Financial Entity Identification and Information Integration (FEIII) 2019 Challenge: The Report of the Organizing Committee

Louisa Raschid
University of Maryland
louisa@umiacs.umd.edu

Douglas Burdick
IBM Research
drburdic@us.ibm.com

Cesar de Pablo
BBVA Data & Analytics
cesar.depablo@bbvadata.com

Mark Flood
University of Maryland
mflood@rhsmith.umd.edu

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

Joe Langsam
University of Maryland
joe@langsam.org

Jay Pujara
USC/ISI
jpujara@isi.edu

Elena Tomas
BBVA Data & Analytics
elena.tomas@bbvadata.com

Ian Soboroff
NIST
ian.soboroff@nist.gov

ABSTRACT

This report presents the goals and outcomes of the 2019 Financial Entity Identification and Information Integration (FEIII) Challenge. We describe two challenge datasets and tasks. FEIII **SHIP** was a bill of lading dataset for incoming shipments to the United States and the task was to identify the major shippers for some product, from some country. FEIII **CALI** included state and local regulatory data from California, and the task was entity linkage for San Francisco restaurants. The report summarizes plans for the 2020 Challenge and the Business Open Knowledge Network (BOKN).

Keywords

FEIII, financial entities, relationships, entity linkage, tariffs, suppliers, supply chains.

1. MOTIVATION AND INTRODUCTION

This is the fourth 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 [Flood et al. 2016]. 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). 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 [Raschid et al. 2017]. 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 was a link prediction task. We provided a complex dataset organized around *ego nodes*, where each node was an entity in the S&P 500 index. We provided some properties for each entity as well as information about their relationships to other entities. Relationships included subsidiaries, branches and other ownership or control associations. We also included information about potential partners, suppliers and competitors. The 2018 challenge was to predict the *IsCompetitorOf* relationship edges obtained from the Thomson Reuters Data Fusion (TRDF) resource [Reuters 2018].

For FEIII 2019, we reached backwards with one dataset and moved forwards with a second. In the spirit of the 2016 entity identification task, the **CALI** task was entity linkage around a set of San Francisco restaurants. The dataset was a compilation of public data sources from California that included information around restaurant inspections, liquor sales, sales taxes, company registrations, etc. In the spirit of more recent link prediction and competitor identification tasks, the **SHIP** challenge was to identify shippers, for some product (identified using the Harmonized Tariff Code), from some country. We provided a dataset of the bills of lading header information from the U.S. Customs and Border Protection Agency's Automated Manifest System (AMS), for incoming US shipments in 2018.

All data are public data. We thank our collaborators from Enigma who provided access to (cleaned) datasets.

<https://public.enigma.com/>

FEIII 2019 data can be obtained from the NIST website:

<https://ir.nist.gov/feiii>

The report is organized as follows: Section 2 presents details of the **SHIP** task and dataset. Section 3 presents details of the **CALI** task and dataset. Section 4 discusses plans for Business Open Knowledge Network (BOKN).

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, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DSMM'19, June 30, 2019, Amsterdam, Netherlands

© 2019 ACM. ISBN 978-1-4503-6823-0/19/06...\$15.00

DOI: <https://doi.org/10.1145/3336499.3338014>

2. 2019 SHIP CHALLENGE DESCRIPTION

2.1 Dataset

A comprehensive dataset of the bill of lading header information from the U.S. Customs and Border Protection Agency Automated Manifest System (AMS), for incoming US shipments in the first 6 months of 2018. The header identifies the shipper, the consignee (recipient), and shipment details including dates, ports, an unstructured description, and some numeric descriptions including count, weight and value. For example, *ASML Holdings (Netherlands) shipped electronic components used in RF sensors to Sonitrol*. We used (5 digit) Harmonized Tariff Codes to identify product categories.

2.2 Task

Given a tariff code and a country, the task was to identify the most significant suppliers of products in that tariff code, or related products in a neighboring tariff code, from that country. There were a range of data cleaning and prediction challenges in addressing the challenge.

- Not all entries have countries; countries may be inferred from the address.
- A more difficult problem was that not all entries have tariff codes. TARIFF codes may be predicted from the item description.
- A major challenge was that the supplier field may not identify the actual supplier of the product. It may instead identify a freight handling company or an agent. This leads to a secondary task to distinguish the real suppliers from freight handling companies.
- A related and more sophisticated task may include predicting the actual shipper or the actual consignee. This information is sometimes provided in other fields like the item description. Shippers (consignees) may also be inferred using features from the network structure of the dataset.

We expand on the task to distinguish real suppliers from agents.

- The item description may identify that a freight handling company is involved and may provide information about the real supplier.
- Freight handling companies may be associated with many tariff codes across multiple entries, in comparison to suppliers of products, who may be associated with a more limited set of tariff codes.
- The industry sector (SIC, NAICS) codes may also be useful features to identify when a shipper is actually a logistics or transportation agent.

3. 2019 CALI CHALLENGE DESCRIPTION

The **CALI** task was entity linkage around a set of San Francisco restaurants. The dataset was a compilation of public data sources from California that included information around restaurant inspections, liquor sales, sales taxes, company registrations, etc. They are as follows:

- San Francisco restaurant inspections (REST).

- California Active Liquor Licences (CALL).
- California Sales Tax Permits (CSTP).
- California Corporate Registrations (CCR).

The **CALI** task was entity linkage from the set of San Francisco based restaurants (REST) to the other three datasets. The precision across all of the matching tasks, as reported by participants after some manual validation, was very good, and typically above 85%. However, all participants noted that the recall was low. For example, there were approximately 6000+ restaurants and over 3 Million CCR records. Participants reported that approximately 15% of the restaurants appeared to have a match with a CCR record. This was surprisingly low coverage, given that the CCR dataset is comprehensive. One explanation may be that many restaurants are owned by companies that are not registered in California. About 400 restaurants had a match to both CALL and CCR records, while less than 100 restaurants had matches to CALL, CSTP and CCR records.

We note that while improved information integration is the direct motivation for successful entity identification and linkage tasks, identifying gaps or an absence of entity linkage can be a positive outcome when the task encounters failure.

4. LESSONS LEARNED AND FUTURE PLANS

The FEIII organizers were invited to submit a proposal to the NSF C-Accel program to develop a Business Open Knowledge Network (BOKN). BOKN is envisioned as a shared resource of curated knowledge. As with past FEIII Challenges, BOKN will have a focus on entity identification and we plan on a symbolic knowledge graph capturing a range of relationship types between entities. BOKN will also include tools to support large-scale data analysis, and interfaces to allow access to additional repositories.

We use the FEIII SHIP dataset to illustrate a use case for BOKN. The re-negotiation of global trade agreements has brought much attention to tariffs that are levied on products that are shipped from a country of origin to a recipient country. This has increased the pressure on small US manufacturers and importers who may need to urgently find alternative suppliers when tariffs limit access to existing options. The impact of tariffs are also of key interest to economists and analysts. For example, increased tariffs on a broad range of goods from China can have a negative impact on some US small businesses in the manufacturing sector and a positive impact on other businesses through market protection.

A BOKN constructed around the FEIII **SHIP** dataset can be used for the following analytical tasks:

- Identify flows, trends, gaps, etc.; some excellent examples and visualization are provided in [Matheson 2019]. For example, there may be a drop of shipments worldwide around the Chinese New Year holiday.
- Aggregate data at the level of industry sector and product category for additional analysis. This will require obtaining functional relationships such as industry sector information (NAICS, SIC codes) for companies, extracting product categories from the shipment entry, mapping product categories to the Harmonized Tariff Schedule, etc.

- Identify bottlenecks and risks [Bonfiglioli and Gancia 2019]. This will require access to additional financial and socio-economic datasets, e.g., identifying all of the manufacturing or supply sources for some product category, estimating the global trading volume and value for some product category / industry sector combination, etc. A specific research question may be determining if there is a potentially unhealthy concentration of trade partners, for some sector.

5. ACKNOWLEDGMENTS

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

6. ADDITIONAL AUTHORS

7. REFERENCES

- [Bonfiglioli and Gancia 2019] Crino R. Bonfiglioli, A. and G. Gancia. 2019. Concentration in International Markets: Evidence from US Imports. In *Barcelona GSE Technical Report*.
- [Flood et al. 2016] 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.
- [Matheson 2019] B. Matheson. 2019. <https://benmatheson.github.io/container/>.
- [Raschid et al. 2017] 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)*.
- [Reuters 2018] Thomson Reuters. 2018. Thomson Reuters Data Fusion. <https://developers.thomsonreuters.com/data-fusion>.