Journal of Web Systems and Applications (2017) Vol. 1, Numuber 1
Clausius Scientific Press, Canada

Prediction of Datasets sameAs Interlinking on Web of Data

Haichi Liu*, Ting Wang and Jintao Tang

College of Computer, National University of Defense Technology, Changsha, Hunan Province, China

liuhaichi@nudt.edu.cn

Keywords: Linked data, Dataset, sameAs interlinking, Link Prediction.

Abstract. In order to be considered as Linked Data, the datasets on the web must be linked to other datasets. We focus on predicting the possible links between datasets with the most important RDF link type, owl:sameAs using link prediction and classification techniques. Since the goal is to discriminate between linked dataset pairs against not-linked ones, we formulate the link prediction problem as a classification problem. We adopt Random Forest as the basic classifier to incorporate features of the scores output by unsupervised predictors, and apply the bagging technique to combine multiple forests to reduce variance and improve the accuracy. Experiments show we can improve the prediction performance by about 10% in AUROC compared with the best unsupervised predictor.

1. Introduction

The emergence of Linked Data approach has led to the availability of a wide variety of structured datasets on the Web [1]. On Web of Data, the use of owl: sameAs [2] predicate is ubiquitous in interlinking Linked Data datasets to support merging distributed descriptions of equivalent RDF resources from different datasets. The considerable scale makes choosing target datasets that should be interlinked with a given dataset a big challenge [3,4]. In most related works [5,6], the authors didn't distinguish the RDF link types, while in real application scenarios, merely identifying target datasets for interlinking without specifying the link type would be of less help as dataset publishers still don't know what kinds of RDF links can be established and furthermore how to configure the data linking algorithms. Nikolov et al. [3] proposed approach of identifying relevant datasets for establish equivalence relations between individuals of datasets depending on the support of a third-party semantic search engine, and they have only conducted experiments on three datasets as examples.

We investigate the problem of identifying which dataset pairs are most likely to be linked together as link prediction and classification problem. We construct the sameAs graphs from LOD Cloud 2014 [1]. We apply 13 unsupervised predictors implementing link prediction measures to the graphs. Since the goal is to discriminate between examples of the linked dataset pairs against examples of the not-linked dataset pairs, we formulate the link prediction as a supervised classification problem. We adopt Random Forest as the basic classifier to incorporate features of the scores output by unsupervised predictors, and we apply bagging technique to combine multiple forests to reduce

variance and improve the accuracy. We have published all scripts and source code for prediction and extraction at GitHub (https://github.com/HaichiLiu/LP-sameAs-networks) with the sameAs graphs, so that all results presented in the following can be verified.

2. Proposed Approach

2.1 sameAs Graph Construction

We construct the experimental graph from the LOD Cloud 2014 dataset published in [1]. The dataset was a crawl of the Web of Linked Data conducted in April 2014, which contains 8,038,396 resources crawled from 900,129 documents. Altogether, the crawled data belongs to 1014 different datasets. For analyzing the sameAs linkage between datasets, we aggregate all sameAs RDF links by dataset. Specifically, we consider two datasets to be linked if there exists at least one sameAs RDF link between resources belonging to the datasets. If we consider the graph is directed, the direction of links can be defined from the dataset that contains the subject of RDF link to the dataset contains the object. Otherwise, the link is undirected. Taking datasets as vertices and links between datasets as edges, we managed to construct a directed graph with 651 edges and 326 vertices. By common agreement, Linked Data publishers use the predicate owl:sameAs stating that two URI aliases refer to the same resource, so the link can also be treated as undirected. We also constructed an undirected graph with 585 edges and 326 vertices. Both graphs were used in our experiment.

2.2 Unsupervised Predictors

Most existing studies in link prediction consider baseline unsupervised methods to assign scores to potential links. The state-of-the-art in these methods is aggregated and compared in [7]. We apply 13 unsupervised predictors implementing link prediction measures to both graphs, the list can be found in Table 1. In experiments, the predictor name comes after the specification of directionality (I/O) and separated from the predictor name by an underscore. In the undirected graph, it will only perform one set of predictions since there is no concept of in-edges and out-edges. Preferential Attachment is explicitly set to use out-edges of source and in-edges of target in the directed graph.

2.3 Bagging with Random Forests

Since the goal is to discriminate between the linked dataset pairs (positive instances) against the not-linked dataset pairs (negative instances), we can formulate the link prediction problem as a supervised classification problem. Supervised algorithms are able to capture interdependency relationships between topological properties of these measures. Classification algorithms, especially unstable ones like decision trees, can benefit from reduced variance by being placed in an ensemble framework. Supervised classification offers many strong options for reducing variance such as bagging and random forests for decision trees, the latter can also increase classification efficiency. We use WEKA bagging (10 bags, default parameters) with random forests (10 trees, default parameters) as supervised classification framework in experiments. For directed graph, we use all the 13 unsupervised predictors' scores as features set for classification method. For undirected graph, the 9 unsupervised predictors' scores are used as features set.

3. Experiments

3.1 Experimental Setup

For evaluation, in both directed/undirected graphs, we randomly divide known links into 10 equally sized sets we randomly divide the edges in graph into 10 equally sized sets, and mark them from L1 to L10. To evaluate the unsupervised predictors, we use L10 for obtaining ground truth, and use the

graph with links in sets L1 to L9 for computing predictive score. To evaluate the supervised method, we have constructed several graphs from the divided links set. From the first graph, $G_x = (V_x, E_x)$, constructed from L1 to L8, we extract topological measures, and potentially node attributes, that serve as features for each pair of nodes (v_i, v_j). From the second graph, $G_y = (V_y, E_y)$, constructed from L9, we examine (v_i, v_j) to discover whether e_{ij} exists and determine the class label. This yields our training set. We construct testing set in the same way as stated above, but use L1 to L9 to construct the first graph, L10 to construct the second graph.

3.2 Results

In both directed and undirected graphs, the ROC curves of each predictor are drawn together in Fig. 1 for comparison. In the directed graph, only the JPageRank_I, Jac-cardCoefficient_I, SimRank_O predictors' ROC curves are below the diagonal line y = x with one (JPageRank_I, JaccardCoefficient_I) or three (SimRank_O) points when FPR (False Positive Rate) is rather small. In the undirected graph, the JaccardCoefficient, SimRank predictors' ROC curves are below the line y = x with one points when FPR is rather small. The ROC curves of all the other predictors with the remaining points are always higher than the line y = x, which indicates that these predictors are effec-tive in predicting the existence of links in dataset sameAs interlinking graphs. The AUROC values of unsupervised predictors are summarized in Table 1. In the directed graph, the AUROC values are between 0.67 and 0.86, and JPageRank_O achieves the highest value of about 0.8506. In the undirected graph, the AUROC values are between 0.70 and 0.85, and PreferentialAttachment achieves the highest value of about 0.8495. These predictors outperform the random predictor by at least 34% AUROC for directed graph, 40% for undirected graph.

We compare the best performed unsupervised method and supervised method with optimized parameter ratio_positive in ROC curves and AUROC, and the results are showed in Fig. 2 and Table 2 respectively. We can see from Fig. 2 that super-vised method outperforms unsupervised method for both directed and undirected graphs. As summarized in Table 2, for the directed graph, supervised method outperforms unsupervised method by 9% AUROC, 13% for the undirected graph.

4. Conclusion

Identifying which two Linked Data datasets can be linked by owl:sameAs links was studied as a link prediction problem in this paper. We constructed sameAs interlinking graphs from newly updated LOD Cloud 2014 dump with directed and undirected edges respectively. We exploited several unsupervised link prediction measures on both graphs and formulate the link prediction as a supervised classification problem. We adopt Random Forest as the basic classifier to incorporate features of the scores output by unsupervised predictors, and apply bagging technique to combine multiple forests. Experiments show that with supervised method we can further improve the prediction performance.

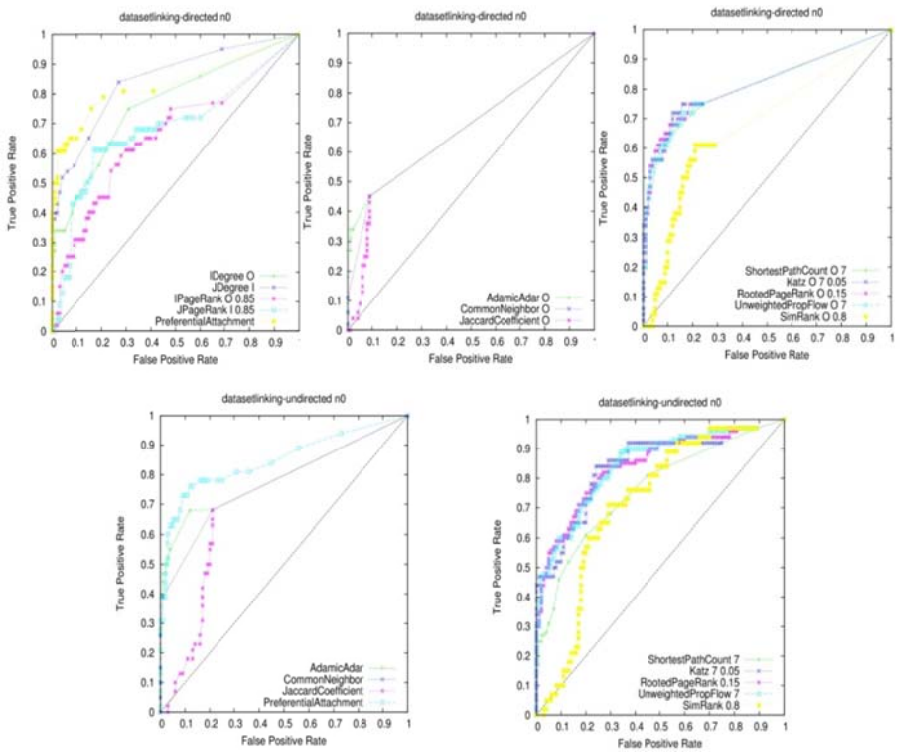

Fig. 1. ROC curves of unsupervised predictors for the directed and undirected graphs.

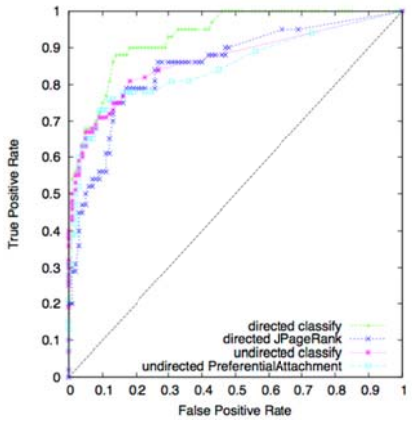

Fig. 2. Comparison of supervised and unsupervised method in ROC curves.

Table 1 AUROC of unsupervised predictors for the directed and undirected graphs.

Directed Graph			
Predictor	***AUROC***	***Predictor***	***AUROC***
IDegree_O	0.7637	AdamicAdar	0.6936
JDegree_I	0.8467	JaccardCoefficient	0.6706
IPageRank_I	0.7333	Katz_O	0.8167
JPageRank_O	0.8506	RootedPageRank_O	0.8164
PreferentialAttachment	0.8367	ShortestPathCount_O	0.8084
CommonNeighbor	0.6846	SimRank_O	0.6745
PropFlow_O	0.8118		
Undirected Graph			

Predictor	***AUROC***	***Predictor***	***AUROC***
PreferentialAttachment	0.8495	Katz	0.8433
CommonNeighbor	0.7737	RootedPageRank	0.8432
PropFlow	0.8470	ShortestPathCount	0.7631
AdamicAdar	0.7916	SimRank	0.7168
JaccardCoefficient	0.7045		

Table 2 Comparison of unsupervised and supervised method in AUROC.

	Directed Graph	**Undirected Graph**
Unsupervised	JPageRank_O 0.8506	PreferentialAttachment 0.8495
Supervised	0.9277	0.9594

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (61472436)

References

[1] Schmachtenberg M, Bizer C, Paulheim H. Adoption of the Linked Data Best Practices in Different Topical Domains[J]. Lecture Notes in Computer Science, 2014:245-260.

[2] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel- Schneider, and L. A. Stein. OWL Web Ontology Language Reference. W3C Recom- mendation. www.w3.org/TR/owl-ref (2004).

[3] Nikolov A, Motta E. What Should I Link to? Identifying Relevant Sources and Classes for Data Linking[J]. Lecture Notes in Computer Science, 2012:284-299.

[4] Liu H. et al. (2016) Identifying Linked Data Datasets for sameAs Interlinking Using Recommendation Techniques. In: Cui B., Zhang N., Xu J., Lian X., Liu D. (eds) Web-Age Information Management. WAIM 2016. Lecture Notes in Computer Science, vol 9658. Springer, Cham

[5] Lopes G. R., Leme L.A.P.P, Nunes B.P., et al. Two Approaches to the Dataset Interlinking Recommendation Problem. 2014. In: 15th International Conference on Web Information System Engineering (WISE 2014).71-74.

[6] HC Liu, PL Liu, JT Tang, H Ning, DP Wei, T Wang, Collaborative Datasets Retrieval for Interlinking on Web of Data, WWW 2015 Companion, May 18–22, 2015, Florence, Italy.

[7] Lü L, Zhou T. Link prediction in complex networks: A survey[J]. Physica A Statistical Mechanics & Its Applications, 2011, 390(6):1150–1170.