Multivariate Prediction for Learning on the Semantic Web

Yi Huang[1], Volker Tresp[1], Markus Bundschus[2]
, Achim Rettinger[3], and Hans-Peter Kriegel[2]

Siemens, Munich (1), University of Munich (2), Technical University of Munich (3)

Abstract. One of the main characteristics of Semantic Web (SW) data is that it is notoriously incomplete: in the same domain a great deal might be known for some entities and almost nothing might be known for others. A popular example is the well known friend-of-a-friend data set where, for privacy concerns and other reasons, some members document exhaustive private and social information whereas almost nothing is known for other members. Although deductive reasoning can be used to complement factual knowledge based on the ontological background, still a tremendous number of potential statements remain to be uncovered. The paper is focused on the prediction of potential relationships and attributes by exploiting regularities in the data using statistical relational learning algorithms. We argue that multivariate prediction approaches are most suitable for dealing with the resulting high-dimensional sparse data matrix. Within the statistical framework, the approach scales up to large domains and is able to deal with highly sparse relationship data. A major goal of the presented work is to formulate an inductive learning approach that can be used by people with little machine learning background. We present experimental results using a friend-of-a-friend data set.

1 Introduction

The Semantic Web (SW) is becoming a reality. Most notably is the development around the Linked Open Data (LOD) initiative. The term Linked Data is used to describe a method of exposing, sharing, and connecting data via dereferenceable Unique Resource Identifiers (URIs) on the Web. Typically, existing data sources are published in the Semantic Web's Resource Description Framework (RDF), where statements are expressed as simple subject-property-object (s, p, o) triples and are graphically displayed as a directed labeled link between a node representing the subject and a node representing the object (Figure 1). Data sources are interlinked with other data sources in the LOD cloud. In some efforts, subsets of the LOD cloud are retrieved in repositories and some form of logical reasoning is applied to materialize implicit triples. The number of inferred triples is typically on the order of the number of explicit triples. One can certainly assume that there are a huge number of true triples which are neither known as facts nor can be derived from reasoning. This might concern triples within one of the contributing data sources such as DBpedia[1] (intralinks), as well as triples describing

[1] http://dbpedia.org/

interlinks between the contributing data sources. The goal of the work presented here is to estimate the truth values of triples exploiting patterns in the data. Here we need to take into account the nature of the SW. LOD data is currently dynamically evolving and quite noisy. Thus flexibility and ease of use are preferred properties if compared to highly sophisticated approaches that can only be applied by a small number of machine learning experts. Reasonable requirements are as follows:

- Machine learning should be "push-button" requiring a minimum of user intervention.
- The learning algorithm should scale well with the size of the SW.
- The triples and their probabilities, which are predicted using machine learning, should easily be integrated into SPARQL-type querying.[2]
- Machine learning should be suitable to the data situation on the SW with sparse data (e.g., only a small number of persons are friends) and missing information (e.g., some people don't reveal private information).

Looking at the data situation, there are typically many possible triples associated with an entity (these triples are sometimes called entity molecules or, in our work, statistical unit node set) of which only a small part is known to be true. Due to the large degree of sparsity of the relationship data in the SW, multivariate prediction is appropriate for SW learning. The rows, i.e., data points in the learning matrix are defined by the key entities or statistical units in the sample. The columns are formed by nodes that represent the truth values of triples that involve the statistical units. Nodes representing aggregated information form the inputs. The size of the training data set is under the control of the user by means of sampling. Thereby the data matrix is typically independent or only weakly dependent on the overall size of the SW and in consequence the time consumption and feasibility of model training is essentially independent of the overall size of the SW. In this paper we use the friend-of-a-friend (FOAF) data set, which is a distributed social domain describing persons and their relationships in SW-format. Our approach is embedded in a statistical framework requiring the definition of a statistical unit and a population. In our experiments we compare different sampling approaches and analyze generalization on a test set.

The paper is organized as follows. In the next section we discuss how machine learning can be applied to derive probabilistic weights for triples whose truth values are unknown and introduce our approach. In Section 3 we present experimental results using friend-of-a-friend (FOAF) data. Finally, Section 4 contains conclusions and outlines further work.

2 Statistical Modeling

2.1 Defining the Sample

We must be careful in defining the statistical unit, the population, the sampling procedure and the features. A statistical unit is an object of a certain type, e.g., a person.

[2] SPARQL is a new standard for querying RDF-specific information and for displaying querying results.

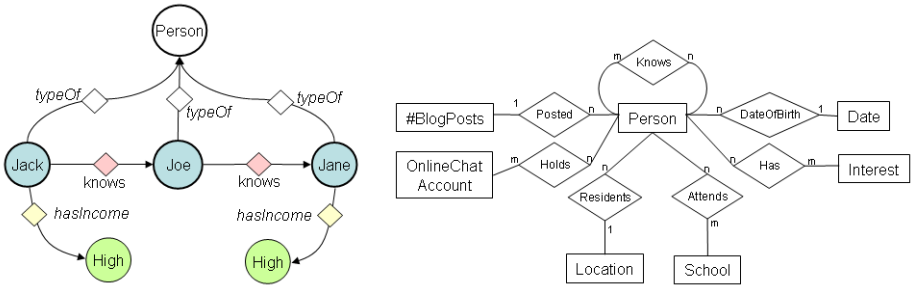

Fig. 1. Left: Example of an RDF graph displaying a social friendship network in which the income of a person is an attribute. Resources are represented by circular nodes and triples represented by labeled directed links from subject node to object node. The diamond-shaped nodes stand for random variables which are in state *one* if the corresponding triples exist. Nodes representing statistical units (here: *Persons*) have a darker rim. Right: Entity-relationship diagram of the LJ-FOAF domain

The population is the set of statistical units under consideration. In our framework, a population might be defined as the set of persons that attend a particular university. For learning we use a subset of the population. In the experimental section we will explore various sampling strategies. Based on the sample, a data matrix is generated where the statistical units in the sample define the rows.

2.2 The Random Variables in the Data Matrix

We now introduce for each potential triple a *triple node* drawn as a diamond-shaped node in Figure 1, left. A triple node is in state *one* (*true*) if the triple is known to exist and is in state *zero* (*false*) if the triple is known not to exist. Graphically, one only draws the triple nodes in state *one*, i.e., the existing triples.

We now associate some triples with statistical units. The idea is to assign a triple to a statistical unit if the statistical unit appears in the triple. Let's consider the statistical unit *Jane*. Based on the triples she is participating in, we obtain $(X, typeOf, Person)$, $(Joe, knows, X)$, and $(X, hasIncome, High)$ where X is a variable that represents a statistical unit. The expressions form the random variables (outputs) and define columns in the data matrix. By considering the remaining statistical units *Jack* and *Joe* we generate the expressions (columns), $(X, knows, Jane)$, $(Jack, knows, X)$. We will not add $(Jane, knows, X)$ since Jane considers no one in the data base to be her friend. We iterate this procedure for all statistical units in the sample and add new expressions (i.e., columns in the data matrix), if necessary. Note that expressions that are not represented in the sample will not be considered. Also, expressions that are rarely true (i.e., for few statistical units) will be removed since no meaningful statistics can be derived

from few occurrences. In [1] the triples associated with a statistical unit were denoted as *statistical unit node set* (SUNS).

2.3 Non-random Covariates in the Data Matrix

The columns we have derived so far represent truth values of actual or potential triples. Those triples are treated as random variables in the analysis. If the machine learning algorithm predicts that a triple is very likely, we can enter this triple in the data store. We now add columns that provide additional information for the learning algorithm but which we treat as covariates or fixed inputs.

First, we derive simplified relations from the data store. More precisely, we consider the expressions derived in the last subsection and replace constants by variables. For example, from $(X, knows, Jane)$ we derive $(X, knows, Y)$ and count how often this expression is true for a statistical unit X, i.e. we count the number of friends of person X.

Second, we consider a simple type of aggregated features from outside a SUNS. Consider first a binary triple $(X, knows, Jane)$. If Jane is part of another binary triple, in the example, $(X, hasIncome, High)$ then we form the expression $(X, knows, Y) \wedge (Y, hasIncome, High)$ and count how many rich friends a person has. A large number of additional aggregated features are possible but so far we restricted ourselves to these two types.

After construction of the data matrix we prune away columns which have *ones* in fewer than ϵ percent of all rows or in more than $(1 - \epsilon)$ of all rows, where ϵ is usually a very small number. Thus, we remove aggregates features that are very rarely true or almost always true, since for those no meaningful statistical analysis is possible. Note that by applying this pruning procedure we reduce the exponential number of random variables to typically a much smaller set.

2.4 Algorithms for Learning with Statistical Units Node Sets

A row in the resulting data matrix contains external inputs based on aggregated information (if available) and typically a large number of binary and sparse outputs. A *one* stands for a triple known to be true and a *zero* for a triple whose truth value is unknown. In this situation, multivariate prediction approaches have been most successful [2]. In multivariate prediction all outputs are jointly predicted such that statistical strength can be shared between outputs. The reason is that some or all model parameters are sensitive to all outputs, improving the estimates of those parameters. The approaches we are employing here are based on a matrix completion of the entire data matrix, including inputs and outputs.[3] We investigate matrix completion based on a singular value decomposition (SVD), matrix completion based on non-negative matrix factorization (NNMF) [3] and matrix completion using latent Dirichlet allocation (LDA) [4]. All three approaches estimate unknown matrix entries via a low-rank matrix approximation. SVD is based on a singular value decomposition and NNMF is a decomposition under the constraints

[3] Although the completion is applied to the entire matrix, only $zeros$ —representing triples with unknown truth values— are overwritten.

that all terms in the factoring matrices are non-negative. LDA is based on a Bayesian treatment of a generative topic model. After matrix completion of the $zero$ entries in the data matrix, the entries are interpreted as certainty values that the corresponding triples are true. After training, the models can be applied to statistical units in the population outside the sample.

3 Experiments

3.1 Data Set and Experimental Setup

Data Set: The experiments are based on friend-of-a-friend (FOAF) data. The FOAF ontology is based on RDFS/OWL and is formally specified in the FOAF Vocabulary Specification 0.91[4].

All extracted entities and relations are shown in Figure 1. In total we collected 32,062 persons and all related attributes. From this triple set, which we call full triple set, we selected 14,425 persons with a "dense" friendship information. On average, a given person has 27 friends. Then we pruned rare attributes which are associated with less than 10 persons. Table 1 lists the number of different individuals (top rows) and their known instantiated relations (bottom rows) in the full triple set, in the pruned triple set and in triples sets in different experiment settings (explained below). The resulting data matrix, after pruning, has 14,425 rows (persons) and 15,206 columns. Among those columns 14,425 ones (friendship attributes) refer to the property *knows*. The remaining 781 columns (general attributes) refer to general information about age, location, number of blog posts, attended school, online chat account and interest.

Data Retrieval and Sampling Strategies: In our experiments we evaluated the generalization capabilities of the learning algorithms given 4 different situations.

Setting 1 describes the situation where the depicted part of the SW is randomly accessible, meaning that all instances can be queried directly from triple stores. Statistical units in the sample for training are randomly sampled and statements for other randomly selected statistical units are predicted for testing (inductive setting). This way, on average persons are barely connected by the $knows$ relation. The $knows$ relation in the training and test set are very sparse (0.18%).

Setting 2 also shows the situation where statistical units in the sample are randomly selected, but this time the truth values of statements concerning the statistical units in the training sample are predicted (transductive setting). Some instances of the $knows$ relation of the selected statistical units are withheld from training and used for prediction. Prediction should be easier here since the statistics for training and prediction match perfectly

Setting 3 assumes that the Web address of one user (i.e., statistical unit) is known. Starting from this random user profile, the profiles of users connected by the $knows$ relation are gathered by crawling breadth-first and are then added to the training set.

[4] http://xmlns.com/foaf/spec/

The test set is gathered by continued crawling (inductive setting). This way all profiles are (not necessarily directly) connected and training profiles show a higher connectivity (1.02%) compared to test profiles (0.44%). In this situation generalization can be expected to be easier than setting 1 and 2 since local properties are more consistent than global ones.

Setting 4 is the combination of setting 2 and 3. The truth values of statements concerning the statistical units in the training sample are predicted (transductive setting). Instances of the $knows$ relation are withheld from training and used for prediction.

Evaluation Procedure and Evaluation Measure: The task is to predict potential friends of a person, i.e., $knows$ statements. For each person in the data set, we randomly selected one $knows$ friendship statement and set the corresponding matrix entry to $zero$, to be treated as unknown (test statement). In the test phase we then predicted all unknown friendship entries, including the entry for the test statement. The test statement should obtain a high likelihood value, if compared to the other unknown friendship entries. Here we use the normalized discounted cumulative gain (NDCG) [5] to evaluate a predicted ranking.

Benchmark methods: *Baseline:* Here, we create a random ranking for all unknown triples, i.e., every unknown triple gets a random probability assigned. *Friends of friends in second depth (FOF, d=2):* We assume that friends of friends of a particular person might be friends of that person too. From the RDF graph point of view the $knows$ relation propagates one step further alongside the existing $knows$ linkages.

		full	pruned	setting 1	setting 2	setting 3	setting 4
Concept	$Person$	32,062	14,425	4,000	2,000	4,000	2,000
#Indivi.	$Location$	5,673	320	320	320	320	320
	$School$	15,744	329	329	329	329	329
	$Interest$	4,695	118	118	118	118	118
	$On.ChatAcc.$	5	5	5	5	5	5
	$Date$	4	4	4	4	4	4
	$\#BlogPosts$	5	5	5	5	5	5
Role	$knows$	530,831	386,327	14,650	7,339	58,399	40,786
#Inst.	(sparsity)	(0.05%)	(0.19%)	train(0.18%)	(0.18%)	train (1.02%)	(1.02%)
				test (0.18%)		test (0.44%)	
	$residence$	24,368	7,964	2,228	1,106	2,389	1,172
	$attends$	31,507	5,088	1,423	747	1,467	718
	has	9,607	1,645	449	245	420	214
	$holds$	19,021	8,319	2,221	1,087	2,243	1,168
	$dateOfBirth$	10,040	5,287	1,492	715	1,563	779
	$posted$	31,959	14,369	3,985	1,992	3,985	1,994

Table 1. Number of individuals and number of instantiated relations in the full triple set, in the pruned triple set (see text) and statistics for the different experimental settings

3.2 Results

In settings 1 and 2 we randomly sampled 2,000 persons for the training set. In addition, in setting 1 we further randomly sampled 2,000 persons for the test set. In setting 3, 4,000 persons were sampled, where the first half were used for training and the second half for testing. Setting 4 only required the 2,000 persons in the training set. In each case, sampling was repeated 5 times such that error bars could be derived. Table 1 reports details of the samples (training set and, if applicable, test set). The two benchmark methods and the three matrix completion methods proposed in Section 2.4 were then applied to the training set. For each sample we repeated the evaluation procedure described above 10 times, i.e., random selection of one $knows$ relation per person to be treated as unknown. Since NNMF is only applicable in a transductive setting, it was only applied in setting 1 and 3.

The best *NDCG all* scores of all algorithms in different settings are shown in Table 2, where z indicates the number of latent variables when the best scores are achieved. Comparing the results over different settings we can easily find that for three matrix completion methods one obtains best performance in setting 4, next best performance in setting 2, then follows setting 1 and setting 3 is the most difficult. The baseline method, random guess, is independent to the settings and achieves almost the same score. A single irregularity is that FOF, d=2 in setting 2 performs better than in setting 4.

The fact that the scores in setting 4 are the best indicates that a link-following sampling strategy increases indeed the performance of learning methods. Similar results in statistical comparisons between random and network-cross sampling have been obtained in other works, e.g., [6]. On one side, the sampled persons are more likely to come from the same communities and have similar profiles so that they likely would want to know each other. On the other side, the $knows$ relation is more dense than in the case of random sampling (see Table 1). In the latter case persons more rarely have common friends. The experimental results confirm the assumption that the more sparse the matrix is, the more difficult the problem becomes since friendship patterns are more rare. In addition, we observe that the prediction performance in setting 1 is not much worse than the prediction performance in setting 2. Although from disjoint sets the statistics in training and testing are similar, leading to comparable results. Interestingly, we see that the performance of setting 3 is much worse than the prediction in setting 4. We attribute this to the general statistics in the training and the test set which are very different in setting 3. In Table 1 it is apparent that in setting 3 the $knows$ relation in the training data set (1.02%) is significantly more dense than in the test data set (0.44%). Intuitively speaking, the people in the training know each other quite well, but the people in the test do not know the people in the training as much.

4 Conclusions and Outlook

In our experiments based on the FOAF data set, LDA showed best performance, which we attribute to the fact that LDA, in contrast to NNMF and SVD, uses a Bayesian approach, which has a smaller tendency to overfitting. Thus LDA can be a default method being insensitive to exact parameter tuning. All three approaches exploited the benefits of multivariate prediction since approaches based on single predictions (not reported

Method	setting 1	setting 2	setting 3	setting 4
$Baseline$	0.1092 ± 0.0003	0.1092 ± 0.0003	0.1094 ± 0.0001	0.1094 ± 0.0001
$FOF, d = 2$	0.2146 ± 0.0095	0.2146 ± 0.0095	0.1495 ± 0.0077	0.1495 ± 0.0077
$NNMF$	NaN	0.2021 ± 0.0058 z=100	NaN	0.2983 ± 0.0197 z=150
SVD	0.2174 ± 0.0061 z=150	0.2325 ± 0.0074 z=100	0.2085 ± 0.0147 z=200	0.3027 ± 0.0179 z=100
LDA	0.2512 ± 0.0049 z=200	0.2988 ± 0.0057 z=200	0.2375 ± 0.0123 z=200	0.3374 ± 0.0117 z=200

Table 2. Best *NDCG all* and standard error where z stands for the number of latent variables

here) did not even reach the performance of the benchmark approaches. We demonstrated how probabilistic statements can be integrated into extended SPARQL queries. As example, based on the learning results for the FOAF data, one could answer queries such as: *Who would likely want to be Jack's friend; which female persons in the north-east US, would likely want to be Jack's friends.*

The approach can be extended in many ways. One might want to allow the user to specify additional parameters in the learning process, if desired, along the line of the extensions described in [7]. Another extension concerns ontological background knowledge. So far, ontological background knowledge was considered by including logically inferred statements into learning. A great advantage of the approach is that ontological knowledge is not required for the generation of the data matrix since the latter is generated based on observed SW triples. Ongoing work explores additional ways of exploiting ontological background information, e.g., for structuring the learning matrix. Similarly, we did not yet address the problem of ontology mapping and the problem of having identical entities represented on the SW under different identifiers.

Acknowledgements: We acknowledge funding by the German Federal Ministry of Economy and Technology (BMWi) under the THESEUS project and by the EU FP 7 Large-Scale Integrating Project LarKC.

References

1. Tresp, V., Huang, Y., Bundschus, M., Rettinger, A.: Materializing and querying learned knowledge. In: Proceedings of the First ESWC Workshop on Inductive Reasoning and Machine Learning on the Semantic Web. (2009)
2. Tresp, V., Yu, K.: Learning with dependencies between several response variables. In: Tutorial at ICML 2009. (2009)
3. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature (1999)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3** (2003)
5. Jarvelin, K., Kekalainen, J.: IR evaluation methods for retrieving highly relevant documents. In: SIGIR'00. (2000)
6. Neville, J., Gallagher, B., Eliassi-Rad, T.: Evaluating statistical tests for within-network classifiers of relational data. In: ICDM 2009. (2009)
7. Kiefer, C., Bernstein, A., Locher, A.: Adding data mining support to sparql via statistical relational learning methods. In: ESWC 2008, Springer-Verlag (2008)