A Knowledge Graph Based Framework in Relationship Modelling and Real-time Monitoring of Market Participants

Wenliang Cheng*, 1, Kaihao Guo*, 2, Tianpei Jiang2, Haipeng Zhang2
1CFETS Information Technology (Shanghai) Co., Ltd, Shanghai, China
chengwenliang_zh@chinamoney.com.cn
2School of Information Science and Technology, ShanghaiTech University, Shanghai, China
{guokh, jiangtp, zhanghp}@shanghaitech.edu.cn

Abstract. Various types of relationships among participants exist in the foreign exchange market, such as transaction, credit relations inside the market and subcompany, investing and shareholding relations outside the market. In the market with complex relationships, an incident happened in one institution may have "butterfly effect". To monitor systemic financial risks, we integrate various types of data and establish a Knowledge Graph describing the relationships between market participants. Based on the Knowledge Graph, we design a framework to measure the impact of various events on the entire market by connecting the internal market relationships to external market related incidents. By using the BiLSTM-CRF model, we extract the relevant entities in the market news with high accuracy and measure the impact on the market with graph related indicators such as closeness and betweenness centralities. In addition, to use external unstructured data, we use a document-level event extraction method to analyse these unstructured datasets and extract new events and relationships from them to dynamically update the Knowledge Graph.

1. Introduction
There are complex relationships between participants in the foreign exchange market. These relationships are partly formed by the trading relationships within the market, such as debtor, creditor, trading. The other comes from relationships outside the market, such as subcompany, credit and investment relationship. These relationships make the market tightly connected. When some market participants are exposed to risks, the impact of the risks may spread by the relationship chain between participants and confuse the market.

In recent years, foreign exchange market participants in China experienced risk events. Their impact on other participants in the market is significant. For example, the risk incident at Baoshang Bank triggered market concerns about the credit of small and medium size banks[1].

In fact, there are certain reasons behind the occurrence of these risk events. For example, the risk of Jinzhou Bank caused by the default of many of its customers, and several of them were listed companies [2]. Simultaneously, before the risk incident of Jinzhou Bank, some of its shareholders frequently
pledged 164 times[1]. Through these incidents, if we establish a monitoring system which can discover the potential problems in time of the market and estimate impact, we can immediately carry out regulatory intervention on relevant institutions to prevent such incidents in advance. In such a system, we would need to perform high dimensional information query and calculation in a complex network structure. This amount of calculation is difficult for traditional database and data processing methods. Knowledge Graph and corresponding knowledge extraction techniques can better complete key contents such as knowledge extraction, knowledge fusion and knowledge reasoning etc. The main contribution in this paper is that we establish a framework for event discovery, analysis and influence evaluation based on the Knowledge Graph of market participants, and achieve good results of the analysis of the influence measurement of external financial information.

2. Related Work

2.1. Knowledge Graph

Knowledge Graph is a semantic network represented by a directed graph structure, where the vertex represents an "entity" in reality, and the edge describes the relationship between entities. The Knowledge Graph is one of the most effective representations of relationships. The earliest Knowledge Graph was proposed by Google and applied to its search engine, through which it understood semantically the information the user wanted to search. Today, most of search engines are supported by other Knowledge Graph such as Freebase, DBpedia [3], Yago [4] and WikiData [5].

The generation method of Knowledge Graph is mainly divided into two types: automatic machine extraction and crowdsourcing. WikiData uses the crowdsourcing method to store the structured data on Wikipedia uniformly to generate a Knowledge Graph. NELL [6] uses machine learning algorithms to continuously generate new extraction templates and extract structured knowledge from unstructured web page text to form Knowledge Graph.

2.2. Application of Knowledge Graph in Financial Area

Financial risk events often involve complex relationships among institutions, accounts, and personnel. Therefore, researchers introduce Knowledge Graph to detect these events. Castelltort et al. [7] proposed a method based on graph database, using a large number of rules to mine historical data and retrieve connections between fraudsters to discover fraudulent gangs. Khrestina et al. [8] used machine learning algorithms to find suspicious transactions in the context of digital payments based on graph theory and related theories. Bak et al. [9] designed a framework for entities and relationships related to payment in economic crimes. Jedrzejek et al. [10] used the company-centric Knowledge Graph to trace back and analysed the association of company accounts in bank transfers to find 90% of suspicious invoices in a money laundering case. Those studies indicate that building a knowledge graph to improve the level of supervision in a financial market is necessary.

![Fig.1 System Framework](image-url)
3. Information Analysis System Framework

In this chapter, we introduce the specific framework of risk discovery and financial information monitor system. The framework of this system is divided into two parts. One is the knowledge base and the other is the financial information analysis module. The knowledge base is expressed in the form of Knowledge Graph. Knowledge Graph provides the system with the priori information such as entity attributes and relationships between entities. The financial information analysis module analysis the external news and announcements and extracts knowledge from it to dynamically update the basic Knowledge Graph.

3.1. System Workflow

The system workflow includes text classification, sentiment analysis, related news matching, named entity recognition and relation extraction. Fig.1 shows the operation process of the system. For example, on March 20, 2019, after receiving the news titled ‘Two employees of Baoshang Bank took more than 500,000 yuan bribes, illegally issued 200 million yuan loans, and failed to recover for many years’, the system first classifies it as risks, negative News. In the second step, the system matches the related news of the past 14 days, such as ‘Baoshang Bank’s 85’s generation manager found another guilty conviction during the sentence, and illegal loans of 200 million were not recovered’ (March 13, 2019). In the third step, the system extracts the lawsuit type event from the text. The composition of the event is the defendant and the court. The two defendants in the event are mapped into the Knowledge Graph, and a lawsuit event is added to the graph. Finally, the two defendants are founded that had an employment relationship with the entity Baoshang Bank. In the last step, the system calculates the graph theory indicators to measure the entity’s influence in the Knowledge Graph, and analyses the related nodes and companies, and characterizes the impact of the risk event on the overall market.

3.2. Knowledge Graph Construction

The preliminary construction of the Knowledge Graph is divided into two steps, knowledge extraction and knowledge integration.

Knowledge extraction is the process of extracting the key information in the data. In this topic, we clean the semi-structured data in Tianyancha and Sina Finance and extract valid information to complete this step.

Knowledge integration is the process of aligning heterogeneous data. The main problem solved is entity disambiguation and attribute normalization. The same entity often has different names on data from different sources. For example, ICBC and Industrial and Commercial Bank of China Co., Ltd. both describe the same entity. The difficulty of knowledge integration lies in the alignment of different descriptions of the same entities, attributes, and relationships in different knowledge bases.

To describe the importance of each entity node in the graph, we apply graph theory values on the nodes and edges of the graph. We use the closeness centrality, betweenness centrality and PageRank [17] to measure the importance of entity nodes in the graph in different dimensions.

3.3. Information Analysis

3.3.1. Text Classification

For the obtained news, we divided the text into six categories, stocks, bonds, macro, risk, financing and others. We choose the Bag of Word (BOW) and the support vector machine (SVM) algorithm. The model of BOW+SVM is widely used in text classification scenarios [18].

3.3.2. Sentiment Analysis

We use text sentiment analysis based on LSTM (Long Short-Term Memory) to divide the input news into three categories: positive, neutral, and negative. The advantage of LSTM is that it considers the order between words in a sentence, and overcomes the problem of ignoring the order and syntax of words in the traditional BOW model. At the same time, LSTM’s memory unit and control gate can improve the ability to process long sequence data.
3.3.3. Related news
In the extraction of relevant news, we adopt the method of matching news of 14 days before the target news to construct the event chain. Vaswani et al. [19] proposed that in the construction of event chains through news, using 14 days as the limit can get the best performance. Compare to the Vaswani’s algorithm that only used article titles as features, we use TF-IDF algorithm to construct the vector of news for both of the title and the content. The vector \( v_a = (\theta_1^a \ldots \theta_n^a) \) is defined as
\[
\theta_w^a = tf_w^a \cdot \log \frac{|C|}{|\{c \in C | w \in c\}|
\]
where \( w \) is the word appearing in the news, \( C \) is the corpus of all news, and \( tf_w^a \) is the frequency of the word \( w \) appearing in news \( a \). We use cosine similarity to describe the similarity between \( news_a, news_b \). The similarity \( sim(a, b) \) is defined as
\[
sim(a, b) = \frac{\sum_{i=1}^{N} \theta_i^a \theta_i^b}{\sqrt{\sum_{i=1}^{N} \theta_i^a \theta_i^a} \sqrt{\sum_{i=1}^{N} \theta_i^b \theta_i^b}}
\]
We select the news with the highest similarity within a given time frame to construct the event chain.

3.3.4. Name Entity Recognition
We use a BiLSTM-CRF model to recognize the entities, namely, the words representing people, organizations, and places in the text. BiLSTM-CRF model is the state-of-the-art method in recent years in NER. It outperforms other methods especially in various NER competitions. The model learns the features of the text context through BiLSTM model and uses CRF to learn the output of the preorder to complete named entity recognition.

LSTM is a kind of recurrent neural network (RNN), but LSTM uses storage unit to replace the hidden layer in RNN. BiLSTM uses a bidirectional LSTM structure, which solves the problem that LSTM cannot encode back-to-front information and can better capture bidirectional semantic dependent information. A single unit of LSTM can be expressed by the following formula:
\[
\begin{align*}
i_t &= \sigma(W_{xt}x_t + W_{ht}h_{t-1} + W_{ct}c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
c_t &= f_tc_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
h_t &= o_t \tanh(c_t)
\end{align*}
\]
Where \( \sigma \) is the sigmoid function, \( i, f, o \) and \( c \) represent the input gate, forget gate, output gate, and cell vector respectively, and their dimensions are the same as the hidden layer vector \( h \). The subscript \( t \) indicates the current input position, \( W \) indicates the parameters of each gate, and \( b \) indicates the offset within each gate. We feed one-hot encoding of each Chinese character into the LSTM model to learn the contextual features of the text.

Conditional random field (CRF) is a discriminative probability model, which is a type of random field and is often used to label or analyze sequence data. The advantage of the CRF model is that it labels the words and infers the tags of the words from the tags of each word. By considering the tagging information of adjacent words, it has a good effect on ambiguous words and words that do not appear in training. Compared with HMM, CRF’s undirected graph mode solves the problem of output independence hypothesis. The global distribution of data is considered during normalization and prevents the situation from falling into a local optimum. The modeling formula of CRF is as follows:
\[
P(I|O) = \frac{1}{Z(O)} \exp \sum_i \sum_k \lambda_k f_k(i, j, l_1, l_2)
\]
where \( i \) is the position of the node, \( k \) means the k-th feature function, \( Z(O) \) is used to normalize the model, \( O = (o_1, \ldots, o_i) \) is the output of the LSTM model, and \( I = (i_1, \ldots, i_j) \) represents the output of the model, that is, the probability of the certain class of input entities, and \( P(I|O) \) represents the probability of hidden state sequence under the condition of the observation sequence \( O \).
3.3.5. DEE Model

In the model building task of announcement extraction, we mainly adopt the Doc2EDAG model, which is an end-to-end DEE model.

The basis of Doc2EDAG is NER. In NER part, BiLSTM-CRF algorithm is used to extract and combine the short name matching with the name of the member unit to map the extracted name to the entity. After the entity is extracted, the entity is serialized into events through a neural network. We decompose the task into two parts. One is chapter-level entity coding, and the other is to construct the extracted entity as a directed acyclic graph. In the entity coding task, the input article is divided into multiple sentences and word vectors are embedded.

In Doc2EDAG, we use directed acyclic graph instead of table filling to extract events. The advantage of generating a directed acyclic graph is that the process of event extraction can be split into multiple subtasks, and each subtask corresponds to a node. After embedding the entity and sentence, we input the entity and sentence vector into a linear classifier to classify the event category, and use the result as the initial node of the directed acyclic graph expansion. For each entity, it is determined whether a new path needs to be generated to describe the event label, which is a binary classification problem. By generating a path for each element in the event, we can get the complete event extraction results.

4. Experiments

4.1. Information Analysis

The information analysis experiment is divided into two parts which is sentiment analysis and text classification. In the experiment of sentiment analysis, we use 7,535 news in various fields to train the LSTM neural network, and test it on a test set composed of 921 related news of member entities, with an accuracy rate of 75.4%. In the text classification model, we used 1,143 member-related news datasets to train the SVM model, and tested on 243 pieces of data, obtaining an accuracy rate of 67.2%.

4.2. NER Experiment

On the NER model, we use MSRA corpus NER dataset for a total of 50,685 sentences to train the BiLSTM-CRF neural network and the process is assisted by the Abbreviation matching technology. We test the model on a dataset of 250 news items related to Inter-bank market participants in China. By extracting 445 related entities, we obtain an accuracy rate 93.33% and a recall rate 97.68%.

4.3. Dataset

We use the previous labelling data and the original public data of the Doc2EDAG model for a total size of 34,568 to train the event extraction model. Table 2 describe the distribution of our dataset.

| Event Type          | Train | Test | Dev  | Total |
|---------------------|-------|------|------|-------|
| Equity Freeze       | 787   | 185  | 203  | 1,175 |
| Equity Repurchase   | 1,862 | 677  | 1,137| 3,676 |
| Equity Overweight   | 5,268 | 566  | 339  | 6,173 |
| Equity Underweight  | 5,101 | 285  | 271  | 5,657 |
| Equity Pledge       | 12,857| 1,491| 1,254| 15,602|
| Total               | 27,857| 3,450| 3,481|34,788|

Table 2. Precision, Recall and F1 score of the test set on the Experiment

| Model    | GreedyDec | Doc2EDAG |
|----------|-----------|----------|
|          | P. | R.  | F1 | P. | R.  | F1 |
| EF.      | 80.0 | 45.2 | 57.8 | 79.6 | 78.1 |78.8 |
| ER.      | 86.4 | 77.2 | 81.5 | 84.2 | 87.5 |85.7 |
| EO.      | 70.1 | 41.9 | 52.5 | 74.1 | 73.9 |74.0 |
### 4.4. Results from Model Experiment

Our baseline model is GreedyDe, which is a simplified version of Doc2EDAG. It uses a greedy algorithm in the part of table filling. There are five types of target events in this experiment, which are Equity Freeze (EF), Equity Repurchase (ER), Equity Underweight (EU), Equity Overweight (EO) and Equity Pledge (EP).

As Table 2 shows, we can see that in the Equity Freeze (21.0%), Equity Repurchase (4.2%), Equity Overweight (21.4%), Equity Underweight (28.1%) and Equity Pledge (23.3%) the Doc2EDAG model obtained a higher F1 value, and the overall performance is also better than the baseline.

### 5. Conclusion

In this paper, we propose a Knowledge Graph based market monitoring system. The system can deal with unstructured text data for text classification, sentiment analysis and approximate text matching to monitor market sentiment and risk events. The system can also measure the impact of risk events on the overall market by calculating graph theory indicators. At the same time, we apply a DEE method in the system to extract more detailed structured information from external data than before, to target entity mapping for risk events, and dynamically update the basic Knowledge Graph.

### Acknowledgments

This project is sponsored by Shanghai Sail Program (No.19YF1433800) and the Key Projects of Shanghai Soft Science Research Program (No. 20692194300).

### References

[1] https://www.reuters.com/article/us-china-banks-baoshang/chinas-baoshang-bank-to-be-taken-over-by-local-governments-state-firms-idUSKBN1ZK0HK

[2] http://finance.sina.com.cn/stock/relnews/cn/2019-08-01/doc-ihytcitm6233933.shtml

[3] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia-A crystallization point for the Web of Data[J]. Web Semantics: science, services and agents on the world wide web, 2009, 7(3): 154-165.

[4] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 697-706.

[5] Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledge base[J]. 2014.

[6] E. P. Wigner, “Theory of traveling-wave optical laser,” Phys. Rev., vol.

[7] Castellstor A, Laurent A. Fuzzy Historical Graph Pattern Matching A NoSQL Graph Database Approach sfor Fraud Ring Resolution[C]//IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, Cham, 2015: 151-167.

[8] Khrestina M P, Dorofeev D I, Kachurina P A, et al. Development of Algorithms for Searching, Analyzing and Detecting Fraudulent Activities in the Financial Sphere[J]. European Research Studies, 2017, 20(4B): 484-498.

[9] Bak J, Cybulka J, Jedrzejek C. Ontological modeling of a class of linked economic crimes[M]//Transactions on Computational Collective Intelligence IX. Springer, Berlin, Heidelberg, 2013: 98-123.

[10] Jedrzejek C, Bak J, Falkowski M. Graph mining for detection of a large class of financial crimes[C]//17th International Conference on Conceptual Structures, Moscow, Russia. 2009, 46.

[11] Program Chair-Boguraev B , Program Chair-Munoz R , Program Chair-Pustejovsky J . Proceedings of the Workshop on Annotating and Reasoning about Time and Events[C]// Workshop on Annotating & Reasoning About Time & Events. Association for Computational Linguistics, 2006.
[12] Liao S, Grishman R. Using Document Level Cross-Event Inference to Improve Event Extraction[C]// ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden. DBLP, 2010.

[13] Hong Y, Zhang J, Ma B, et al. Using Cross-Entity Inference to Improve Event Extraction[C]// Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011.

[14] Chen Y, Xu L, Liu K, et al. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks[C]. international joint conference on natural language processing, 2015: 167-176.

[15] Yang H, Chen Y, Liu K, et al. DCFEE: A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data[C]. meeting of the association for computational linguistics, 2018: 50-55.

[16] Zheng S, Cao W, Xu W, et al. Doc2EDAG: An end-to-end document-level framework for chinese financial event extraction[J]. arXiv preprint arXiv:1904.07535, 2019.

[17] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web.[C]. the web conference, 1999: 161-172.

[18] Ma, Xuezhe, and Eduard Hovy. "End-to-end sequence labeling via bi-directional lstm-cnns-crf." arXiv preprint arXiv:1603.01354 (2016).

[19] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[20] Gottschalk S, Demidova E. Eventkg: A multilingual event-centric temporal knowledge graph[C]//European Semantic Web Conference. Springer, Cham, 2018: 272-287.