

A Gaussian Process Model for Knowledge Propagation in Web Ontologies

Pasquale Minervini, Claudia d’Amato, Nicola Fanizzi, Floriana Esposito
Department of Computer Science - Università degli Studi di Bari Aldo Moro, Italy
{firstname.lastname}@uniba.it

Abstract—We consider the problem of predicting missing class-memberships and property values of individual resources in Web ontologies. We first identify which relations tend to link similar individuals by means of a finite-set Gaussian Process regression model, and then efficiently propagate knowledge about individuals across their relations. Our experimental evaluation demonstrates the effectiveness of the proposed method.

I. INTRODUCTION

Standard query answering and reasoning services for the Semantic Web [1] (SW) rely on deductive inference. However, purely deductive approaches may suffer from limitations, owing to: i) the complexity of reasoning tasks on expressive representations, ii) the inherent incompleteness of SW knowledge bases (KBs), and iii) the presence of logically conflicting knowledge fragments. Deciding on the truth of specific facts (assertions) in SW KBs requires to take into account the *open-world* semantics adopted when reasoning in this context: a failure on deciding the truth value of a given fact does not imply that such fact is false, but rather that its truth value cannot be deductively inferred from the KB; this differs from the *Negation As Failure*, commonly used with databases. Other issues are related to the distributed nature of the data across the Web: mutually conflicting pieces of knowledge may lead to flawed inferences and contradictory answers. Estimating the truth value of an assertion can be cast as a *statistical inference* problem: individual resources in an ontology can be regarded as statistical units, and their properties can be statistically inferred, even when they cannot be deduced from the knowledge base. Several approaches have been proposed in the SW literature (see [2] for a survey). A major issue with the methods proposed so far is that the induced statistical models are either difficult to interpret and understand by experts and to integrate in logic-based SW infrastructures, or computationally impractical when used on real KBs.

Related Work: A variety of methods have been proposed for predicting the truth value of assertions in Web ontologies. They include generative models, kernel methods (e.g. [3] and [4]), and matrix or tensor factorization methods (e.g. [5] and [6]). An issue with existing methods is that they either rely on a possibly expensive search process, or induce statistical models that are often not easy to interpret by human experts. Kernel methods induce models in a high-dimensional feature space implicitly defined by a kernel function. The underlying kernel function itself usually relies on purely syntactic features of the neighborhood graphs of two individual resources, such as their common subtrees [3], or isomorphic subgraphs [4]: in both cases, there is not necessarily a direct interpretation of such features in terms of domain knowledge. Latent variable

models and matrix or tensor factorization methods such as [5] and [6] try to explain the observations in terms of latent classes and attributes, which also may be non-trivial to interpret in terms of the domain’s vocabulary.

Contribution: We propose a transductive inference method for predicting the truth value of assertions, which is based on the following intuition: examples that are *similar* in some aspects tend to be linked by specific relations. Our method aims at identifying such relations, and permits the efficient propagation of information along chains of related entities. It is especially useful with real world *shallow* ontologies, i.e. those with a relatively simple fixed terminology and populated by very large amounts of data, where related individuals tend to influence each other, such as social or citation networks. In particular, this paper makes the following contributions:

- A method for efficiently *propagating* knowledge among similar examples: it leverages a similarity graph which plays a critical role in the knowledge propagation process.
- A method for *learning* an optimal similarity graph for a given prediction task, by leveraging a set of semantically heterogeneous relations among examples.

This paper is organized as follows: in Sect. II we introduce the problem of *transductive learning* in the context of semantic knowledge bases; in Sect. III we discuss the proposed method, which is based on the efficient propagation of knowledge among similar examples, and address the problem of identifying which relations are likely to link similar examples; in Sect. IV we provide empirical evidence for the effectiveness of the proposed method; in Sect. V we summarize this work.

II. TRANSDUCTIVE LEARNING WITH WEB ONTOLOGIES

We assume the knowledge base is encoded in a syntactic variant of some *Description Logic* [7] (DL). Basic elements are *atomic concepts* $N_C = \{C, D, \dots\}$ interpreted as subsets of a domain of objects (e.g. `Person` or `Article`), and *atomic roles* $N_R = \{R, S, \dots\}$ interpreted as binary relations on such a domain (e.g. `friendOf` or `authorOf`). Domain objects are represented by *individuals* $N_I = \{a, b, \dots\}$, each associated to a domain entity. Specifically, we consider KBs in the OWL 2 language¹, which has its theoretical foundations in DLs: concepts and roles are referred to as *classes* and *properties*, respectively. A DL *knowledge base* (KB) $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ is composed by two main components: a *TBox* \mathcal{T} , which contains terminological axioms, and an *ABox* \mathcal{A} , which contains ground axioms (*assertions*) about individuals. In the following, we

¹OWL 2 W3C Recommendation: <http://www.w3.org/TR/owl-overview/>

denote the set of individuals occurring in \mathcal{A} as $\text{Ind}(\mathcal{A})$. Let Q be a query concept and a an individual in a KB \mathcal{K} : *Instance Checking* consists in deciding whether $\mathcal{K} \models Q(a)$ holds. SW inference services make the *Open-World Assumption* (OWA), i.e. it might happen that $\mathcal{K} \not\models Q(a)$ and $\mathcal{K} \not\models \neg Q(a)$, where $\neg Q$ is the complement of Q . Given a (infinite) set of variables N_V , a *Conjunctive Query* (CQ) q is a conjunction of concept or role atoms $C(v)$ or $R(v, v')$, with $v, v' \in N_V \cup N_I$, built on the signature of \mathcal{K} . A binding of the variables w.r.t. some model of \mathcal{K} determines a result via the value of answer variables.

In this work we focus on a transductive learning problem: given a set of labeled and unlabeled examples, the problem consists in learning a *labeling function* for a given target class that can be used for predicting whether individuals belong to a class C (positive class) or to its complement $\neg C$ (negative class) when this cannot be inferred deductively. Formally:

Definition 2.1 (Transductive Class-Membership Learning):

Given: a *target class* C in a KB \mathcal{K} , and a set of examples $X \subseteq \text{Ind}(\mathcal{A})$, partitioned into:

- *positive examples:* $X_+ \triangleq \{a \in X \mid \mathcal{K} \models C(a)\}$;
- *negative examples:* $X_- \triangleq \{a \in X \mid \mathcal{K} \models \neg C(a)\}$;
- *neutral (unlabeled) examples:*

$$X_0 \triangleq \{a \in X \mid a \notin X_+ \wedge a \notin X_-\}.$$

Find: a labeling function $\mathbf{f}^* : X \mapsto \{-1, +1\}$, where $+1$ (resp. -1) corresponds to the positive (resp. negative) class.

III. A GAUSSIAN PROCESS REGRESSION MODEL FOR KNOWLEDGE PROPAGATION

In this section, we discuss a new method, named *Gaussian Process Knowledge Propagation* (GPKP), for solving the prediction problem in Def. 2.1 in the context of Web ontologies. In Sect. III-A we show that a similarity graph between examples can be used to define a finite-set Gaussian Process over their labels, which allows to efficiently propagate class information. In Sect. III-B we propose a solution to the problem of learning the similarity graph. In Sect. III-C we discuss how to retrieve relations among examples expressed as conjunctive queries.

A. Transductive Learning as an Optimization Problem

We now propose a solution to the transductive learning problem in Def. 2.1. Following Sect. II, we aim at finding a labeling function \mathbf{f}^* defined over examples X , which is both *consistent* with the training labels, and *varies smoothly* among similar individuals; we assume that a similarity graph over examples in X has already been provided. Such a graph is represented by its symmetric adjacency matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$, with $n \triangleq |X|$, such that $\mathbf{W}_{ij} \geq 0$ if $x_i, x_j \in X$ are *similar*, and 0 otherwise; for simplicity, we assume that $\mathbf{W}_{ii} = 0$.

Formally, each labeling function can be represented as a finite-size vector $\mathbf{f} \in \{-1, +1\}^n$, where \mathbf{f}_i is the label for the i -th element in the set of examples X . According to [8], labels can be enforced to vary smoothly among similar individuals by means of the following penalty function defined over \mathbf{f} :

$$E(\mathbf{f}) \triangleq \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{W}_{ij} (\mathbf{f}_i - \mathbf{f}_j)^2 + \epsilon \sum_{i=1}^n \mathbf{f}_i^2, \quad (1)$$

where the first term enforces the labeling function to vary smoothly among similar examples, and the second term is a ℓ_2 regularizer (with weight $\epsilon > 0$) over \mathbf{f} . Let $L \triangleq X_+ \cup X_-$ and $U \triangleq X_0$ represent labeled and unlabeled examples, respectively. In [8], authors propose a continuous relaxation of \mathbf{f} , where $\mathbf{f} \in [-1, +1]^n$ also encodes a measure of the classification confidence. This allows for a simple, closed-form solution to the problem of minimizing $E(\cdot)$ for a given value of \mathbf{f}_L , providing a solution to the problem in Def. 2.1. Note that the penalty function $E(\cdot)$ in Eq. 1 can be rewritten as:

$$E(\mathbf{f}) = \mathbf{f}^T (\mathbf{L} + \epsilon \mathbf{I}) \mathbf{f}, \quad (2)$$

where \mathbf{D} is a diagonal matrix such that $\mathbf{D}_{ii} = \sum_{j=1}^{|X|} \mathbf{W}_{ij}$ and $\mathbf{L} \triangleq \mathbf{D} - \mathbf{W}$ is the *graph Laplacian* of \mathbf{W} . Reordering the matrix \mathbf{W} , the graph Laplacian \mathbf{L} and the vector \mathbf{f} w.r.t. the membership to L and U , they can be rewritten as:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{LL} & \mathbf{W}_{LU} \\ \mathbf{W}_{UL} & \mathbf{W}_{UU} \end{bmatrix}, \mathbf{L} = \begin{bmatrix} \mathbf{L}_{LL} & \mathbf{L}_{LU} \\ \mathbf{L}_{UL} & \mathbf{L}_{UU} \end{bmatrix}, \mathbf{f} = \begin{bmatrix} \mathbf{f}_L \\ \mathbf{f}_U \end{bmatrix}. \quad (3)$$

The problem of finding a labeling \mathbf{f}_U^* for unlabeled examples which minimizes the penalty function $E(\cdot)$ for a given value for \mathbf{f}_L has the following closed form solution:

$$\mathbf{f}_U^* = (\mathbf{L}_{UU} + \epsilon \mathbf{I})^{-1} \mathbf{W}_{UL} \mathbf{f}_L. \quad (4)$$

Efficiency: A solution for Eq. 4 can be computed efficiently in nearly-linear time. Indeed computing \mathbf{f}_U^* can be reduced to solving a linear system in the form $\mathbf{A} \mathbf{x} = \mathbf{b}$, with $\mathbf{A} = (\mathbf{L}_{UU} + \epsilon \mathbf{I})$, $\mathbf{b} = \mathbf{W}_{UL} \mathbf{f}_L$ and $\mathbf{x} = \mathbf{f}_U^*$. A linear system $\mathbf{A} \mathbf{x} = \mathbf{b}$ with $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be solved in nearly linear time if the coefficient matrix \mathbf{A} is *symmetric diagonally dominant*² (SDD). An algorithm for solving a SDD linear system is discussed in [9]: its time-complexity is $\approx O(m \log^{\frac{1}{2}} n)$, where m is the number of non-zero entries in \mathbf{A} . In Eq. 4, the matrix $(\mathbf{L}_{UU} + \epsilon \mathbf{I})$ is SDD, since \mathbf{L}_{UU} is a principal submatrix of the graph Laplacian \mathbf{L} which is also SDD [10].

B. A Relations-based Similarity Graph

The method for propagating knowledge across similar examples discussed in Sect. III-A relies on a similarity graph, represented by its adjacency matrix \mathbf{W} . The underlying assumption in this work is that some relations among individuals in the KB might encode a similarity relation w.r.t. a specific target property or class: identifying such relations allows to propagate information across similar individuals. In literature, this phenomenon is referred to as *homophily*: related entities tend to influence each other, and some relations (e.g. *friendship* in a social network) can encode some form of similarity w.r.t. a set of properties (such as hobbies or musical tastes). However, depending on the learning task, not all relations are equally effective at encoding similarities between examples (e.g. quiet people tend to prefer talkative friends and vice versa).

In this work, we represent each relation type by means of a symmetric *adjacency matrix* $\tilde{\mathbf{W}}$, such that $\mathbf{W}_{ij} = \tilde{\mathbf{W}}_{ji} = 1$ iff either the relation $\text{rel}(x_i, x_j)$ or $\text{rel}(x_j, x_i)$ hold in the ontology, and 0 otherwise; rel represents a type of relation between examples (such as friendship or co-authorship). For simplicity, we assume that $\tilde{\mathbf{W}}_{ii} = 0, \forall i$. Given a set of matrices

²A matrix \mathbf{A} is SDD iff \mathbf{A} is symmetric and $\forall i : \mathbf{A}_{ii} \geq \sum_{i \neq j} |\mathbf{A}_{ij}|$.

$\mathcal{W} \triangleq \{\tilde{\mathbf{W}}_1, \dots, \tilde{\mathbf{W}}_r\}$, one for each relation type, we can define \mathbf{W} as a linear combination of matrices in \mathcal{W} :

$$\mathbf{W} \triangleq \sum_{i=1}^r \mu_i \tilde{\mathbf{W}}_i, \text{ with } \mu_i \geq 0, \forall i \quad (5)$$

where μ_i is a parameter representing the weight of $\tilde{\mathbf{W}}_i$ in the adjacency matrix of the similarity graph \mathbf{W} .

Learning as Inverse Covariance Estimation: Let us consider the continuous relaxation of the penalty function $E(\cdot)$ in Eq. 2 (with $\mathbf{f} \in \mathbb{R}^n$). It corresponds to the *energy function* of the following probability density p over \mathbf{f} :

$$p(\mathbf{f}) = (2\pi)^{-\frac{1}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} E(\mathbf{f}) \right\} \\ = \mathcal{N}(\mathbf{0}, (\mathbf{L} + \epsilon \mathbf{I})^{-1}). \quad (6)$$

The probability density function in Eq. 6 defines a finite-set Gaussian process [11] $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Omega} = (\mathbf{L} + \epsilon \mathbf{I})$ and $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$ are respectively its *inverse covariance* (or *precision*) and *covariance* matrix, and $|\boldsymbol{\Sigma}|$ indicates the determinant of $\boldsymbol{\Sigma}$. The covariance matrix and its inverse fully determine the independence relations among variables in a multivariate Gaussian distribution [11]: if $\boldsymbol{\Omega}_{ij} \neq 0$, then there is an edge between variables \mathbf{f}_i and \mathbf{f}_j in the minimal I-map Gaussian Markov Random Field (GMRF) of p . The parametric form of \mathbf{W} is fully specified by the hyperparameters $\boldsymbol{\mu}$ in Eq. 5, which may be unknown. Estimating the inverse precision matrix in a GMRF given a set of observations is referred to as *inverse covariance estimation* [12]: in this work, we estimate the hyperparameters by maximizing a regularized *marginal likelihood* [11] of \mathbf{f}_L induced by the probability density p in Eq. 6. Following [12], we add a ℓ_1 -norm regularization term for controlling the sparsity of hyperparameters (and dealing with the curse of dimensionality). To comply with noisy training labels, we also assume independent and identically distributed Gaussian noise [11] with variance $\sigma^2 \geq 0$. The resulting set of hyperparameters $\boldsymbol{\Theta} \triangleq \{\boldsymbol{\mu}, \epsilon, \sigma^2\}$ fully determines a probability density p , and we estimate $\boldsymbol{\Theta}$ by maximizing the following ℓ_1 -regularized log-marginal likelihood of \mathbf{f}_L :

$$\mathcal{L}(\boldsymbol{\Theta} | \mathbf{f}_L) = -\frac{1}{2} \mathbf{f}_L^T \mathbf{K}_L^{-1} \mathbf{f}_L - \frac{1}{2} \log |\mathbf{K}_L| - \frac{1}{2} \log 2\pi - \lambda \|\boldsymbol{\Theta}\|_1,$$

where $\boldsymbol{\Sigma} = (\mathbf{L} + \epsilon \mathbf{I})^{-1}$ is the covariance matrix ($\boldsymbol{\Sigma}$ has a block structure analogue to the one in Eq. 3), $\mathbf{K}_L \triangleq (\boldsymbol{\Sigma}_{LL} + \sigma^2 \mathbf{I})$, $|\mathbf{K}_L|$ indicates the determinant of \mathbf{K}_L , and $\lambda \geq 0$ controls the sparsity (i.e. the complexity) of the solution. We propose using gradient-based optimization methods for finding the parameters $\boldsymbol{\Theta}_{ML}$ maximizing the marginal log-likelihood $\mathcal{L}(\boldsymbol{\Theta} | \mathbf{f}_L)$. The subgradient of $\mathcal{L}(\boldsymbol{\Theta} | \mathbf{f}_L)$ w.r.t. a hyperparameter $\theta \in \boldsymbol{\Theta}$ can be calculated as follows [11]:

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta} | \mathbf{f}_L)}{\partial \theta} = \frac{1}{2} \text{Tr} \left[(\boldsymbol{\alpha} \boldsymbol{\alpha}^T - \mathbf{K}_L^{-1}) \frac{\partial \mathbf{K}_L}{\partial \theta} \right] - \lambda \frac{\theta}{\sqrt{\theta^2}},$$

where $\boldsymbol{\alpha} = \mathbf{K}_L^{-1} \mathbf{f}_L$ and we assume that $0/0 = 0$. The partial derivatives of \mathbf{K}_L w.r.t. hyperparameters in $\boldsymbol{\Theta}$, according to the parametrization of \mathbf{W} proposed in Eq. 5, are:

$$\frac{\partial \mathbf{K}_L}{\partial \mu_i} = \left[\frac{\partial \boldsymbol{\Sigma}}{\partial \mu_i} \right]_{LL} = - \left[\boldsymbol{\Sigma} \tilde{\mathbf{L}}_i \boldsymbol{\Sigma} \right]_{LL}, \\ \frac{\partial \mathbf{K}_L}{\partial \epsilon} = - \left[\boldsymbol{\Sigma} \mathbf{I} \boldsymbol{\Sigma} \right]_{LL} \text{ and } \frac{\partial \mathbf{K}_L}{\partial \sigma^2} = \mathbf{I},$$

using the property $\partial(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}(\partial \mathbf{X})\mathbf{X}^{-1}$, where $\tilde{\mathbf{L}}_i$ is the Laplacian of the graph corresponding to the matrix $\tilde{\mathbf{W}}_i$.

C. Retrieving Meaningful Relations Between Examples

In this work, we consider the relations encoded by CQs for the construction of the similarity graph. However, the number of relations that can be expressed using CQs is very large: to overcome this problem, in empirical evaluations (see Sect. IV), we considered two types of such relations holding between pairs of examples $a, b \in X$:

- *Simple* relations, i.e. those corresponding to CQs in the form $\mathbf{r}(a, b)$, where $\mathbf{r} \in N_R$ is an atomic role;
- *Composite* relations, corresponding to CQs in the form:

$$\exists z. (\mathbf{r}(a, z) \wedge \mathbf{r}(b, z)) \quad \text{or} \quad \exists z. (\mathbf{r}(z, a) \wedge \mathbf{r}(z, b)),$$

where $\mathbf{r} \in N_R$ is an atomic role and $z \in N_V$ is a variable.

Several efficient query answering services can be used for retrieving complex relations holding among examples. In particular, CQs (see Sect. II) can be expressed in the SPARQL-DL [13] query language. SPARQL-DL seems particularly convenient for the task: it is a specialization of SPARQL, sharing its syntax and specific for OWL's *Direct Model-Theoretic Semantics*³. SPARQL-DL queries generalize CQs, as they admit variables in place of property names, thus answering multiple CQs at once, and *non-distinguished variables*, i.e. those that are bound to entities that need not be interpreted as specific individuals of the queried ontology.

Using SPARQL-DL queries for retrieving complex relations among examples is particularly convenient: a single SPARQL-DL query can answer a set of CQs, thanks to the use of variables in place of role names. When describing the outcome of empirical evaluations, we will use the following short-hand notations to concisely describe the relations retrieved during the learning process:

$$\text{rel}_1 \circ \text{rel}_2^{-1}(a, b) \equiv \exists z. (\text{rel}_1(a, z) \wedge \text{rel}_2(z, b)), \\ \text{rel}_1^{-1} \circ \text{rel}_2(a, b) \equiv \exists z. (\text{rel}_1(z, a) \wedge \text{rel}_2(b, z)),$$

where $\text{rel}_1, \text{rel}_2 \in N_R$, $a, b \in N_I$ and $z \in N_V$.

D. Summary of the Method

The proposed method, named *Gaussian Process Knowledge Propagation* (GPKP), can be summarized as follows:

- 1) Retrieve (possibly complex) relations among examples in X using SPARQL-DL queries, and then use them to create a set of adjacency matrices $\mathcal{W} = \{\tilde{\mathbf{W}}_1, \dots, \tilde{\mathbf{W}}_r\}$.
- 2) Learn the hyperparameters by maximizing the regularized marginal log-likelihood of labeled examples in Eq. III-B:

$$\boldsymbol{\Theta}_{ML} = \arg \max_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\Theta} | \mathbf{f}_L). \quad (7)$$

- 3) Use the learned parameters $\boldsymbol{\Theta}_{ML} = \{\boldsymbol{\mu}, \epsilon, \sigma^2\}$ to find the most likely labels for unlabeled examples \mathbf{f}_U :

$$\mathbf{f}_U^* = \mathbb{E}[\mathbf{f}_U | \mathbf{f}_L, \mathcal{W}, \boldsymbol{\Theta}_{ML}] = \boldsymbol{\Sigma}_{UL} \boldsymbol{\Sigma}_{LL}^{-1} \mathbf{f}_L,$$

where $\boldsymbol{\Sigma} = [(\mathbf{L} + \epsilon \mathbf{I})^{-1} + \sigma^2 \mathbf{I}]$, and \mathbf{L} is the graph Laplacian of the learned similarity graph with adjacency matrix $\mathbf{W} = \sum_{i=1}^r \mu_i \tilde{\mathbf{W}}_i$.

³<http://www.w3.org/TR/owl2-direct-semantics>

TABLE I: Ontologies considered in the experiments

Ontology	DL Lang.	#Axioms	#Inds.	#Props.
AIFB PORTAL [3]	$\mathcal{AL}\mathcal{E}\mathcal{H}\mathcal{O}(\mathcal{D})$	268540	44328	285
DBPEDIA 3.9 [14] Frag.	$\mathcal{AL}\mathcal{CH}$	78795	16606	132
BGS [4]	$\mathcal{AL}\mathcal{I}(\mathcal{D})$	825133	87555	154

IV. EMPIRICAL EVALUATION

The method discussed in Sect. III was experimentally evaluated in comparison with other approaches proposed in the literature on a variety of assertion prediction problems. Sources and datasets are available at <https://code.google.com/p/gpkp/>.

Ontologies: We considered three real, shallow ontologies: the AIFB PORTAL Ontology ⁴, the DBPEDIA 3.9 Ontology [14] and the BRITISH GEOLOGICAL SURVEY (BGS) Ontology ⁵. AIFB PORTAL models the key concepts (and relations) within a research community, BGS represents knowledge collected by the British Geological Survey, and DBPEDIA contains structured information extracted from Wikipedia. The characteristics of these ontologies are outlined in Tab. I.

Experimental Setting: GPKP is summarized in Sect. III-D. We used *Projected Gradient Ascent* for solving the optimization problem in Eq. 7, jointly with an intermediate line search to assess the step size. In each learning task, the L_1 regularization parameter λ was selected via cross validation within the training set (with λ ranging in $\lambda \in \{0.0, 10^{-2}, 10^{-1}, \dots, 10^2\}$). Before each experiment, all knowledge inherent to the target class was removed from the ontology. Following the related evaluation procedures in [3], [4], members of the target concepts were considered as *positive examples*, while an equal number of *negative examples* was randomly sampled from unlabeled examples. Remaining instances (i.e. neither positive nor negative) were considered as *neutral* (or unlabeled) *examples*.

Results are reported in terms of *Area Under the Precision-Recall Curve* (AUC-PR), a measure to evaluate rankings also used in e.g. [6]. In each experiment, we considered the problem of predicting the membership of examples to several classes; for each of such classes, we performed a 10-fold cross validation (CV), and report the average AUC-PR obtained using each of the considered methods. We used the same 10-folds partitioning across experiments related to each of the datasets; for such a reason, we report statistical significance tests using a paired, non-parametric difference test (Wilcoxon T test). We also report diagrams showing how using a limited quantity of randomly sampled labeled training instances (i.e. 10%, 30%, 50%, \dots , a plausible scenario for real world settings with limited labeled training data), and using the remaining examples for testing, affects the resulting AUC-PR values.

We compare GPKP with state-of-the-art approaches proposed for learning from ontological KBs. We considered two kernel methods: Soft-Margin SVM (SM-SVM) and Kernel Logistic Regression (KLR), jointly with different kernel functions suited for ontological KBs. We also considered two relational prediction models, namely SUNS [5] and RESCAL [6]. The RDF graph used by both kernels methods and relational

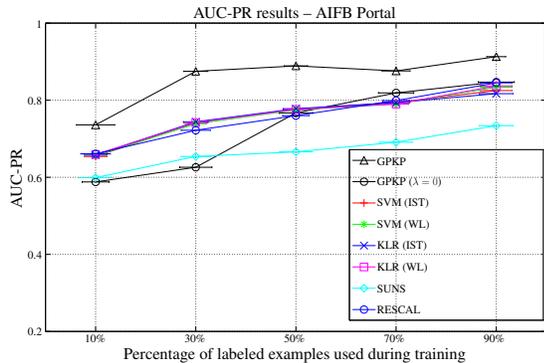
prediction models was materialized as follows: all $\langle s, p, o \rangle$ triples were retrieved by means of SPARQL-DL queries (where p was either an object or a data-type property) together with all *direct type* and *direct sub-class* relations. In our experiments, we used the *Intersection SubTree* [3] (IST) and the *Weisfeiler-Lehman* [4] (WL) kernels for ontological KBs. For each kernel/algorithm and each learning task, parameters have been selected via 10-fold CV. IST kernel parameters were ranging in $d \in \{1, 2, 3, 4\}$ and $\lambda_{ist} \in \{0.1, 0.3, \dots, 0.9\}$, and WL kernel parameters in $d, h \in \{1, 2, 3, 4\}$ (where d is the depth of the neighborhood graph). In SM-SVM, in order to obtain a ranking among instances (provided by continuous labels f in GPKP), we applied the logistic function s to the decision boundary f instead of the sign function, which is commonly used in the classification context (thus obtaining $s(f(\cdot)) : \mathcal{X} \rightarrow [0, 1]$). In SM-SVM, $C \in \{0.0, 10^{-6}, 10^{-4}, \dots, 10^4, 10^6\}$, while in KLR the weight λ_k associated to the L_2 regularizer was found considering $\lambda_k \in \{10^{-4}, 10^{-3}, \dots, 10^4\}$. The SUNS relational prediction model relies on a low-rank approximation of the matrix representing the relational multigraph. Parameters were selected by means of a 10-fold CV within the training set by grid optimization, with $t \in \{2, 4, 6, \dots, 24\}$ and $\lambda_s \in \{0, 10^{-2}, 10^{-1}, \dots, 10^6\}$. In RESCAL, each evaluation for the regularization parameters t and λ_r requires a tensor factorization step, thus model selection may be unfeasible for large domains. Also, factorized tensor representations are dense, thus the proposed approach might become too memory demanding for large values of t . Each experiment with RESCAL used the ALS algorithm [6].

AIFB PORTAL Ontology: Similarly to other experiments conducted on this ontology (such as [3] and [4]), the learning task consisted in predicting the affiliations of AIFB staff members to research groups. Specifically, in a set of 316 examples (each representing a researcher in the ontology), the task consisted in predicting missing affiliations to 5 distinct research groups. Due to the computational cost of RESCAL, the number of iterations for the ALS algorithm was fixed to 8 and the graph was composed only by statistical units and their immediate neighborhoods; parameter selection was performed via 10-fold CV using the training set, with $t \in \{12, 16, \dots, 32\}$ and $\lambda_r \in \{10^{-8}, 10^{-4}, 1\}$.

Fig. 1 summarizes the AUC-PR results on the research group affiliation prediction task, obtained via 10-fold CV (one per research group, in a *one-versus-all* setting). The plot shows average AUC-PR values describes results obtained with an increasing number of training examples, and leaving the rest to the test: error bars (pictured horizontally) represent twice the standard deviation. In general, results slightly varied among research groups. The proposed method (GPKP) yields significantly higher AUC-PR values than those observed with other methods, where statistical significance was assessed by a Wilcoxon T test with $p < 0.05$. We also compared GPKP with its variant which does not make use of a sparsity-enforcing L_1 regularizer, denoted GPKP ($\lambda = 0$). Results provided by GPKP were significantly higher than those observed with GPKP ($\lambda = 0$), showing how enforcing sparsity in the parameters vector μ can be beneficial to the learning task. GPKP can also be used to elicit new knowledge on a domain: Tab. II shows a sample of the relations considered for the affiliation prediction task, among a total of 77 retrieved (all *composite*) relations, together with a measure of their relevance (given by their

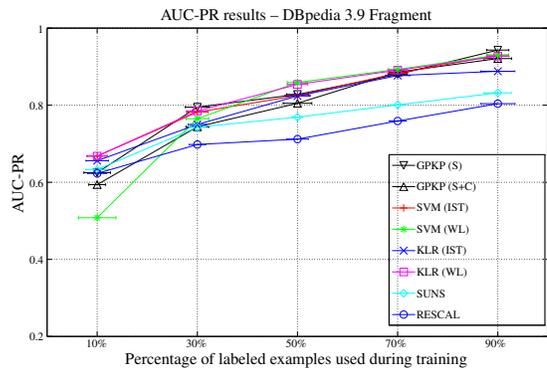
⁴<http://www.aifb.kit.edu/web/Wissensmanagement/Portal>

⁵<http://data.bgs.ac.uk/>, as of March 2014



Method	AUC-PR (mean \pm var.)	
GPKP	0.913 \pm 0.009	
GPKP ($\lambda = 0$)	0.847 \pm 0.032	▼
SUNS	0.734 \pm 0.030	▼
RESCAL	0.845 \pm 0.025	▼
SM-SVM (IST)	0.825 \pm 0.025	▼
SM-SVM (WL)	0.834 \pm 0.025	▼
KLR (IST)	0.817 \pm 0.029	▼
KLR (WL)	0.837 \pm 0.025	▼

Fig. 1: AIFB PORTAL Ontology – Plot: AUC-PR results (mean, std.dev.) estimated by 10-fold CV, obtained varying the percentage of labeled examples used for training – Table: AUC-PR results estimated by 10-fold CV: ▼/▼ (resp. ▲/▲) indicates that GPKP’s mean is significantly higher (resp. lower) in a paired Wilcoxon T test with $p < 0.05 / p < 0.10$



Method	AUC-PR (mean \pm var.)	S	S+C
GPKP ^S	0.943 \pm 0.012		
GPKP ^{S+C}	0.921 \pm 0.019		
SUNS	0.832 \pm 0.019	▼	▼
RESCAL	0.804 \pm 0.029	▼	▼
SM-SVM (IST)	0.930 \pm 0.011		
SM-SVM (WL)	0.930 \pm 0.011		
KLR (IST)	0.888 \pm 0.029		
KLR (WL)	0.927 \pm 0.012		

Fig. 2: DBPEDIA 3.9 Ontology – Plot: AUC-PR results (mean, std.dev.) estimated by 10-fold CV, obtained varying the percentage of labeled examples used for training – Table: AUC-PR results estimated by 10-fold CV and the corresponding paired Wilcoxon T significance tests (described as in Fig. 1)

TABLE II: Relations considered in the AIFB PORTAL and the DBPEDIA 3.9 Ontologies and their corresponding weight

AIFB PORTAL	
High μ_i	Low μ_i
publications ⁻¹ \circ publications	title \circ title ⁻¹
interest \circ interest ⁻¹	mobile \circ mobile ⁻¹
lecturer ⁻¹ \circ lecturer	road \circ road ⁻¹

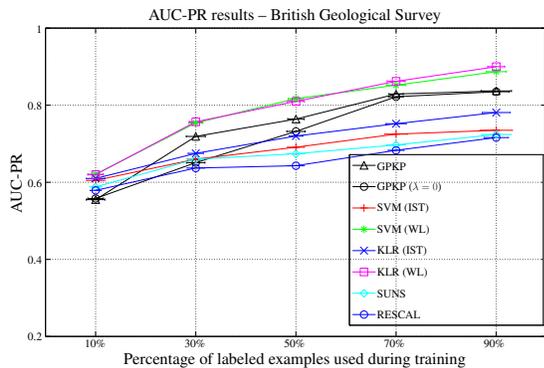
DBPEDIA 3.9	
High μ_i	Low μ_i
vicePresident	successor
president	profession \circ profession ⁻¹
region \circ region ⁻¹	religion \circ religion ⁻¹

associated weight μ_i , described as either Low if $\mu_i \approx 0$, and High otherwise). GPKP successfully recognizes that authors sharing publications or interests, teaching the same courses or sharing the office space are likely to be affiliated to the same research group (unlike e.g. sharing the same academic title).

Evaluation on the DBPEDIA 3.9 Fragment: Similarly to [6], we evaluated the proposed approach on two prediction tasks, namely predicting party affiliations to either the Democratic and the Republican party for 82 US presidents and vice-presidents from DBPEDIA 3.9. The experiment illustrated in [6] uses a small RDF fragment containing the `president` and `vicePresident` predicates only. In this experiment, we used a fragment of DBPEDIA 3.9, obtained by means of a crawling process. Following the extraction procedure in [15], the DBPEDIA 3.9 RDF graph was traversed starting from resources representing US Presidents and Vice-Presidents: all immediate neighbors, i.e. those with a recursion depth of 1, were retrieved, together with their related schema information

(direct classes and their super-classes, together with their hierarchy). All extracted knowledge was used to create a KB whose characteristics are outlined in Tab. I. In RESCAL, the number of iterations of the ALS algorithm was fixed to 16, with parameters $t = 32$ and $\lambda_r = 10^{-8}$ (given by an analysis of the dataset), while for WL $d = 1$ and $h = 1$. In this experiment, the total number of retrieved relations (both *simple* and *composite*) was higher than the number of instances itself: 82 US presidents and vice-presidents were interlinked by 25 simple relations and 149 composite relations. This differs from other empirical evaluations discussed in this paper, in which instances are linked by a more limited number of, exclusively composite, relations. For such a reason, we evaluated two variants of the proposed method: GPKP^S, which only uses simple relations, and GPKP^{S+C}, which uses both simple and composite relations among examples.

Experimental results are summarized in Fig. 2. We observe that AUC-PR values obtained with GPKP^S are higher than results obtained by other methods considered in comparison. However the difference was not always statistically significant: only in two cases we observed that $p < 0.05$. AUC-PR values were slightly lower in the case of GPKP^{S+C}, which might be explained by the *curse of dimensionality*. GPKP was able to identify which relations are likely to link same party affiliates, some of which are summarized in Tab. II. It successfully identified that the vice president is likely to belong to the same party of the president; that representatives covering a role under the same president are likely to belong to the same party; or that representatives elected in the same region are likely to belong to the same party. On the other hand, sharing the same religion, profession, nationality or awards does not necessarily mean sharing the same party affiliation.



Method	AUC-PR (mean ± var.)	
GPKP	0.837 ± 0.023	
GPKP (λ = 0)	0.835 ± 0.016	
SUNS	0.724 ± 0.022	▼
RESCAL	0.716 ± 0.015	▼
SM-SVM (IST)	0.735 ± 0.026	▼
SM-SVM (WL)	0.887 ± 0.010	
KLR (IST)	0.781 ± 0.020	
KLR (WL)	0.900 ± 0.007	△

Fig. 3: BGS Ontology – Plot: AUC-PR results (mean, std.dev.) estimated by 10-fold CV, obtained varying the percentage of examples used for training – Table: AUC-PR results estimated by 10-fold CV and the corresponding paired Wilcoxon T significance tests (described as in Fig. 1)

Evaluation on the BRITISH GEOLOGICAL SURVEY Ontology: As in [4], we focused on the *Lithogenesis* prediction problem in the BRITISH GEOLOGICAL SURVEY (BGS) Ontology. The problem consisted in predicting the value of the property *hasLithogenesis* in a set of 159 named rock units labeled with their corresponding lithogenetic type. As in [4], we focus on two tasks, consisting in the prediction of two major lithogenetic types: “Alluvial” and “Glacial”. For efficiency reasons, in SUNS and RESCAL the relational graph was composed only by statistical units and their immediate neighborhoods. In RESCAL, the number of iterations for the ALS algorithm was fixed to 16, while parameter selection was performed via 5-fold CV within the training set, with $t \in \{12, 16, \dots, 32\}$ and $\lambda_r \in \{10^{-8}, 10^{-4}, 1\}$.

Results are summarized in Fig. 3: AUC-PR values observed with GPKP are higher than those observed with kernel methods using the IST kernel, SUNS or RESCAL. Kernel methods relying on the WL kernel provided slightly higher AUC-PR results than GPKP ($p < 0.10$ for KLR), confirming the effectiveness of the WL kernel on this specific dataset [4]. However, it is not clear how to interpret statistical models induced by using the WL kernel in terms of domain knowledge. On the other hand, models learned by GPKP explicitly indicate the importance of each relation in the knowledge propagation process. Also in this case, GPKP was able to extract relations between rock units that encode a form of similarity w.r.t. their lithogenetic type. For example, among a total of 23 relations (all *composite*) it emerged that rocks with similar geographical distributions, thickness and lithological components were likely to share their lithogenetic type, while their geological theme and oldest geological age were not considered informative.

V. SUMMARY

Starting from the assumption that some relations among examples in a Web ontology might influence each other, we propose a method, named *Gaussian Process Knowledge Propagation* (GPKP), for efficiently learning the importance of each relation in a knowledge propagation process. In the proposed method, the joint distribution over labels for a set of examples is modeled through a finite-set Gaussian Process regression model, where the inverse covariance matrix is learned by exploiting relations holding between examples in the ontology. We experimentally show that GPKP is competitive with state-of-the-art methods in terms of AUC-PR. It also provides an interpretable statistical model, proving to be an effective instrument for mining new knowledge from Web ontologies.

Acknowledgments: This work fulfills the objectives of the PON 02_00563_3489339 project “PUGLIA@SERVICE - Internet-based Service Engineering enabling Smart Territory structural development” funded by the Italian Ministry of University and Research (MIUR).

REFERENCES

- [1] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Scientific American*, vol. 284, no. 5, pp. 34–43, May 2001.
- [2] A. Rettinger, U. Lössch, V. Tresp, C. d’Amato, and N. Fanizzi, “Mining the Semantic Web: Statistical learning for next generation knowledge bases,” *Data Min. Knowl. Discov.*, vol. 24, no. 3, pp. 613–662, 2012.
- [3] U. Lössch, S. Bloehdorn, and A. Rettinger, “Graph kernels for RDF data,” in *Proceedings of ESWC’12*, ser. LNCS, E. Simperl *et al.*, Eds., vol. 7295. Springer, 2012, pp. 134–148.
- [4] G. K. D. de Vries, “A Fast Approximation of the Weisfeiler-Lehman Graph Kernel for RDF Data,” in *ECML/PKDD (1)*, ser. LNCS, H. Blockeel *et al.*, Eds., vol. 8188. Springer, 2013, pp. 606–621.
- [5] V. Tresp, Y. Huang, M. Budschus, and A. Rettinger, “Materializing and querying learned knowledge,” in *Proceedings of IRML’09*, 2009.
- [6] M. Nickel, V. Tresp, and H.-P. Kriegel, “A Three-Way Model for Collective Learning on Multi-Relational Data,” in *Proceedings of ICML’11*, L. Getoor *et al.*, Eds. Omnipress, 2011, pp. 809–816.
- [7] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds., *The Description Logic Handbook*. Cambridge University Press, 2007.
- [8] X. Zhu, Z. Ghahramani, and J. D. Lafferty, “Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions,” in *Proceedings of ICML’03*, T. Fawcett *et al.*, Eds. AAAI Press, 2003, pp. 912–919.
- [9] M. B. Cohen, R. Kyng, G. L. Miller, J. W. Pachocki, R. Peng, A. Rao, and S. C. Xu, “Solving SDD linear systems in nearly $m \log^{1/2} n$ time,” in *Proceedings of STOC 2014*, D. B. Shmoys, Ed. ACM, 2014, pp. 343–352.
- [10] D. A. Spielman, “Algorithms, Graph Theory, and Linear Equations in Laplacian Matrices,” in *Proceedings of ICM’10*, 2010, pp. 2698–2722.
- [11] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. MIT Press, 2005.
- [12] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
- [13] E. Sirin and B. Parsia, “SPARQL-DL: SPARQL Query for OWL-DL,” in *OWLED*, ser. CEUR Workshop Proceedings, C. Golbreich *et al.*, Eds., vol. 258. CEUR-WS.org, 2007.
- [14] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, “DBpedia - a crystallization point for the Web of Data,” *J. Web Sem.*, vol. 7, no. 3, pp. 154–165, 2009.
- [15] S. Hellmann, J. Lehmann, and S. Auer, “Learning of OWL Class Descriptions on Very Large Knowledge Bases,” *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 2, pp. 25–48, 2009.