mapping matrix," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics*, 2015, pp. 687–696.
[40] S. Zwicklbauer, C. Seifert, and M. Granitzer, "Doser-a knowledge-base-agnostic framework for entity disambiguation using semantic embeddings," in *Proc. Eur. Semantic Web Conf.*, 2016, pp. 182–198.
[41] "The power of graph-based search," Neo4j, White Paper, 2015. Accessed: May 15, 2019. [Online]. Available: https://go.neo4j.com/rs/710-RRC-335/images/Power-of-Graph-Based-Search.pdf
[42] Y. Jia, Y. Qi, H. Shang, R. Jiang, and A. Li, "A practical approach to constructing a knowledge graph for cybersecurity," *Engineering*, vol. 4, no. 1, pp. 53–60, Feb. 2018.
[43] Z. Wang, M. Guo, Z. Li, M. Tang, and J. Yu, "Knowledge graph construction for payment data risk control," in *Innovation Computing* (Lecture Notes in Electrical Engineering), vol. 675, C. Yang, J. Pei, and Y. Chang, Eds. Singapore: Springer, 2020, pp. 1901–1907.
[44] I. Astrova, "How the anti-trustRank algorithm can help to protect the reputation of financial institutions," in *Research Challenges in Information Science* (Lecture Notes in Business Information Processing), vol. 385, F. Dalpiaz, J. Zdravkovic, and P. Loucopoulos, Eds. Cham, Switzerland: Springer, 2020, pp. 503–508.
[45] J. Ren, J. Long, and Z. Xu, "Financial news recommendation based on graph embeddings," *Decis. Support Syst.*, vol. 125, Oct. 2019, Art. no. 113115.
[46] H. Yun, Y. He, L. Lin, Z. Pan, and X. Zhang, "Construction research and application of poverty alleviation knowledge graph," *Web Information Systems and Applications* (Lecture Notes in Computer Science), vol. 11817, W. Ni, X. Wang, W. Song, and Y. Li, Eds. Cham, Switzerland: Springer, 2019, pp. 430–442.
[47] N. Yerashenia and A. Bolotov, "Computational modelling for bankruptcy prediction: Semantic data analysis integrating graph database and financial ontology," in *Proc. IEEE 21st Conf. Bus. Informat. (CBI)*, Moscow, Russia, Apr. 2019, pp. 84–93, doi: 10.1109/CBI.2019.00017.
[48] Z. Chen, S. Yin, and X. Zhu, "Research and implementation of QA system based on the knowledge graph of chinese classic poetry," in *Proc. IEEE 5th Int. Conf. Cloud Comput. Big Data Analytics (ICCBDA)*, Chengdu, China, Oct. 2020, pp. 495–499, doi: 10.1109/ICCBDA49378.2020.9095587.
[49] S. Wang, Z. Wang, and T. Xiaofeng. (2020). *Intelligent Investment Decision Based on Knowledge Graph*. Accessed: Nov. 9, 2020. [Online]. Available: https://ssrn.com/abstract=3449419
[50] S. Hu, L. Zou, J. X. Yu, H. Wang, and D. Zhao, "Answering natural language questions by subgraph matching over knowledge graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 5, pp. 824–837, May 2018, doi: 10.1109/TKDE.2017.2766634.
[51] P. Verhoef, E. Kooge, and N. Walk, *Creating Value with Big Data Analytics*. London, U.K.: Routledge, 2015, doi: 10.4324/9781315734750.
[52] *Lemon: An Ontology-Lexicon Model for the Multilingual Semantic Web*. Accessed: Jan. 19, 2021. [Online]. Available: https://www.w3.org/International/multilingualweb/madrid/slides/declerck.pdf
[53] MCB Bank Limited. *Financial Statements For the year ended December 31, 2017*. Accessed: Jan. 19, 2021. [Online]. Available: https://www.mcb.com.pk/assets/Accounts%20Dec%202017.pdf
[54] (2017). *United Bank Limited*, Accessed: Jan. 19, 2021. [Online]. Available: https://www.ubldirect.com/corporate/resources/ubl/aboutus/financial_report/report_2017/UBLAnnualReport2017.pdf

**SAMREEN ZEHRA** received the B.S. degree in computer science from the University of Karachi, Karachi, Pakistan, in 2005, and the M.S. degree in computer science from Mohammad Ali Jinnah University, Karachi, in 2019.

Since 2008, she has been a Senior Software Engineer with the National Bank of Pakistan, Pakistan. She has a vast experience in application development using the best technical practices. Her former conference paper was about ontology-based sentiment analysis. Her research interest includes efficient use of knowledge management and retrieval techniques to facilitate decision making for the stakeholders in the financial industry.

Ms. Zehra has been also a member of the Australian Computer Society, since July 2018.

**SYED FARHAN MOHSIN** was born in Karachi, Pakistan, in 1976. He received the M.C.S. degree from Al-Khair University (AJK), and the M.S. degree in computer science from the PAF KIET, Karachi, in 2013. He is currently pursuing the Ph.D. degree in computer science with Muhammad Ali Jinnah University (MAJU), Karachi.

Since 2011, he has been working as a Web Developer with the Dow University of Health Sciences. He has published his recent research article "A Survey on Distributed Information Systems using Semantic Web Techniques" in *NUST Journal of Engineering Sciences*, in 2019. His first publication "Web based multimedia recommendation system for e-learning website" was published, in 2010.

**SHAUKAT WASI** received the intermediate certificate from the Cadet College Petaro, in 1998, the degree in computer science from the University of Karachi, and the master's and Ph.D. degrees in computer science from the FAST-National University of Computer and Emerging Sciences (NUCES).

He started his professional career at FAST. He was one of the founding faculty members of the Computer Science Department, DHA Suffa University, Karachi. He is currently an Associate professor and the Associate Dean with the Faculty of Computing (FOC), Mohammad Ali Jinnah University (MAJU), Karachi. He is heading the Interactive and Intelligent Natural Language Processing (IINLP) Research Group, FOC, MAJU. He has published 19 publications in local and international conferences and journals. He has his expertise in text classification and mining, information retrieval and extraction, and human computer interaction.

Dr. Wasi is honored to be a Program Evaluator for the National Computing Education Accreditation Council (NCEAC), Pakistan. He has planned an international computing conference in collaboration with the IEEE Karachi Section, in 2021.

**SYED IMRAN JAMI** received the B.S. degree in computer science from the University of Karachi, in 2000, the M.S. degree in computer science from the Lahore University of Management Sciences, in 2004, and the Ph.D. degree in computer science from the National University of Computer and Emerging Sciences, in 2011.

Since 2006, he has been the founding members of the Research Centre for ubiquitous Computing. He also worked with the Haptics Research Laboratory and the Pervasive and Networked Systems Research Group, Deakin University, Australia. He has authored 16 journal articles and ten conference papers. He worked on several funded research projects and supervised 12 graduate and doctoral students.

**MUHAMMAD SHOAIB SIDDIQUI** (Member, IEEE) received the B.S. degree from the Department of Computer Sciences, University of Karachi, in 2004, and the M.S. and Ph.D. degrees in computer engineering from Kyung Hee University, South Korea, in 2008 and 2012, respectively.

He is currently an Associate Professor with the Islamic University of Madinah, Kingdom of Saudi Arabia. His research interests include routing, security, and management in wireless networks, sensor networks, IP traceback, secure provenance, blockchain technologies, and remote monitoring using the IoT.

Dr. Siddiqui is a member of ACM.

**MUHAMMAD KHALIQ-UR-RAHMAN RAAZI SYED** (Member, IEEE) received the B.E. degree in computer software from the National University of Sciences and Technology (NUST), in 2004, the M.S. degree in computer science from the Lahore University of Management Sciences (LUMS), in 2006, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2010.

From 2011 to 2016, he worked as an Assistant Professor with the Karachi Institute of Economics and Technology. He has been an Associate Professor and the Head of the Department with the Department of Computer Science, Mohammad Ali Jinnah University, since 2016. His research interests include information security, the Internet of Things, content based communications, ad hoc networks, and architectures for engineering applications.