Scalable Relation Prediction Exploiting Both Intrarelational Correlation and Contextual Information

Xueyan Jiang[2], Volker Tresp[1,2], Yi Huang[1,2],
Maximilian Nickel[2], and Hans-Peter Kriegel[2]

[1] Siemens AG, Corporate Technology, Munich, Germany
[2] Ludwig Maximilian University of Munich, Munich, Germany

Abstract. We consider the problem of predicting instantiated binary relations in a multi-relational setting and exploit both intrarelational correlations and contextual information. For the modular combination we discuss simple heuristics, additive models and an approach that can be motivated from a hierarchical Bayesian perspective. In the concrete examples we consider models that exploit contextual information both from the database and from contextual unstructured information, e.g., information extracted from textual documents describing the involved entities. By using low-rank approximations in the context models, the models perform latent semantic analyses and can generalize across specific terms, i.e., the model might use similar latent representations for semantically related terms. All the approaches we are considering have unique solutions. They can exploit sparse matrix algebra and are thus highly scalable and can easily be generalized to new entities. We evaluate the effectiveness of nonlinear interaction terms and reduce the number of terms by applying feature selection. For the optimization of the context model we use an alternating least squares approach. We experimentally analyze scalability. We validate our approach using two synthetic data sets and using two data sets derived from the Linked Open Data (LOD) cloud.

1 Introduction

There recently has been a growing interest in the prediction of the truth values of (instantiated) binary relations, i.e., grounded statements. A major reason is the growing amount of data that is published in the Linked Open Data (LOD) cloud where information is represented in the form of subject-predicate-object (s, p, o) triples. In the associated RDF graph (Resource Description Framework), entities (i.e., subjects and objects) are represented as nodes and statements are represented as directed labeled links from subject node to object node. Thus relation prediction becomes equivalent to the prediction of labeled links. In this paper we focus on the prediction of statements with a common predicate p and with defined sets of subject nodes and object nodes. We then generalize to entities not in the training set. For predicting instantiated binary relations we exploit both

intrarelational correlations and contextual information. Intrarelational correlations exploits dependencies within the relation of interest and would correspond to the data dependencies exploited in typical collaborative learning systems. Contextual information consists of all other information sources.

In the concrete examples we consider models that exploit two sources of contextual information. The first one is multi-relational contextual information that is derived from the database. The second one concerns information from unstructured sources, e.g., in form of textual documents describing the involved entities (e.g., from the entities' Wikipedia pages). As a new contribution we exploit nonlinear interactions between the associated information sources. By using low-rank approximations in the context models, the models perform latent semantic analyses and can generalize across specific terms, i.e., the model might use similar latent representations for semantically related terms. In [12] we have introduced a hierarchical Bayesian approach that is highly scalable by exploiting sparse matrix algebra, can easily generalize to new entities and does not suffer from local optima. [13] describes the additive modelling approach in greater detail. In this paper we compare the two approaches and also consider simple heuristic solutions.

The paper is organized as follows. The next section discusses related work. Section 3 describes our different ways of combining contextual information with intrarelational correlations. In section 4 we discuss how context information can be modeled and we introduce an alternating least squares solution for combining intrarelational correlations with contextual information. Section 5 contains our experimental results on synthetic data sets and on two data sets derived from the Linked Open Data (LOD) cloud. We also perform extensive experiments on scalability. Section 6 presents our conclusions.

2   Related Work

Some standard models for relational learning are, e.g., Probabilistic Relational Models [16,9], Markov Logic Networks [24] and the infinite models in [29,15]. Although conceptionally elegant, they are difficult to apply and often involve complex structural learning.[1] Our approach is related to link prediction, which is reviewed in [22,8]. SVD-based decompositions, as used in our approach, were compared to nonnegative matrix factorization (NMF) and latent Dirichlet allocation (LDA) in [10]. All three approaches benefitted greatly from regularization and then gave comparable performance. We used SVD-based decompositions since they can efficiently be computed using highly optimized packages, since predictions for new entities can be calculated easily and since they have unique solutions.

The winning entries in the Netflix competitions are based on matrix factorization [25,1,4]. The main difference is that, in those applications, unknown ratings

---

[1] As an example, we were not successful in getting the structural learning in MLNs to work in our domains.

can be treated as missing entries. In contrast, in relation prediction an instanti-
ated relationship not known to be true is very likely untrue. In the experiments
in our paper we include the hierarchical Bayesian model developed in [12]. An
advantage of that model is that it is based on a probabilistic generative model.

RFD graphs also map elegantly to a tensor representation. Tensor models for
relational learning have been explored in [20] and [21], showing both scalability
and state-of-the-art results on benchmark datasets.

Recently, there has been quite some work on the relationship between kernels
and graphs [5,27,7,3,18]. Kernels for semi-supervised learning, for example, have
been derived from the spectrum of the Graph-Laplacian. In [30,28] approaches
for Gaussian process based link prediction have been presented. Link prediction
in relational graphs has also been studied by the relational learning communities
and by the ILP communities [26,19,17]. Kernels for semantically rich domains
have been developed by [6]. Link prediction is covered and surveyed in [22,8].
Inclusion of ontological prior knowledge to relational learning has been discussed
in [23].

3 Relation Prediction by Exploiting Both Intrarelational Correlation and Context Information

3.1 Notation and Contextual Information

In this paper we assume that binary relations are presented by RDF triples of
the form (s, p, o) where subject s and object o stand for entities in a domain
and where p is the predicate. In an RDF graph, entities are nodes and a triple is
a labeled directed link from subject node to object node. Let $Z_{i,j,k}$ be a variable
assigned to the triple (s $= i$, p $= j$, o $= k$). $Z_{i,j,k} = 1$ stands for the fact that
the corresponding triple is known to exist and $Z_{i,j,k} = 0$ stands for the fact that
the corresponding triple is not known to exist.

We are now interested in a particular set of triples $\{(s = i, p = p, o = k)\}_{i,k}$ where p $= p$ is fixed and where the sets of subject and object entities
are known. Let X be the matrix of Z-values where $(X)_{i,k} = 1$ if (s $= i$, p
$= p$, o $= k$) is known to exist; otherwise $(X)_{i,k} = 0$. In the following we will
derive a number of matrices where the zeros are replaced by continuous numbers
that can be interpreted as confidence values for a relation being true, based on
the available evidence. In a probabilistic sense, we can interpret the continuous
numbers as $P((X)_{i,k} = 1|Data)$. These confidence values can then be the basis
for classification and ranking tasks as described in Section 5.

In this paper we assume that contextual information is available from which
we can derive an estimate of how likely a target relation is true, denoted by

$$f_{i,k}.$$

Contextual information might consist of other statements in the knowledge
base relevant for the relation under consideration, but could also include un-
structured information, e.g., textual documents describing the involved entities

(see Section 4). The corresponding matrix F with $(F)_{i,k} = f_{i,k}$ has the same dimensionality as X and entries typically assume values between zero and one, although this is not enforced.

Our first and most simple estimate for the confidence values for statements would be simply derived from this estimate and represents a pure context-based model of the form

$$X_{\mathrm{F}} = F. \tag{1}$$

Here we do not exploit intrarelational correlations and this solution is sensible if X is very sparse. Alternatively, we might trust the ones in the X matrix (which represent certain facts) and use

$$X_{\mathrm{M}} = \max(X, F) \tag{2}$$

where max is applied componentwise or, if we are willing to tolerate confidence values greater than one,

$$X_{\mathrm{S}} = X + F. \tag{3}$$

In both solutions, F is mostly relevant to the zero entries of X.

3.2   Intrarelational Correlations

In many applications the correlations in the relational matrix can be exploited to derive predictions, an effect often associated with collaborative filtering. The leading approaches exploiting intrarelational correlations are based on a factorization of X, which is also the approach we are taking.

We propose to minimize the cost function

$$\min_{X_{\mathrm{r}}} \|X - X_{\mathrm{r}}\|_F^2$$

where we impose the constraint on X_{r} to have a maximum rank of r.

It is well known that one specific solution can be derived from a singular value decomposition (SVD) with

$$X = UDV^T \tag{4}$$

where U and V are matrices with orthonormal columns and where D is a diagonal matrix. The diagonal entries $d_i \geq 0$ are ordered according to magnitude. The optimal r-rank reconstruction can be written as

$$X_{\mathrm{r}} = U_r D_r V_r^T$$

where U_r and V_r contain the first r columns of the respective matrices and where D_r is a diagonal matrix with the r leading components of D. Low-rank reconstructions are used in latent semantic analysis to generalize from observed terms to related terms, they are an important ingredient in the winning entries in the Netflix competition [25,1,4], and they also give very good results in predicting links in semantic graphs [10].

We also consider a regularized version which typically improves predictions significantly by using the cost function

$$\min_{\tilde{W}} \left(\|X - \tilde{W} X_{\mathrm{r}}\|_F^2 + \lambda \|\tilde{W}\|_F^2 \right)$$

where X_{r} is fixed and where the parameter matrix $\tilde{W}$ is optimized.

The overall solution is then[2]

$$X_{\mathrm{CF}} = U_r \ \mathrm{diag} \left\{ \frac{d_i^3}{d_i^2 + \lambda} \right\}_{i=1}^r V_r^T$$

$$= U_r \ \mathrm{diag} \left\{ \frac{d_i^2}{d_i^2 + \lambda} \right\}_{i=1}^r U_r^T X = X V_r \ \mathrm{diag} \left\{ \frac{d_i^2}{d_i^2 + \lambda} \right\}_{i=1}^r V_r^T \qquad (5)$$

where $\mathrm{diag} \left\{ \frac{d_i^2}{d_i^2 + \lambda} \right\}_{i=1}^r$ is an $r \times r$ diagonal matrix with r diagonal entries. In the following we assume that X has fewer rows than columns such that U_r is fast to compute based on an SVD of the kernel matrix XX^T, but one should simply apply the reconstruction most suitable.

We can easily generalize to a new subject entity with x^{new} (as column vector) using

$$x_{\mathrm{CF}}^{new} = V_r \ \mathrm{diag} \left\{ \frac{d_i^2}{d_i^2 + \lambda} \right\}_{i=1}^r V_r^T x^{new} = X^T U_r \ \mathrm{diag} \left\{ \frac{1}{d_i^2 + \lambda} \right\}_{i=1}^r U_r^T X x^{new}. \quad (6)$$

We can now include contextual information by adding the context matrix and the intrarelational module and obtain as a heuristics

$$X_{\mathrm{H}} = X_{\mathrm{CF}} + F. \qquad (7)$$

Note that in contrast to X_{S}, here we use X_{CF} instead of X and we obtain a combination model that exploits correlations in X. Thus we will get a high score for a link, if either the context model or the intrarelational model (or both) is positive about the link.

3.3   Hierarchical Bayes

So far, the combination scheme in Equation 7 might be considered a plausible heuristic. In this section and in the next section we consider two combination schemes that can be derived from principled approaches.

In [12] we described a hierarchical Bayesian (HB) approach for the combination of contextual information with intrarelational correlation. It motivates the following approach: We are searching for the low-rank approximation X_{HBS} that minimizes

$$\min_{X_{\mathrm{HBS}}} \|X_S - X_{\mathrm{HBS}}\|_F^2$$

---

[2] Here and in the following we have typically these three ways of formulating the solution. One should take the one which is most efficient considering the dimensionalities of the involved matrices.

where $X_S = X + F$ was defined in Equation 3. Again, the solution can be based on the SVD, in this case in the form of

$$X + F = U^{MF} D^{MF} V^{MF^T}$$

and a regularized low-rank approach now leads to the model

$$X_{\text{HBS}} = U_r^{MF} \text{ diag} \left\{ \frac{(d_i^{MF})^2}{(d_i^{MF})^2 + \lambda} \right\}_{i=1}^r U_r^{MF^T} (X + F). \tag{8}$$

A similar solution, i.e., X_{HBM}, is obtained if we use X_M instead of X_S. Note that for X_H we first smooth X via a regularized low-rank approximation and then add F, whereas for M_{HBS} we first add X and F and then smooth the resulting matrix.

Alternatively we can use as a basis the decomposition of X instead of the decomposition of $X + F$ and obtain

$$X_{\text{HBS2}} = U_r \text{ diag} \left\{ \frac{d_i^2}{d_i^2 + \lambda} \right\}_{i=1}^r U_r^T (X + F)$$

$$= X_{\text{CF}} + U_r \text{ diag} \left\{ \frac{d_i^2}{d_i^2 + \lambda} \right\}_{i=1}^r U_r^T F. \tag{9}$$

For X_{HBS2} we can exploit sparse matrix algebra for calculating the decomposition of X (whereas $X + F$ is typically non sparse) and F only needs to be calculated for the entities of interest. Interestingly, the solution consists of adding to X_{CF} a regularized projections of F using the *largest* singular values, so we add to X_{CF} a "low-frequency" version of F.

3.4   Additive Models

The idea here is that the intrarelational correlations should only model the residual difference after F has been subtracted from X. The goal is then to minimize the cost function

$$\min_{X_{\text{CFa}}} \|X - (F + X_{\text{CFa}})\|_F^2.$$

A regularized low-rank approximation where the basis is calculated from the decomposition of X is then

$$X_{\text{CFa}} = U_r \text{ diag} \left\{ \frac{d_i^2}{d_i^2 + \lambda} \right\}_{i=1}^r U_r^T (X - F) \tag{10}$$

and the overall prediction is

$$X_{\text{add}} = X_{\text{CFa}} + F$$

such that

$$X_{\text{add}} = X_{\text{CFa}} + F = U_r \ \text{diag} \left\{ \frac{d_i^2}{d_i^2 + \lambda} \right\}_{i=1}^r U_r^T (X - F) + F$$

$$= X_{\text{CF}} + U \left(I - \text{diag} \left\{ \frac{d_i^2}{d_i^2 + \lambda} \right\}_{i=1}^r \right) U^T F. \tag{11}$$

Interestingly, the solution consists of adding to X_{CF} a "high frequency" version of F.

In the next section we derive specific models for F. An overall additive model where F and X_{add} are adapted in turn (the latter using Equation 10) and where Equation 11 is used for overall prediction is defined as X_{global}.

4   Context Models for Our Applications

4.1   Context Models Based on the Database

So far, f could have been an arbitrary function of context information. We see this as a great advantage of our approach since it permits a great modularity and the context model and the intrarelational model can be optimized independently.

Now we derive a specific context model that we will use in the applications. Let's consider a multi-relational database of triples (i.e., a triple store). Let A be a matrix with as many rows as X, i.e., with one row for each subject entity in X. The columns of A represent features describing the subjects in X. In the simplest case they consist of the truth value of all (relevant) triples with the same subject. Consider the example that rows are users and columns are movies and the task is to predict if a user watches a movie. In this example, a particular column in A might indicate if a user is of young age and the model would be able to exploit the preference of young people for certain movies.

Similarly, B is a matrix. The number of rows of B is equal to the number of columns of X. The columns of B represent features describing the objects in X. In the simplest case they consist of the truth values of all (relevant) triples, where the object of X is the subject. Following the example, a column in B might indicate if a movie is an action movie and the model can exploit the preference of some people for action movies. Thus B is suitable to model personal preferences.

Finally, we introduce the matrix C formed by the Kronecker product $C = A \otimes B$, i.e., C contains all possible product terms of the elements of A and B. The number of rows in C is the number of rows of A times the number of rows of B and the number of columns in C is the number of columns of A times the number of columns of B. Following the example, a column in C might indicate if a movie is an action movie and, at the same time, the user is young and the model might learn that young people like action movies.

We now write a least squares cost function

$$\|X - F\|_F^2$$

where

$$F = AW^A + (BW^B)^T + \text{matrix}(Cw^C) \tag{12}$$

and where $\text{matrix}(\cdot)$ transforms the vector into a matrix of appropriate dimensions. $\|\cdot\|_F$ is the Frobenius norm. The matrices W^A and W^B and the vector w^C contain the parameters to be optimized. Thus we predict the entries in F as a linear combination of the subject features in A, the column features in B and the interaction features in C.

To control overfitting, we add to the cost functions the penalty terms $\lambda_A\|W^A\|_F^2$, $\lambda_B\|W^B\|_F^2$, and $\lambda_C\|w^C\|_F^2$.

To reduce the amount of computation and also as a means to prevent overfitting, we are looking for low-rank solutions with ranks r_A, r_B, and r_C as discussed in the next subsection. By using low-rank models, the models perform latent semantic analyses and can generalize across specific terms, i.e., the model might use similar latent representation for semantically related terms.

The number of interaction terms in C can easily be several millions, so we perform fast feature selection strategies by evaluating the Pearson correlation between targets and features.

4.2   Alternating Least Squares

An easy way to optimize the cost function is to repeatedly iterate over all three terms where in each iteration we keep the other two fixed.

Let

$$X^{-A} = X - \left((BW^B)^T + \text{matrix}(Cw^C)\right)$$
$$X^{-B} = X - \left(AW^A + \text{matrix}(Cw^C)\right)$$
$$X^{-C} = X - \left(AW^A + (BW^B)^T\right)$$

and let $x^{-C} = \text{vec}(X^{-C})$.

The individuals contributions are the calculated as

$$AW^A = U_{A,r_A} \text{diag}\left\{\frac{(d_i^{(A)})^2}{(d_i^{(A)})^2 + \lambda_A}\right\}_{i=1}^{r_A} U_{A,r_A}^T X^{(-A)} \tag{13}$$

$$BW^B = U_{B,r_B} \text{diag}\left\{\frac{(d_i^{(B)})^2}{(d_i^{(B)})^2 + \lambda_B}\right\}_{i=1}^{r_B} U_{B,r_B}^T (X^{(-B)})^T \tag{14}$$

$$Cw^C = U_{C,r_C} \text{diag}\left\{\frac{(d_i^{(C)})^2}{(d_i^{(C)})^2 + \lambda_C}\right\}_{i=1}^{r_C} U_{C,r_C}^T x^{(-C)} \tag{15}$$

where we have used the singular value decompositions (SVD)

$$A = U_A D_A V_A^T \quad B = U_B D_B V_B^T \quad C = U_C D_C V_C^T \tag{16}$$

and where U_{A,r_A} contains the first r_A columns of U_A, U_{B,r_B} contains the first r_B columns of U_B, and U_{C,r_C} contains the first r_C columns of U_C.

Again we can exploit sparse matrix algebra for calculating the decompositions. The convergence of the alternating least squares algorithm is quite fast, requiring fewer than 10 iterations.

Note again that we can include the intrarelational model as an additional fourth component to be optimized with alternating least squares leading to the model X_{global} introduced in Section 3.4. A more extensive analysis of the additive models can be found in [13] where also additional feature candidates are discussed. Since the bases for the decompositions (calculated in Equations 4 and 16) are calculated before the optimization of the parameters, the alternating least squares iterations converge to unique solutions. [3]

4.3   Incorporating External Information Sources and Aggregation

In the applications we are considering we sometimes have available textual data describing the involved entities. We simply treat the keywords in the textual descriptions as additional features describing subjects, resp. objects. In some applications, it is useful to add aggregated information. This can be represented as additional features as well.

## 5   Experiments

5.1   Scalability

For the kind of relational data that we are considering, X is very sparse and the reduced-rank reconstruction can be calculated efficiently. Figure 1 shows experimental results. Note that for a sizable X-matrix with 10^5 rows, 10^6 columns, 10^7 nonzero elements and a rank of $r = 50$, the computation only takes approximately 10 minutes on a standard laptop computer. For matrices where $K = XX^T$ becomes dense one might employ the alternating least squares solution described in [20] that does not rely on a sparsity of K in the factorization and does not enforce orthogonality constraints.

5.2   Tuning of Hyperparameters

The approaches contain up to 8 hyperparameters (r, r_A,r_B,r_C, λ, λ_A,λ_B,λ_C) which are tuned using cross-validation sets (i.e. they are not tuned on the test set). We follow the approach described in [2] and perform a random search for the best hyperparameters.

---

[3] Recall that we first calculate the kernel matrix $K = XX^T$ and then perform the SVD decomposition. Naturally, we could start with a kernel matrix suitable for the RDF graph. In this view our alternating least squares solution is an efficient way of calculating a kernel solution with a kernel $k(s, s', o, o') = k_{CF}(s, s') + k_A(s, s') + k_B(o, o') + k_C(s, s', o, o')$ where $k_{CF}(s, s')$ is the intrarelational kernel, $k_A(s, s')$ is a kernel for subject nodes, $k_B(o, o')$ is a kernel for object nodes, and $k_C(s, s', o, o')$ is a kernel for modeling interactions.
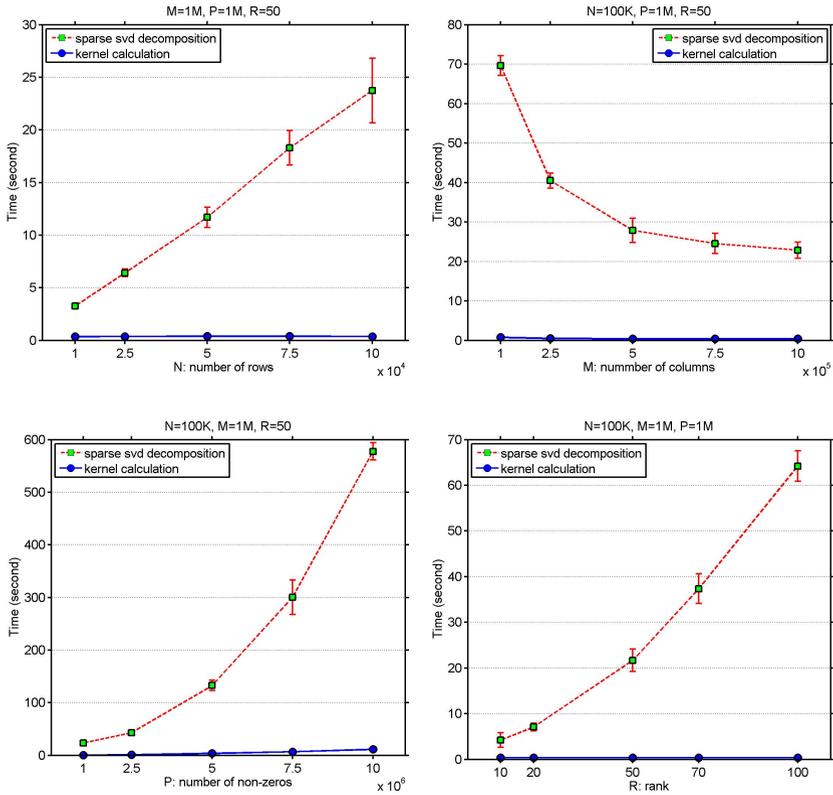
Fig. 1. We consider a sparse random $N \times M$ matrix X. First we construct the kernel matrix via $K = XX^T$ and then use sparse SVD to obtain U_r. The top left figure shows computational time for the SVD as a function of N (red dashed). We see approximately a linear dependency which is related to the fact that the number of rows of U is N as well. In this experiment, $r = 50$, $M = 10^6$ and the number of nonzero entries in X is $p = 10^6$. The top right figure shows computation time for the SVD as a function of M (red dashed). We see a decrease: the reason is that with increasing M, K becomes less dense. We used $p = 10^6$, $N = 10^5$, and $r = 50$. The bottom left shows an approximately quadratic dependency of the computational time for the SVD on p ($M = 10^6$, $N = 10^5$, $r = 50$) (red dashed). Note that the last data point in the plot is a system with $p = 10^7$ requiring only 10 minutes of computation. Finally, the bottom right figure shows the dependency on r ($M = 10^6$, $N = 10^5$, $p = 10^6$) (red dashed). A 10 fold increase in r approximately displays a 10 fold increase in computational cost. Each figure also shows the computational time for calculating $K = XX^T$, which, in comparison, is negligible (blue continuous). A prediction for data for a novel subject (i.e., a new row in X) can efficiently be calculated using Equation 6.

5.3   Synthetic Data

The synthetic data has been generated according to our modeling assumptions. The target relation is a sum of four components: the first one is modeling the intrarelations correlation, the second one uses features describing the subject entities, the third one uses features describing the object entities, and the fourth one uses interaction terms. In the first experiment, both the intrarelational correlation and the context models have predictive power and all six combination schemes improve upon the subsystems. The additive models X_{add} and X_{global} seem to be more robust and perform well on both experiments.

We randomly selected one true relation to be treated as unknown (test statement) for each subject entity in the data set. In the test phase we then predicted all unknown relations for the entity, including the entry for the test statement. The test statement should obtain a high likelihood value, if compared to the other unknown entries. The normalized discounted cumulative gain (nDCG@all) [11] is a measure to evaluate a predicted ranking.

5.4   Associating Diseases with Genes

As the costs for gene sequencing are dropping, it is expected to become part of clinical practice. Unfortunately, for many years to come the relationships between genes and diseases will remain only partially known. The task here is to predict diseases that are likely associated with a gene based on knowledge about gene and disease attributes and about known gene-disease patterns.

Disease genes are those genes involved in the causation of, or associated with a particular disease. At this stage, more than 2500 disease genes have been discovered. Unfortunately, the relationship between genes and diseases is far from simple since most diseases are polygenic and exhibit different clinical phenotypes. High-throughput genome-wide studies like linkage analysis and gene expression profiling typically result in hundreds of potential candidate genes and it is still a challenge to identify the disease genes among them. One reason is that genes can often perform several functions and a mutational analysis of a particular gene reveals dozens of mutation cites that lead to different phenotype associations to diseases like cancer [14]. An analysis is further complicated since environmental and physiological factors come into play as well as exogenous agents like viruses and bacteria.

Despite this complexity, it is quite important to be able to rank genes in terms of their predicted relevance for a given disease as a valuable tool for researchers and with applications in medical diagnosis, prognosis, and a personalized treatment of diseases.

In our experiments we extracted information on known relationships between genes and diseases from the LOD cloud, in particular from Linked Life Data and Bio2RDF, forming the triples (Gene, related_to, Disease). In total, we considered 2462 genes and 331 diseases. For genes we extracted 11332 features and for the diseases 1283 features from the LOD cloud. In addition, we retrieved 8000 textual features describing genes and 3800 textual features describing diseases
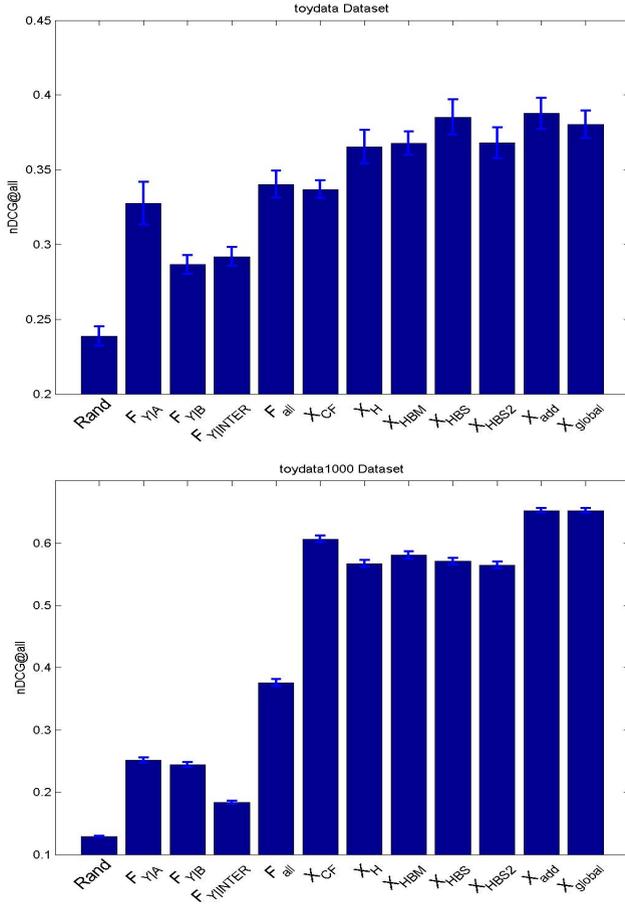
Fig. 2. Test results on synthetic data. In the first experiment (top), we had 100 subjects and 80 objects, A had 80 columns, B had 80 columns, and C had 8000 rows and 4000 columns. The three context models $F_{Y|A}$, $F_{Y|B}$, and $F_{Y|INTER}$ make valuable predictions significantly above random. The combination of all three context models, i.e., $X_F = F_{\text{all}}$, is better than any of the individual context models. The context model and the intrarelation correlation are comparable strong in prediction: X_{CF} gives comparable results to F_{all}. All six combination schemes are better than the intrarelational model or the context model on their own, so all combination schemes are sensible. The additive model X_{add} and the additive model where the context model and the context model are jointly optimized (X_{global}) perform best, although there is no statistical significant difference between the 6 combination models. In the second experiment (bottom), we had 1000 subjects and 1000 objects, A had 6 columns, B had 7 columns, and C had 1000000 rows and 42 columns. Thus the intrarelation correlation is stronger than the contextual model. X_{add} and X_{global} show better performance than X_{CF} and F_{all} individually.
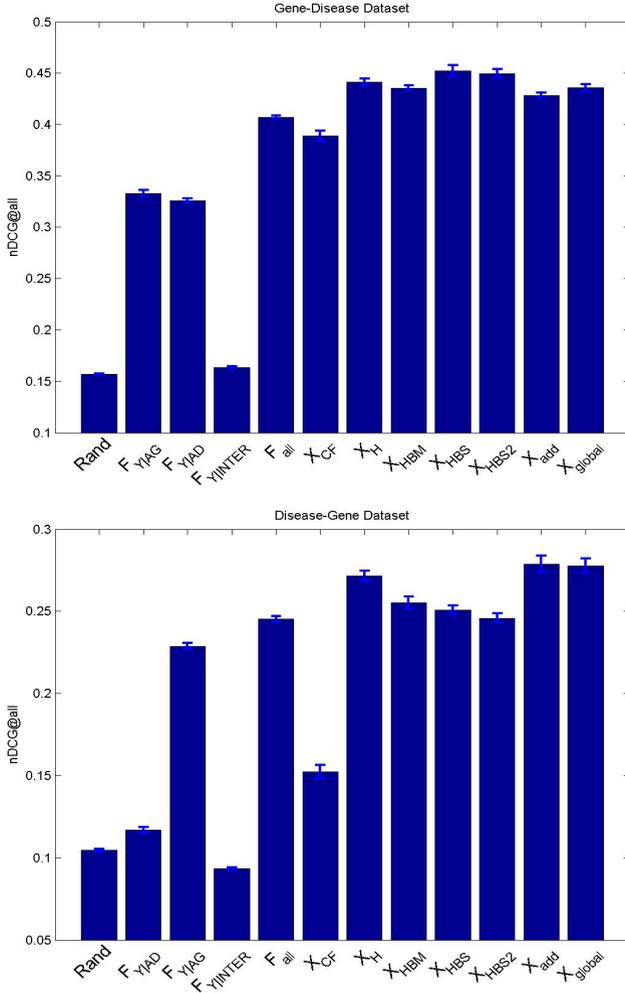
Fig. 3. The goal is to predict the relationship between genes and diseases. On the top we ranked recommended diseases for genes and on the bottom we ranked recommended genes for diseases. We considered contextual features from disease attributes $F_{Y|AD}$ and from gene attributes $F_{Y|AG}$, and contribution from the interaction term $F_{Y|INTER}$. The combination of all contextual models in F_{all} is better than the individual context models where $F_{Y|INTER}$ is not better than random. All six combination schemes are better than the intrarelational model or the context model on their own, so all combination schemes are sensible. In this experiment, two of the hierarchical Bayes models, i.e., X_{HBS} and X_{HBS2} give best results. The results are generally better than the results reported in [12] since, there, only contextual features from text documents were used. The second task, predicting genes for diseases, is more difficult due to the great number of potential genes. Intrarelational correlation on its own is relatively weak (X_{CF}). Again, all combination schemes give good results.
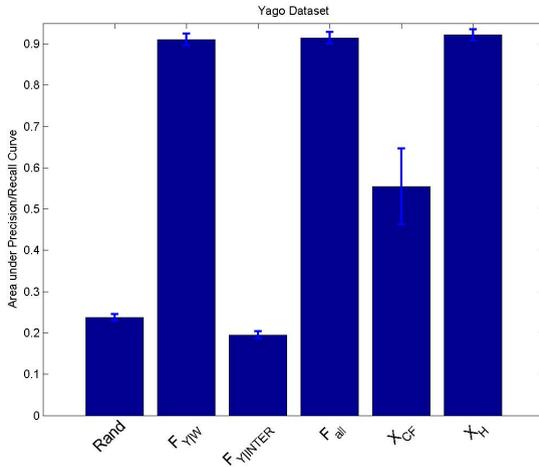
Fig. 4. The task is to predict the nationalities of writers. The writer attributes $F_{Y|W}$ have considerable predictive power. The intrarelational correlation (X_{CF}) benefits from the imbalance of classes. We display the area under precision/recall curve on writers not in the training set (induction). None of the combination models is significantly better than $F_{Y|W}$, which in this experiment is reasonable, since very few writers have more than one nationality (of the combination schemes, we only show X_H).

from corresponding text fields in Linked Life Data and Bio2RDF. After applying feature selection, the interaction matrix C had 814922 rows and 1133 columns.

Figure 3 shows the results. This is a very interesting data set: when predicting diseases for genes, the contextual information (reflected in F_{all}) and the intrarelation correlational (reflected in X_{CF}) are both equally strong; in most data sets, one of the two is dominating. All six combination schemes are effective and provide results significantly better than F_{all} or X_{CF} on their own. Predicting genes for diseases generally gives a weaker nDCG score and the leading approaches are X_{add} and X_{global}.

5.5   Predicting Writer's Nationality in YAGO2

The final set of experiments was done on the YAGO2 semantic knowledge base. YAGO2 is derived from Wikipedia and also incorporates WordNet and GeoNames. There are two available versions of YAGO2: core and full. We used the first one which currently contains 2.6 million entities, and describes 33 million facts about these entities. Our experiment was designed to predict the nationalities of writers. We choose four different types of writers: American, French, German and Japanese.

We obtained 440 entities representing the selected writers. We selected 354 entities (i.e., writers) and added textual information describing the writers and the countries. We performed 10-fold cross validation for each model, and

evaluated them with the area under precision and recall curve. Figure 4 shows the results. As there are only 4 nationalities, which are almost always mutual exclusive (there is a small number of writers with more than one nationality), the intrarelational correlation is quite weak and the country attributes were not used. Interestingly, the interaction term is reasonable strong ($F_{Y|INTER}$). In fact, no model is better than $F_{Y|W}$ which only exploits the contextual information of the writers.

6   Conclusions

In this paper we have considered the problem of predicting instantiated binary relations in a multi-relational setting and exploit both intrarelational correlations and contextual information. We have presented a number of sensible algorithms. The algorithms are all modular and have unique solutions. As contextual information we consider information extracted from the database and textual data describing the entities. To include contextual information we use an alternating least squares approach that includes models for subject features, object features and an interaction model. By using low-rank approximations in the context models, the models perform latent semantic analyses and can generalize across specific terms, i.e., the model might use similar latent representation for semantically related terms. The approaches can exploit sparse matrix algebra and, as we have demonstrated experimentally, are highly scalable. The models can easily be applied to new entities not considered in model training. We presented experimental results on synthetic data, on life science data from the Linked Open Data (LOD) cloud. All the presented combination schemes are effective and there is no clear best approach, although there seems to be a general advantage for the additive models X_{add} and X_{global}.

References

1. Bell, R.M., Koren, Y., Volinsky, C.: All together now: A perspective on the netflix prize. Chance (2010)
2. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research (2012)
3. Bloehdorn, S., Sure, Y.: Kernel methods for mining instance data in ontologies. In: ESWC (2007)
4. Candes, E.J., Recht, B.: Exact matrix completion via convex optimization. Computing Research Repository - CORR (2008)
5. Cumby, C.M., Roth, D.: On kernel methods for relational learning. In: ICML (2003)
6. D'Amato, C., Fanizzi, N., Esposito, F.: Non-parametric statistical learning methods for inductive classifiers in semantic knowledge bases. In: IEEE International Conference on Semantic Computing, ICSC (2008)
7. Gärtner, T., Lloyd, J.W., Flach, P.A.: Kernels and distances for structured data. Machine Learning (2004)
8. Getoor, L., Diehl, C.P.: Link mining: a survey. SIGKDD Explorations (2005)

9. Getoor, L., Friedman, N., Koller, D., Pfeffer, A., Taskar, B.: Probabilistic relational models. In: Introduction to Statistical Relational Learning (2007)
10. Huang, Y., Tresp, V., Bundschus, M., Rettinger, A., Kriegel, H.-P.: Multivariate Prediction for Learning on the Semantic Web. In: Frasconi, P., Lisi, F.A. (eds.) ILP 2010. LNCS, vol. 6489, pp. 92–104. Springer, Heidelberg (2011)
11. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: SIGIR 2000 (2000)
12. Jiang, X., Huang, Y., Nickel, M., Tresp, V.: Combining Information Extraction, Deductive Reasoning and Machine Learning for Relation Prediction. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 164–178. Springer, Heidelberg (2012)
13. Jiang, X., Tresp, V., Huang, Y., Nickel, M., Kriegel, H.-P.: Link Prediction in Multi-relational Graphs using Additive Models (submitted, 2012)
14. Kann, M.G.: Advances in translational bioinformatics: computational approaches for the hunting of disease genes. In: Briefings in Bioinformatics (2010)
15. Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. In: AAAI (2006)
16. Koller, D., Pfeffer, A.: Probabilistic frame-based systems. In: AAAI (1998)
17. Landwehr, N., Passerini, A., De Raedt, L., Frasconi, P.: kFOIL: Learning simple relational kernels. In: AAAI (2006)
18. Lösch, U., Bloehdorn, S., Rettinger, A.: Graph Kernels for RDF Data. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 134–148. Springer, Heidelberg (2012)
19. Muggleton, S.H., Lodhi, H., Amini, A., Sternberg, M.J.E.: Support Vector Inductive Logic Programming. In: Hoffmann, A., Motoda, H., Scheffer, T. (eds.) DS 2005. LNCS (LNAI), vol. 3735, pp. 163–175. Springer, Heidelberg (2005)
20. Nickel, M., Tresp, V., Kriegel, H.-P.: A three-way model for collective learning on multi-relational data. In: ICML (2011)
21. Nickel, M., Tresp, V., Kriegel, H.-P.: Factorizing YAGO: scalable machine learning for linked data. In: WWW (2012)
22. Popescul, A., Ungar, L.H.: Statistical relational learning for link prediction. In: Workshop on Learning Statistical Models from Relational Data (2003)
23. Rettinger, A., Nickles, M., Tresp, V.: Statistical Relational Learning with Formal Ontologies. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part II. LNCS, vol. 5782, pp. 286–301. Springer, Heidelberg (2009)
24. Richardson, M., Domingos, P.: Markov logic networks. In: Machine Learning (2006)
25. Takacs, G., Pilaszy, I., Nemeth, B., Tikk, D.: On the gravity recommendation system. In: Proceedings of KDD Cup 2007 (2007)
26. Taskar, B., Wong, M.F., Abbeel, P., Koller, D.: Link prediction in relational data. In: NIPS (2003)
27. Vishwanathan, S.V.N., Schraudolph, N., Kondor, R.I., Borgwardt, K.: Graph kernels. Journal of Machine Learning Research - JMLR (2008)
28. Xu, Z., Kersting, K., Tresp, V.: Multi-relational learning with gaussian processes. In: IJCAI (2009)
29. Xu, Z., Tresp, V., Yu, K., Kriegel, H.-P.: Infinite hidden relational models. In: UAI (2006)
30. Yu, K., Chu, W., Yu, S., Tresp, V., Xu, Z.: Stochastic relational models for discriminative link prediction. In: NIPS (2006)