Hybrid Link Prediction for Competitor Relationships

Jay Pujara
Information Sciences Institute
jpujara@isi.edu

1. INTRODUCTION

Competitor relationships are integral to many important financial applications. Example use cases include understanding regulatory impacts, investing in new business areas, and building economic models. Competitor relationships can be defined based on several aspects, including valuations and asset returns, industrial processes, or offerings of products and services. Determining these relationships is often challenging due to the diverse and complex interactions between companies which must be mined from vast datasets with varying degrees of credibility. In this paper, we approach this problem by constructing a *hybrid* knowledge graph capturing financial relationships and applying a link prediction model to identify missing competitor relationships.

Knowledge graphs are a popular knowledge representation choice for capturing entities and the relationships between them. Knowledge graph construction typically uses only a single type of input data, such as relationships mined from text using information extraction techniques or curated relationships from relational databases. In contrast, for the FEIII Challenge[1], we are provided with several sources of relationships from different types of input, including expert judgments, mined relationships, and statistical features. Our approach creates a hybrid knowledge graph that includes relationships derived from three very different types of data in a single knowledge graph.

We construct a hybrid knowledge graph using data provided for the FEIII Challenge and one additional source, the webpages of companies included in the challenge. The first data source we use are expert judgments curated by the Thomson Reuters Data Fusion (TRDF) platform[2]. The second data source we are provided in the challenge are relationships extracted from text found in SEC filings. Finally, we introduce a third set of statistical signals, derived primarily from collecting webpages of the companies in the knowledge graph and applying natural language processing (NLP) tools to the text of the webpages.

After constructing a hybrid knowledge graph, we develop a link prediction model to predict competitor relationships missing from the knowledge graph. Our model captures dependencies between different relationships in the knowledge graph, supports collective inference of all competitor relationships simultaneously, and is capable of combining diverse signals as part of the predictive process. Our model uses probabilistic soft logic (PSL), a modeling framework that supports efficient, collective inference and expressive specification of rules using a first-order logic syntax.

To understand the value of different signals in the FEIII Challenge dataset, we perform an ablation study to determine the value of each type of input. We find that while expert and mined relationships provide high-precision relationships, recall is extremely low. In contrast, statistical signals provide a similar level of precision with substantially higher recall. Overall, our approach yields an F1 score of 0.74 in a cross-validated experiment on the challenge training data and achieves an F1 score 0.15 on the challenge task.

2. RELATED WORK

Competition has been a longstanding area of interest in economic and financial research Hotelling [1929], Chamberlin [1933]. Approaches to predicting competition have used industry classifications Pearce [1957], financial results Fama and French [1997], Bhojraj and Lee [2002], product offerings Rauh and Sufi [2011], and textual analysis of company documents Hoberg and Philips [2016]. Our knowledge fusion approach can benefit from all of these disparate approaches, combining these signals in a single model.

Knowledge graphs and knowledge base construction have long been seen as a critical approach for capturing and organizing useful knowledge Lenat et al. [1990], Dong et al. [2014], Nickel et al. [2015], Carlson et al. [2010], Bollacker et al. [2008]. Prior work Pujara et al. [2013] has demonstrated that collective models, such as PSL Bach et al. [2017], provide superior performance on cleaning and completing knowledge graphs. We adopt this approach and apply it to financial datasets.

3. APPROACH

We model the FEIII Challenge data using a hybrid knowledge graph. We process the raw data provided in the challenge, and identify several types of relationships from each data source. We refer to relationships from the TRDF platform as Expert data, relationships from text extracted from

[1] http://dsmmworkshop.org/

[2] https://developers.thomsonreuters.com/data-fusion

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

ISBN 978-1-4503-5883-5/18/06…$15.00
DOI: 10.1145/3220547.3220559

Source	Relationships
Expert	Same, Competitor, Joint Venture Strategic Alliance, Supplier, Industry
Mined	Competitor, Competitive, Compete, Competing, Trustee, Agent, Control, Affiliate, Parent, Issuer
Statistical	Webpage Similar, TNIC

Table 1: Expert, Mined, and Statistical Relationships used in our model

SEC filings as Mined data, and relationships found using statistical techniques (including TNIC) as Statistical data. Table 1 contains a list of these relationships used in our model.

After identifying relationships in the knowledge graph, we define a PSL model for using these relationships. The PSL model consists of a series of rules that specify interdependencies between relationships in the knowledge graph. In our model, we are interested in predicting the Competitor relationship using information from expert, mined, and statistical features. After assembling these rules, the weight, or importance, of each rule is learned using training data.

Expert Rules.

Rules based on expert data exploit several relationship patterns. One pattern is that if two identifiers refer to the same company, the identifiers do not define a competitive relationships, and all competitors are shared between these two identifiers. Another pattern is that if two companies share a common supplier, they are likely to be competitors. A third type of pattern is that if two companies have a joint venture (or strategic alliance, or supplier relationship) they are unlikely to compete. Finally, if two companies operate in the same industry, they are likely to compete. A sample of these rules are listed below:

$$\begin{aligned}
\text{TRDF}\textsc{Same}(\mathtt{C1},\mathtt{C2}) \quad & \land\ \textsc{Competitor}(\mathtt{C1},\mathtt{T}) \\
& \rightarrow \textsc{Competitor}(\mathtt{C2},\mathtt{T}) \\
\text{TRDF}\textsc{Same}(\mathtt{C1},\mathtt{C2}) \quad & \rightarrow \neg\textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
\text{TRDF}\textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \quad & \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
\text{TRDF}\textsc{Supplier}(\mathtt{S},\mathtt{C1}) \quad & \land\ \text{TRDF}\textsc{Supplier}(\mathtt{S},\mathtt{C2}) \\
& \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
\text{TRDFJV}(\mathtt{C1},\mathtt{C2}) \quad & \rightarrow \neg\textsc{Competitor}(\mathtt{C2},\mathtt{C2}) \\
\text{TRDF}\textsc{Industry}(\mathtt{C1},\mathtt{I}) \quad & \land\ \text{TRDF}\textsc{Industry}(\mathtt{C2},\mathtt{I}) \\
& \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2})
\end{aligned}$$

Mined Rules.

The second type of rule we introduce uses relationships mined from SEC filings on the basis of keywords identified by domain experts. In some cases there are multiple keywords surrounding the same concept (such as “competitor”, “compete”, and “competing”) which may have differing precision and are associated with different rules to capture these potential differences. The first rules capture these different indicators of competition. Other relationships, such as trustee (or agent) may be negatively correlated with competition. When companies share a common parent (or controlling entity), they may also be less likely to be competitors. However, sharing other types of relationships, such as competitors, agents, affiliates, issuers, or trustees, may increase the probability of a competitive relationship. We encapsulate these types of patterns in the sample rules provided below:

$$\begin{aligned}
\text{SEC}\textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \quad & \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
\text{SEC}\textsc{Compete}(\mathtt{C1},\mathtt{C2}) \quad & \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
\text{SEC}\textsc{Trustee}(\mathtt{C1},\mathtt{C2}) \quad & \rightarrow \neg\textsc{Competitor}(\mathtt{C!},\mathtt{C2}) \\
\text{SEC}\textsc{Parent}(\mathtt{C1},\mathtt{P}) \quad & \land\ \text{SEC}\textsc{Parent}(\mathtt{C2},\mathtt{P}) \\
& \rightarrow \neg\textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
\text{SEC}\textsc{Affiliate}(\mathtt{C1},\mathtt{A}) \quad & \land\ \text{SEC}\textsc{Affiliate}(\mathtt{C2},\mathtt{A}) \\
& \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
\text{SEC}\textsc{Competitor}(\mathtt{C1},\mathtt{C}) \quad & \land\ \text{SEC}\textsc{Competitor}(\mathtt{C2},\mathtt{C}) \\
& \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2})
\end{aligned}$$

Statistical Rules.

The last type of rule we introduce uses relationships mined from statistical patterns and NLP tools. One statistical signal is a competition probability derived from the similarity of SEC filings, which is provided by the TNIC resources. Another similarity score was derived by retrieving the webpage of each company, performing a set of standard normalizations to remove stopwords and identify salient terms, and computing a statistical similarity between webpages. We experiment with several normalization techniques for computing these similarities and include all of them as separate rules. Finally, we encoded transitivity and common-competitor relationships, so that competitor relationships are propagated and competitive cliques are completed. We list the sample rules based on these signals below:

$$\begin{aligned}
\text{TNIC}\textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \quad & \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
\textsc{WebpageSimilar}(\mathtt{C1},\mathtt{C2}) \quad & \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
\textsc{Competitor}(\mathtt{C1},\mathtt{C}) \quad & \land\ \textsc{Competitor}(\mathtt{C},\mathtt{C2}) \\
& \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
\textsc{Competitor}(\mathtt{C1},\mathtt{C}) \quad & \land\ \textsc{Competitor}(\mathtt{C2},\mathtt{C}) \\
& \rightarrow \textsc{Competitor}(\mathtt{C1},\mathtt{C2})
\end{aligned}$$

Standard Rules.

Three standard rules are included in the model. The first is a prior rule which states that, in the absence of other evidence, no two companies are competitors. The second rule enforces the constraint that a company does not compete against itself. The last rule forces competitor relationships to be symmetric. are likely to compete. A sample of these rules are listed below:

$$\begin{aligned}
& \rightarrow \neg\textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \\
& \rightarrow \neg\textsc{Competitor}(\mathtt{C},\mathtt{C}) \\
\textsc{Competitor}(\mathtt{C1},\mathtt{C2}) \quad & \rightarrow \textsc{Competitor}(\mathtt{C2},\mathtt{C1})
\end{aligned}$$

In the next section, we show the power of each of these sets of rules individually, as well as investigating how performance changes as different signals are combined in the model.

4. EVALUATION

To evaluate our approach, we used the labeled training data provided as part of the FEIII Challenge. The scripts used for data preparation and testing are publicly available at https://github.com/puuj/pujara-dsmm18. We used the competitor relationships reported by TRDF (in TRDF edges) as the ground truth. To perform a meaningful cross-validation, we represented these ground truth relationships as a graph, and used the METIS package Karypis and Kumar [2009] to perform a minimum-weighted equal vertex edge-cut to produce five connected components. This procedure minimized the overlap of information across different validation folds. We created five separate validation folds using this graph partitioning output. In our cross-validated experiments, we conditioned on three folds of data, used one fold as the training targets and supervision, and held out the last fold of data for evaluation.

We evaluated each of the three models (Expert, Mined, Statistical) separately, considered three variants where two of the three signals were used, and a model combining all of the data sources. In the interest of time, we only learn rule weights from training data when using the combined model containing all of the rules. We compare to a baseline that predicts all competitor edges as true, maximizing recall at the expense of precision. We report precision, recall, F1 score, and area under the precision-recall curve in Table 4.

We observe that the baseline method has low precision, since most company pairs are not competitors. Using the expert knowledge in the TRDF resources results in high precision (0.78) but low recall (0.03), which is expected in a manually-curated resource. The mined rules have much lower precision (0.56) and even lower recall (0.01) highlighting the difficulty of mining high-precision relationships using only keywords and a limited textual corpus. Statistical signals have an advantage with precision approach that of the expert rules (0.77) and much higher recall (0.68). In our experiments, combining different combinations of signals failed to improve the results of either model alone, possibly because the training data and procedure were not sufficient to balance the input features in these cases. When combining all of the rules, we find that the training procedure does help with knowledge fusion by retaining the high precision of the best models, boosting the recall, and improving the F1 and AUC. Surprisingly, applying this trained model on the FEIII challenge test data produced significantly worse results, with an F1 of 0.16 and an AUC of 0.09. This suggests that the training data, patterns, and graph structure between these two datasets may differ substantially.

5. CONCLUSION

In this paper, we develop a system for combining diverse data sources, such as expert knowledge, keyword-based information extraction, and statistical predictions, in a single, hybrid knowledge graph We introduce a model for identifying missing competitor relationships in this hybrid knowledge graph, and show how different data sources contribute to the model, finding that statistical signals offer high precision and coverage, and are an important component for predicting competitor relationships. A surprising result was that applying a model with strong performance on the FEIII challenge training data on the challenge problem resulted in far lower performance. Generalizing the competitor model across datasets and incorporating new rules and types of signals remain areas of open research and future work.

Method	Prec.	Recall	F1	AUC
Baseline	0.06	1.00	0.10	0.06
Expert	0.78	0.03	0.05	0.39
Mined	0.56	0.01	0.02	0.36
Statistical	0.77	0.68	0.72	0.83
E + M	0.63	0.02	0.04	0.38
E + S	0.73	0.60	0.62	0.81
M + S	0.77	0.68	0.72	0.83
All	0.77	0.72	0.74	0.81
Challenge	0.16	0.17	0.16	0.09

Table 2: Results for single-source and two-source models, as well as a combined model. All results except the challenge results are averaged from a 5-fold cross-validated experiment. Expert rules have the best precision, while statistical rules demonstrate high recall without sacrificing precision. Combining all models improves the recall and F1 score.

Acknowledgements

This material is based on research sponsored in part by the National Science Foundation under Grant Nos. 1561057 and 1561068.

References

S. H. Bach, M. Broecheler, B. Huang, and L. Getoor. Hinge-loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research (JMLR)*, 2017.

S. Bhojraj and C. M. C. Lee. Who is my peer? a valuation-based approach to the selection of comparable firms. *Journal of Accounting Research*, 40(2):407–439, 2002.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference On Management Of Data*, pages 1247–1250. ACM, 2008.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell. Toward an Architecture for Never-Ending Language Learning. In *AAAI*, 2010.

E. Chamberlin. *A Theory of Monopolistic Competition.* Harvard University Press, 1933.

X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 601–610. ACM, 2014.

E. F. Fama and K. R. French. Industry costs of equity. *Journal of financial economics*, 43(2):153–193, 1997.

G. Hoberg and G. Philips. Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, 124(5), 2016.

H. Hotelling. Stability in competition. *Economic Journal*, 39(153):41–57, 1929.

G. Karypis and V. Kumar. MeTis: Unstructured Graph Partitioning and Sparse Matrix Ordering System, Version 4.0. http://www.cs.umn.edu/˜metis, 2009.

D. B. Lenat, R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd. Cyc: toward programs with common sense. *Commun. ACM*, 33(8):30–49, Aug. 1990. ISSN 0001-0782. . URL http://doi.acm.org/10.1145/79173.79176.

M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A Review of Relational Machine Learning for Knowledge Graphs: From Multi-Relational Link Prediction to Automated Knowledge Graph Construction. *arXiv preprint arXiv:1503.00759*, 2015.

E. Pearce. *History of the Standard Industrial Classification*. Bureau of the Budget, Office of Statistical Standards, 1957.

J. Pujara, H. Miao, L. Getoor, and W. Cohen. Knowledge Graph Identification. In *ISWC*, 2013.

J. D. Rauh and A. Sufi. Explaining corporate capital structure: Product markets, leases, and asset similarity. *Review of Finance*, 16(1):115–155, 2011.