Financial Entity Identification and Information Integration (FEIII) 2018 Challenge: The Report of the Organizing Committee

Louiqa Raschid
University of Maryland
louiqa@umiacs.umd.edu

Douglas Burdick
IBM Research
drburdic@us.ibm.com

John Grant
University of Maryland
grant@cs.umd.edu

Joe Langsam
University of Maryland
joe@langsam.org

Jay Pujara
USC/ISI
jpujara@isi.edu

Elizabeth Roman
Thomson Reuters
Elizabeth.Roman@thomsonreuters.com

Ian Soboroff
NIST
ian.soboroff@nist.gov

Mohammed Zaki
RPI
zaki@cs.rpi.edu

Elena Zotkina
University of Maryland
ezotkina@umiacs.umd.edu

ABSTRACT

This report presents the goals and outcomes of the 2018 Financial Entity Identification and Information Integration (FEIII) Challenge. We describe the challenge task and the training dataset. The report summarizes the process, outcomes and plans for the 2019 Challenge.

KEYWORDS

FEIII, financial entities, relationships, competitor relationships, link prediction.

ACM Reference Format:

Louiqa Raschid, Douglas Burdick, John Grant, Joe Langsam, Jay Pujara, Elizabeth Roman, Ian Soboroff, Mohammed Zaki, and Elena Zotkina. 2018. Financial Entity Identification and Information Integration (FEIII) 2018 Challenge: The Report of the Organizing Committee. In *DSMM'18: Data Science for Macro-Modeling with Financial and Economic Datasets, June 15, 2018, Houston, TX, USA*. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3220547.3225218>

1 MOTIVATION AND INTRODUCTION

This is the third edition of the Financial Entity Identification and Information Integration (FEIII) Challenge. The 2016 Challenge was to develop technologies to automatically align diverse financial entity identification schemes and identifiers [6]. This included data from the Federal Financial Institutions

Examination Council (FFIEC), the Securities and Exchange Commission (SEC), and the Global Legal Entity Identifier Foundation (GLEIF).

In 2017, we had a focus on identifying and understanding relationships among financial entities that reflect activities, e.g., participation within a financial contract, and that have an impact on the behavior and performance outcomes of the financial entities [10]. The FEIII 2017 dataset was created from 10-K and 10-Q filings retrieved from the Securities and Exchange Commission (SEC) EDGAR website and additional resources from the National Information Center (NIC) of the FFIEC. We extracted sentences (context sentences) from the filings that provide evidence for a specific role / relationship between the filing financial entity and another mentioned financial entity. The challenge task was a ranked classification task, to identify relevant and interesting sentences in the filing that provide evidence for a specific relationship, described by a role keyword, between the filing financial entity and a mentioned financial entity. The scored evaluation task was to identify those sentences that (1) validated the relationship and / or (2) provided highly relevant and interesting knowledge that further describes financial activities, behavior and performance outcomes.

The 2018 FEIII challenge task builds upon our previous experiences, datasets and tasks. We provide a complex dataset organized around *ego nodes*, where each ego node is a financial entity. The 2018 seed entities are financial entities in the S&P 500 index and include 95 financial entities in the NAICS Sector 52 (Finance and Insurance) and 35 entities in Sector 51 (Information and Communication).

We provide some properties for each entity as well as information about their relationships to other entities. Relationships include subsidiaries, branches and other ownership or control associations. We also include information about potential partners, suppliers and competitors. The ego network is drawn from multiple resources. Some parts of the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, or to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DSMM'18, June 15, 2018, Houston, TX, USA

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5883-5/18/06...\$15.00

<https://doi.org/10.1145/3220547.3225218>

network and properties are not validated and do not have high confidence.

The long term intent of the 2018 challenge is that participants will use the dataset to identify future challenges for 2019 and beyond. However, we also define a 2018 **SCORED** challenge for the short term: Predict the *IsCompetitorOf* relationship edges obtained from the Thomson Reuters Data Fusion (TRDF) resource [3]. For the 95 seed entities in NAICS Sector 52, there are 826 *IsCompetitorOf* edges that were included as training data. The prediction task was to consider 37 seed entities in Sector 51, and to predict 211 *IsCompetitorOf* edges.

The report is organized as follows: Section 2 presents details of the task and the dataset. Section 2.3 summarizes the participant scores. Section 3 discusses lessons learned and plans for the 2019 challenge.

2 2018 CHALLENGE DESCRIPTION

2.1 Dataset

We start with a set of seed companies (financial entities) that are in the S&P 500 index, from NAICS sectors 51 (Information and Communication) and 52 (Finance and Insurance). We have obtained data about these seed companies from the following sources:

- 10-K reports filed with the Securities and Exchange Commission from 2012 through 2016.
- The Global Legal Entity Identifier Foundation [1].
- The TNIC competitor dataset [7].
- The Open Corporates portal [2].
- Thomson Reuters Data Fusion portal [3].

The properties of the ego dataset can be summarized as follows:

- Properties of financial entities from Open Corporate (OC), TRDF and GLEIF, e.g., name, address, jurisdiction where the entity is incorporated, etc.
- (Subsidiary) Relationships between financial entities from OC. This includes parent, branch and control relationships.
- Level 2 relationship information from GLEIF.
- **Training data for scored task:** A subset of properties and relationships from the TRDF knowledge graph, including the following: *IsParentOf*; *IsUltimateParentOf*; *hasStrategicAlliance*; *hasJointVenture*; *IsJointVentureOf*; *IsCompetitorOf*; *IsSupplierOf*.

We note that the semantics for these relationships may not be strictly defined.

There are approximately 45K edges in this dataset, commencing with one of the seed entities.

There are approximately 19K *IsCompetitorOf* edges; we make special note of these edges since they are directly related to the scored prediction task.

While the *IsCompetitorOf* edges are directed edges, participants may assume that they are undirected edges and represent a symmetric relationship.

We filter out those edges (from the 45K edges) where

Count Entities	Count Edges	Within Sector	Cross Sector
95 sector 52	826	794	32
37 Sector 51	211	172	39

Table 1: Statistics for the *IsCompetitorOf* Edges.

both nodes have an SEC CIK value (approximately 17K edges).

- Relationships between a filing financial entity and a mentioned financial entity, and corresponding text, from SEC 10K filings.
- Predicted pairs of competitor financial entities from the TNIC dataset; the TNIC predictions are made based on the text similarity of the corresponding 10 K annual report of each entity.

A visualization of an exemplar dataset of seed entities:

<https://karsha.umiacs.umd.edu/FinNetwork>.

FEIII 2018 data can be obtained from the NIST website: <https://ir.nist.gov/dsfin>.

2.2 Task

The 2018 FEIII challenge task is to enhance a given ego network dataset. A simple enhancement could be to perform entity resolution and link financial entities across two or more resources. This task could also include entity-pair alignment, where both a pair of parent entities, and a pair of child entities, are linked across resources.

One could also confirm a known relationship or further validate a predicted relationship, i.e., increase confidence in the property or relationship. Confidence is with respect to some ground truth dataset, e.g., from Thoms Reuters [3]. A more difficult challenge, and hence a more valuable enhancement to the ego network, would be to predict a new and unknown relationship.

Of particular interest is identifying competitors for each seed financial entity. We start with 132 seed entities with data in TNIC and the TRDF dataset. There are 6113 (distinct) edges in TNIC between any pair of the 132 seed entities; 7 are recursive edges and may be discounted. 17 of the 132 seed entities have 0 competitor entities in TRDF. The remaining 115 entities have 1038 *IsCompetitorOf* edges in TRDF.

Table 1 reports on the statistics of the ground truth *IsCompetitorOf* edges.

The overlap between the 6113 TNIC edges and the 1038 TRDF edges is 957. There is a gap in coverage of 81 TRDF edges starting from TNIC.

Note: TNIC is a symmetric graph. The current TRDF dataset is a sample that does not include symmetric edges. We are working with TRDF to obtain a complete dataset. For the purposes of the scored task, participants could assume that the TRDF graph was symmetric. In this case, there would be 2077 TRDF edges and the overlap with the 6113

Participant	Precision	Recall	F-Score
P1	41.3	18.01	25.08
P2	19.59	40.28	26.36
P3	1.34	4.74	2.09
P4	14.9	14.69	14.8
P7	18.89	69.67	29.73

Table 2: Precision, Recall and F-Score for Participants.

Participant	True Positive	False Positive	True Negative	False Negative
P1	38	54	4656	173
P2	85	349	4361	126
P3	10	736	3974	201
P4	31	177	4533	180
P7	147	631	4079	64

Table 3: Breakdown of the Predictions for Participants.

TNIC edges would be 1914, with a gap in coverage of 163 TRDF edges starting from TNIC.

We provided the count of TNIC edges, TRDF edges and the overlap and gap for each of the 132 seed entities with the original TRDF graph. The MAP over all seed entities was 0.35, and the average Precision@5 was 0.32, respectively. We also provided statistics where we assumed that the TRDF graph was symmetric. The MAP over all seed entities was then 0.63, with an average Precision@5 of 0.70, respectively.

2.3 Summary of Participant Solutions

A total of five organizations representing academia and industry in the US, Europe and Asia, participated in the challenge [4, 5, 8, 9, 11]. We briefly summarize their performance.

Table 2 reports on the precision, recall and F-score.

Table 3 further reports on the accuracy of the predictions.

3 LESSONS LEARNED AND FUTURE PLANS

The 2018 challenge was an excellent example of community building around a shared task. All authors of this report, as well as some additional individuals, participated in weekly conference calls over a period of several weeks, to identify potential datasets and to define the challenge task. Several individuals also contributed time and effort to create the dataset.

The dataset that was created was complex with many types of relationships. We hope that it will be a useful shared resource, and we hope to re-purpose the dataset in future challenges, in particular, aligning subsidiary graphs.

A key lesson from the scored task was that the task was very difficult. While the graph around the seed entities was rich and dense, it was also a small dataset around 132 ego nodes. This limited the size of the training and test dataset. In addition, the training data did not have sufficient features for learning. This was clearly reflected in the F-scores of the participants. We hope to have a more detailed discussion about this during the 2018 DSMM Workshop.

4 ACKNOWLEDGMENTS

This research was partially supported by NIST award 70NANB15H194 and NSF grants CNS1305368 and DBI1147144 and the Smith School of Business.

REFERENCES

- [1] [n. d.]. The Global Legal Entity Identifier Foundation. <https://www.gleif.org/en/>.
- [2] [n. d.]. The Open Corporates Portal. <https://opencorporates.com>.
- [3] [n. d.]. Thomson Reuters Data Fusion. <https://developers.thomsonreuters.com/data-fusion>.
- [4] Hyunsook Cho, Hyeop Lee, Geonsoo Kim, Hyungjun Kim, and Jongseon Park. 2018. PREFER: PREDiction Model for Financial Entity Relation. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*.
- [5] Cesar de Pablo, Alberto Rubio, and Elena Tomas. 2018. Discovering Company Relationships - BBV Data & Analytics. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*.
- [6] Mark Flood, John Grant, Haiping Luo, Louiqa Raschid, Ian Soboroff, Kyungjin Yoo, and Elena Zotkina. 2016. Financial Entity Identification and Information Integration (FEIII) Challenge: The Report of the Organizing Committee. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*. 1:1–1:4.
- [7] Gerard Hoberg and Gordon Phillips. [n. d.]. The TNIC Competitor Dataset. <http://hobergphillips.usc.edu>.
- [8] Min Li, Lucian Popa, Doug Burdick, and Rajasekar Krishnamurthy. 2018. Defining and Capturing Competitor Relationship using HIL in the Financial Domain. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*.
- [9] Jay Pujara. 2018. Hybrid Link Prediction for Competitor Relationships. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*.
- [10] Louiqa Raschid, Doug Burdick, Mark Flood, John Grant, Joe Langsam, Ian Soboroff, and Elena Zotkina. 2017. Financial Entity Identification and Information Integration (FEIII) Challenge 2017: The Report of the Organizing Committee. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*.
- [11] Vipula Rawte, Mohammed Zaki, and Aparna Gupta. 2018. Using Supervised Learning Techniques for Entity Relationships. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*.