CEUR-WS.org/Vol-3254/paper397.pdf

Knowledge Graph - Our Secret Recipe for Entity Disambiguation*

Dmytro Dolgopolov[1], Elena Romanova[1], Lev Yarmovsky[1] and Chen Zhang[1]

[1]*FINRA (the Financial Industry Regulatory Authority), Rockville, MD, USA*

Abstract

Being able to correctly disambiguate persons and organizations mentioned in various documents can have a great impact on helping FINRA achieve its mission - protecting investors and ensuring the integrity of U.S. securities markets. Our solution was to disambiguate entities and make connections through FINRA Knowledge Graph [1]. Various text analytics techniques were used to improve data quality. The sparsely populated data led us to a home-grown "seed" algorithm - a bottom-up community detection technique that grows communities around reliable clusters (seeds). We adapted a clustering algorithm used for DNA sequencing to improve the quality of disambiguation. Using these approaches, we were able to find a practical solution for clique detection problem which is foundational for entity disambiguation.

Keywords

Knowledge Graph, Entity Disambiguation, Clustering Algorithms, Graph Clique, Machine Learning, Community Detection

1. Introduction

FINRA receives hundreds of thousands of various documents each year from stockbrokers and investors to be reviewed and analyzed by investigators. They are looking for the information about Who, What, Where, When and How mentioned in these documents. It is a very labor-intensive process to identify persons and organizations in free text, structured and semi-structured data, let alone connect those mentions across all the input data to identify regulatory significant patterns. On top of that, the amount of information about persons and organizations varies across the sources (e.g., some documents may only contain postal address while others only contain the phone number). Traditional disambiguation techniques failed for these scenarios.

Fortunately, knowledge graph shows more potential in our scenario. First of all, the graph technique is an easy way to integrate data with various data formats in various dimensions. Second, knowledge graph allows involving in relationship information as additional data for entity disambiguation process. Lastly, graph clustering algorithms provide more flexible classification experiences than traditional grouping or clustering methods such as K-Means.

ISWC2022: The 21st International Semantic Web Conference, 23–27 October 2022, Hangzhou, China

* FINRA Knowledge Graph Team

dmytro.dolgopolov@finra.org (D. Dolgopolov); elena.romanova@finra.org (E. Romanova); lev.yarmovsky@finra.org (L. Yarmovsky); chen.zhang@finra.org (C. Zhang)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

2. Methods

Some traditional algorithms are used for data preparation, for instance "Data Bucketing" algorithm is used for initial grouping, Machine Learning (ML) solutions are used for name similarity comparison etc.

For entity disambiguation we experimented with graph community detection clustering algorithm through connected components. Based on our experience, when applied to tightly connected graphs, it usually creates connections between loosely associated data, introducing "long chain" relationships. For targeted community detection we developed Seed Algorithm, with which "seeds" are generated by connecting nodes via reliable edges. Other potentially less reliable connections were added later on to grow the seed clusters into bigger communities. As a result, clusters built in this way represent less ambiguous groups.

However some remaining ambiguous connections need to be eliminated. We designed a home grown edge cutting algorithm similar to CLICK algorithm [2], that was first developed in bio engineering for successful gene patterns clustering. All edges assigned weights that represent how strongly the nodes are connected by a specific type of an edge. The algorithm recursively cuts the lowest weighted edges between communities, representing the weakest links, until graph is reaching a desired threshold and the communities are separated.

3. Results and Conclusion

In 2021 alone our solution processed 13 million documents and tens of billions of facts found in structured and unstructured sources. Some 36 million references of organizations and 4.7 million persons mentioned in these sources were identified. Leveraging our approach, we were able to uniquely identify 1 million organizations and 700 thousand persons of interest.

Past experience taught us that common approaches don’t work for sparsely populated data. The high quality of entity extraction results play an important role. More importantly, Knowledge Graph allows us to connect entities with unrelated features through graph connections. As for the two types of entities that we are analyzing, person disambiguation turns out to be much harder than organization disambiguation.

In the future, we plan to use Graph Neural Network (GNN) to analyze the quality of disambiguated communities. In addition, we plan to use GNN to identify patterns within the knowledge graph that represent business relevant scenarios before they become apparent for investigators.

References

[1] D. Dolgopolov, E. Romanova, Using knowledge graph to improve enterprise search experience, ISWC (2018).

[2] R. Sharan, R. Shamir, Click: A clustering algorithm with applications to gene expression analysis, Algorithm with Applications to Gene Expression Analysis, AAAI (2000).