CI-P: Developing the Next Generation of Community Financial CyberInfrastructure for Monitoring and Modeling of Financial Eco-Systems and for Managing Systemic Risk

Louiqa Raschid **University of Maryland**

Keywords:
Include descriptors of the CISE sub-discipline research that the infrastructure enables. You may also include a basic descriptor of the type of infrastructure. These keywords will help in classifying your proposal during for review.

Overview:

For CI-P proposals

- Does the proposal document the potential community involvement in the planning process?
- Comment on the national need for and validity of the research infrastructure being explored.

The **intellectual merits** of monitoring and modeling financial eco-systems..

- One
- Two
- Three
- What
-

The **broader impact** of the next generation of financial cyberinfrastrucuture….

Community and Steering Committee ….

1. Introduction
1.1 Community Description
1.2 Need for a Community Infrastructure
1.3 Vision and Architecture for Community Infrastructure

- computing research infrastructure concept, noting whether it is new infrastructure to be created or existing infrastructure to be enhanced, and provide an estimate of its cost to deploy and operate,
- compelling new research and education opportunities envisioned as being enabled by the infrastructure, and
- steps you will take to identify the consensus needs of the research and education community to be served by the proposed infrastructure, including the process you plan to follow to identify the major characteristics and features of the infrastructure, its useful lifetime, and its cost to create/enhance and operate.

2. Operational Plan
2.1 Plan of Activities
2.2 Plan to Engage Relevant Communities

2.3 Advisory Committee

- Lewis Alexander, Chief U.S. Economist, Nomura. Formerly Counselor to the Secretary of the Treasury.
- Richard Anderson, Economist, Federal Reserve Bank of St. Louis.
- Mike Atkin, CEO, Enterprise Data Management Council.
- Andrei Kirilenko, Professor of the Practice of Finance at the Sloan School of Management, Massachusetts Institute of Technology. Formerly Chief Economist, CFTC.
- Michael Bennett, Semantic Technologies, Enterprise Data Management Council.
- Albert "Pete" Kyle, Robert H. Smith Professor of Finance at the Smith School of Business, University of Maryland.
- Joe Langsam
- Andrew Lo, Charles E. and Susan T. Harris Professor at the Sloan School of Management, Massachusetts Institute of Technology.
- Chester Spatt, Pamela R. and Kenneth B. Dunn Professor at the Tepper School of Business, Carnegie Mellon University
- Nancy Wallace, Lislie and Roslyn Payne Professor at the Haas School of Business, University of California, Berkeley.

2.4 Steering Committee

Elisa Bertino
Michael Franklin
Johannes Gehrke
Lise Getoor
Eduard Hovy
Vagelis Hristidis
H.V. Jagadish
Ben Shneiderman
Amitabh Varhsney
Michael Wellman

2.5 Partners
IBM
Morgan Stanley
Yahoo!
Enterprise Data Management Council
FINRA? SIFMA?

3. From Individual Resources to Community Infrastructure
3.1 Big Picture

DATASETS

- Ground truth datasets a la TDT4.
- Starter or seed datasets that have been manually curated and enriched.
- Large representative collections, e.g., for sampling, de-identification, etc.

TOOLS

USE CASES / SIMULATION SCENARIOS / OTHER ARTEFACTS

3.2 Knowledge Extraction and Network Creation using MIDAS – Shiv

- Analyzing and integrating public data across multiple sources.
- Extract from research reports and correlate with internal data time-series.

In all of the above the quality of extraction and integration is of the utmost importance.
Research must address a range of challenges from from extraction to linking to correlation to predictive algorithms.

3.3 Language, Intent and Semantics - Analysis and Prediction from SEC Filings - Hoberg
3.4 Social Media Modeling and Prediction – Hristidis

3.5 Assessing Information Quality in the pre-CICI and post-CICI/LEI Eras

The lack of unique and potentially immutable identifiers to represent legal entities (organizations) and financial instruments is a major impediment to information sharing and improving information quality. Addressing this issue correctly can single-handedly resolve many data quality issues around systemic risk. For example, CUSIP [] was developed to identify securities, but it is proprietary, and a fee-per-usage model has been developed around it. The proprietary nature of the CUSIP prevents federal agencies from sharing information that is linked to a CUSIP, leading to a major barrier to quality improvement. Post the passage of Dodd Frank, the CFTC wrote several rules around the adoption of a CICI (CFTC Interim Compliant Identifier). Something about the Legal Entity Identifier (LEI).

Consider the following three scenarios/eras:
Current status: Company X (Morgan Stanley) maintains an internal database of entity identifiers and organizational hierarchies.
Short term future: CICI is widely deployed so that (public) financial contracts can be *marked up* using the CICI. Marked up means that if the same entity is a counterparty on several contracts, these contracts can be easily retrieved in response to a query against this entity.
Some future (ideal) state: LEIs are widely deployed.

Next consider the types of queries of interest to a federal regulator:
(1) A federal regulator asks Company X (Morgan Stanley) to report on its complete exposure to Company Y.
(2) A federal regulator asks Company X (Morgan Stanley) to report on its assessment of risk with respect to some position that X holds that involves an exposure to Company Y.

We must develop tools and datasets to answer the queries above as well as to address some of the following interesting research questions:
What information advantage does Company X (which has full knowledge of its inventory and positions) have over the federal regulator (which has full knowledge of the LEI database as well as confidential information reported

historically by company and other institutions.
Conversely, what information advantage does the federal regulator have over Company X?

3.6 Karsha Annotation Recommendation and Markup Tool Using the Financial Industry Business Ontology (FIBO)

Karsha is a markup and recommendation tool to curate a repository of financial documents. Annotation can be done using the Financial Industry Business Ontology (FIBO) as well as other financial ontologies or thesauri. Raschid and colleagues are developing a sample repository comprising a collection of bond prospectus (corporate and municipal bonds) and their supplemental information. Karsha constructs a (Lucene) index over sections of the document (indexing the keywords within sentences). It uses Okapi cosine keyword based similarity [] to compare the sections (sentences) of the document with definitions for FIBO ontology terms and chooses/recommends the Top K terms. We focus on the FIBO [] since it provides an excellent set of definitions for each FIBO term. Karsha is already producing excellent initial results in providing Top K recommendations of FIBO terms using unsupervised methods, *without the use of training data or semi-supervised methods to tune the recommendation system.*

Potential use cases include the following:

- Rank and retrieve documents using FIBO search terms.
- Cluster documents to better understand the contents of a repository.
- Compare pairs of documents for similarities as well as gaps or dissimilarity.

Karsha can be extended to include sentence understanding so that one can answer more refined questions such as *Which of these instruments in this repository is likely to be impacted by a fluctuation of the price of crude oil futures*?

4. Broader Impact

5. Prior NSF Support

References

Supplementary Documents: In the Supplementary Documents Section, provide a list of PIs, Co-PIs, Senior Personnel, paid Consultants, Collaborators and Postdocs to be involved in the project. This list should be numbered and include (in this order) Full name, Organization(s), and Role in the project, with each item separated by a semi-colon. Each person listed should start a new numbered line. For example: