An Experimental Evaluation of a Scalable Probabilistic Description Logic Approach for Semantic Link Prediction

José Eduardo Ochoa Luna[1], Kate Revoredo[2], and Fabio Gagliardi Cozman[1]

[1] Escola Politécnica, Universidade de São Paulo,
Av. Prof. Mello Morais 2231, São Paulo - SP, Brazil
[2] Departamento de Informática Aplicada, Unirio
Av. Pasteur, 458, Rio de Janeiro, RJ, Brazil
eduardo.ol@gmail.com,katerevoredo@uniriotec.br,fgcozman@usp.br

Abstract. In previous work, we presented an approach for link prediction using a probabilistic description logic, named CR$\mathcal{ALC}$. Inference in CR$\mathcal{ALC}$, considering all the social network individuals, was used for suggesting or not a link. Despite the preliminary experiments have shown the potential of the approach, it seems unsuitable for real world scenarios, since in the presence of a social network with many individuals and evidences about them, the inference was unfeasible. Therefore, we extended our approach through the consideration of graph-based features to reduce the space of individuals used in inference. In this paper, we evaluate empirically this modification comparing it with standard proposals. It was possible to verify that this strategy does not decrease the quality of the results and makes the approach scalable.

1 Introduction

Many social, biological, and information systems can be well described by networks, where nodes represent objects (individuals), and links denote the relations or interactions between nodes. Predicting a possible link in a network is an interesting issue that has gained attention, due to the growing interest in social networks. For instance, one may be interested on finding potential friendship between two persons in a social network, or a potential collaboration between two researchers. Thus link prediction [13] aims at predicting whether two nodes (i.e. people) should be connected given that we know previous information about their relationships or interests.

In [13] a survey with some representative link prediction methods, categorized in three groups, was presented. In the first group, feature-based methods construct pair-wise features to use in a classification task. The majority of the features are extracted from the graph topology computing the similarity based on the neighborhood of the pair of nodes or based on ensembles of paths between the pair of nodes [10]. Recently, semantic informations have also being used as features [21, 17]. The second group includes probabilistic approaches which

model the joint-probability among the entities in a network by Bayesian graphical models [20]. And, finally the third group concerns linear algebraic approaches which computes the similarity between the nodes in a network by rank-reduced similarity matrices [9].

In [15] we presented an approach that uses a Bayesian graphical model together with semantic-based features for semantic link prediction. Therefore, our proposal lies on the first two categories described previously. To model the domain and thus consider semantic-based features, the proposal adopted a probabilistic description logic called Credal $\mathcal{ALC}$ (CR$\mathcal{ALC}$) [5], that extends the popular logic $\mathcal{ALC}$ [3] with *probabilistic inclusions*. These are sentences, such as $P(\mathsf{Professor}|\mathsf{Researcher}) = 0.4$, indicating the probability that an element of the domain is a Professor given that it is a Researcher. Exact and approximate inference algorithms have been proposed [5], using ideas inherited from the theory of Relational Bayesian Networks (RBN) [8]. In [14], we extended our proposal to also consider graph-based approaches in order to scale for large social network. In this paper we conduct some experimental analisis in order to verify the benefits of our proposal.

The paper is organized as follows. Section 2.1 reviews basic concepts of probabilistic description logics, CR$\mathcal{ALC}$ and our proposal for a scalable semantic link prediction approach. Section 3 describes the experiments we conducted bringing some discussions. Section 4 concludes the paper.

2 Background

In this section, probabilistic description logic CR$\mathcal{ALC}$ and our former proposal for semantic link prediction are reviewed.

2.1 Probabilistic Description Logics and CR$\mathcal{ALC}$

Description logics (DLs) form a family of representation languages that are typically decidable fragments of first order logic (FOL) [3]. Knowledge is expressed in terms of *individuals*, *concepts*, and *roles*. The semantics of a description is given by a *domain* $\mathcal{D}$ (a set) and an *interpretation* $\cdot^{\mathcal{I}}$ (a functor). Individuals represent objects through names from a set $N_{\mathsf{I}} = \{a, b, \ldots\}$. Each *concept* in the set $N_{\mathsf{C}} = \{C, D, \ldots\}$ is interpreted as a subset of a domain $\mathcal{D}$. Each *role* in the set $N_{\mathsf{R}} = \{r, s, \ldots\}$ is interpreted as a binary relation on the domain.

Several probabilistic descriptions logics have appeared in the literature [11]. Heinsohn [7] and Sebastiani [18] consider probabilistic inclusion axioms such as $P_{\mathcal{D}}(\mathsf{Professor}) = \alpha$, meaning that a randomly selected object is a Professor with probability α. This characterizes a *domain-based* semantics: probabilities are assigned to subsets of the domain $\mathcal{D}$. Sebastiani also allows inclusions such as $P(\mathsf{Professor}(\mathsf{John})) = \alpha$, specifying probabilities over the interpretations themselves. For example, one interprets $P(\mathsf{Professor}(\mathsf{John})) = 0.001$ as assigning 0.001 to be the probability of the set of interpretations where John is a Professor. This characterizes an *interpretation-based* semantic.

The probabilistic description logic CR$\mathcal{ALC}$ is a probabilistic extension of the DL $\mathcal{ALC}$ that adopts an interpretation-based semantics. It keeps all constructors of $\mathcal{ALC}$, but only allows concept names on the left hand side of inclusions/definitions. Additionally, in CR$\mathcal{ALC}$ one can have probabilistic inclusions such as $P(C|D) = \alpha$ or $P(r) = \beta$ for concepts C and D, and for role r. If the interpretation of D is the whole domain, then we simply write $P(C) = \alpha$. The semantics of these inclusions is roughly (a formal definition can be found in [5]) given by:

$$\forall x \in \mathcal{D} \ : \ P(C(x)|D(x)) = \alpha,$$

$$\forall x \in \mathcal{D}, y \in \mathcal{D} \ : \ P(r(x,y)) = \beta.$$

We assume that every terminology is acyclic; no concept uses itself. This assumption allows one to represent any terminology $\mathcal{T}$ through a directed acyclic graph. Such a graph, denoted by $\mathcal{G}(\mathcal{T})$, has each concept name and role name as a node, and if a concept C directly uses concept D, that is if C and D appear respectively in the left and right hand sides of an inclusion/definition, then D is a *parent* of C in $\mathcal{G}(\mathcal{T})$. Each existential restriction $\exists r.C$ and value restriction $\forall r.C$ is added to the graph $\mathcal{G}(\mathcal{T})$ as nodes, with an edge from r and C to each restriction directly using it. Each restriction node is a *deterministic* node in that its value is completely determined by its parents. The graph $\mathcal{G}(\mathcal{T})$ is a Relational Bayesian Network (RBN) [8].

Example 1. Consider a terminology $\mathcal{T}_1$ with concepts $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$. Suppose $P(\mathsf{A}) = 0.9, \mathsf{B} \sqsubseteq \mathsf{A}, \mathsf{C} \sqsubseteq \mathsf{B} \sqcup \exists r.\mathsf{D}, P(\mathsf{B}|\mathsf{A}) = 0.45, P(\mathsf{C}|\mathsf{B} \sqcup \exists r.\mathsf{D}) = 0.5$, and $P(\mathsf{D}|\forall r.\mathsf{A}) = 0.6$. The last three assessments specify beliefs about partial overlap among concepts. Suppose also $P(\mathsf{D}|\neg\forall r.\mathsf{A}) = \epsilon \approx 0$ (conveying the existence of exceptions to the inclusion of D in $\forall r.\mathsf{A}$). Figure 1 in the left depicts $\mathcal{G}(\mathcal{T})$, while the graph in the right illustrates the grounding of $\mathcal{G}(\mathcal{T})$ for a domain with two individuals ($\mathcal{D} = \{a, b\}$).

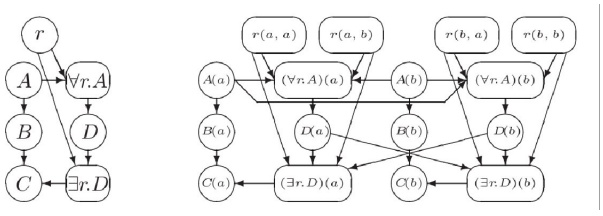

Fig. 1. $\mathcal{G}(\mathcal{T})$ for terminology $\mathcal{T}$ in Example 1 and its grounding for domain$\mathcal{D} = \{a, b\}$.

The semantics of CR$\mathcal{ALC}$ is based on probability measures over the space of interpretations, for a fixed domain. Inferences, such as $P(\mathsf{A}_o(\mathsf{a}_0)|\mathcal{A})$ for an ABox $\mathcal{A}$, can be computed by propositionalization, generating a grounding RBN, where one slice is built for each individual. Therefore, not always exact probabilistic in-

ference is possible. In [5], a first order loopy propagation algorithm was proposed for approximate calculations.

2.2 Link Prediction with CR$\mathcal{ALC}$

Given a social network $\mathcal{N}$, where nodes are entities (represented by letters $a, b, c, \ldots$), one is interested in defining whether a link between a and b is suitable given that there is no link between these nodes in $\mathcal{N}$. In [15], interests, i.e., semantics between the nodes were modeled through a probabilistic ontology represented by the probabilistic description logic CR$\mathcal{ALC}$. In addition, in [14] graph path information was used to improve probabilistic inference. In summary, the semantic link prediction task proposed in [15] (and improved in [14]) can be described as:

Given:
- a network $\mathcal{N}$ defining relationship between objects;
- an ontology $\mathcal{O}$ in CR$\mathcal{ALC}$ describing the domain of the objects;
- the ontology concept $\mathcal{C}$ that defines the semantics of the network objects;
- the ontology role $r(\_, \_)$ that defines the semantics of the relationship between network objects;

Find:
- a revised network $\mathcal{N}_f$ with new relationship between objects.

The proposed algorithm for link prediction receives a network of a specific domain. For instance, in a co-authorship network the nodes represent researchers and the relationship can have the semantics "has a publication with" or "is advised by". Therefore, the ontology represented by CR$\mathcal{ALC}$ describes the domain of publications between researchers, having concepts like Researcher and Publication and roles like hasPublication, hasSameInstitution and sharePublication. This ontology can be learned automatically through a learning algorithm as the ones proposed in [16]. Thus, the nodes represent instances of one of the concepts described in the probabilistic description logic CR$\mathcal{ALC}$ and the semantics of the links is described by one of the roles in the probabilistic description logic CR$\mathcal{ALC}$. These concept and role must be informed as inputs to the proposed algorithm. The link prediction algorithm is described in Algorithm 1.

The algorithm starts looking for all pairs of instances of the concept $\mathcal{C}$ defined as the concept that provides the semantics for the network nodes — this is a general setting, as a rule the set of possible pairs is restricted. For each pair, it checks whether a link between the corresponding nodes exist in the network. If not the probability of the link is calculated through the probability of the defined role conditioned on evidences (step 5). The evidences are provided by the instances of the ontology. The number of instances in an ontology has a great impact in inference. Usually one considers that more instances better inference. However, evidences for different individuals can turn out the inference process computationally expensive, since in a RBN a slice is created for each individual,

Require: a network $\mathcal{N}$, an ontology $\mathcal{O}$, the role $r(\_,\_)$ representing the semantics of the network link, the concept $\mathcal{C}$ describing the objects of the network and a *threshold*.
Ensure: a revised network $\mathcal{N}_f$
1: define $\mathcal{N}_f$ as $\mathcal{N}$;
2: **for all** pair of instances (a, b) of concept $\mathcal{C}$ **do**
3: **if** does not exist a link between nodes a and b in the network $\mathcal{N}$ **then**
4: compute *evidence* based on a, b and nodes in their path;
5: infer probability $P(r(a, b)|evidence)$ using the RBN created through the ontology $\mathcal{O}$;
6: **if** $P(r(a, b)|evidence) > threshold$ **then**
7: add a link between a and b in network $\mathcal{N}_f$;
8: **end if**
9: **end if**
10: **end for**

Algorithm 1: Algorithm for link prediction through CR$\mathcal{ALC}$.

and then inference should be done for each slice. In [5], an approximate inference algorithm was proposed where all slices without evidence are consolidated in a unique slice, thus making inference feasible in real domains. Therefore, less individuals with evidence faster inference is. From another perspective we are interested in predicting a relationship between two individuals, a and b. Therefore, evidences for these two individuals and other individuals strongly related to them are more relevant for link prediction than evidences from other individuals in the network. Thus, in [14] we extended our semantic link prediction approach in order to consider evidences about a, b and the individuals in their path, which makes the link prediction problem scalable for large networks. Therefore, in step 4 the nodes (individuals) belonging to the path between a and b are found. The inference is then performed through CR$\mathcal{ALC}$ lifted variational method on ontology $\mathcal{O}$. If the probability inferred is greater than a threshold then the corresponding link is added to the network. Alternatively, when the threshold to be considered is not known a priori, a rank of the inferred links based on their probability is done and the top-k, where k would be a parameter, are chosen.

3 Experiments

In order to evaluate our previously proposed approach for semantic link prediction empirical experiments were performed. To do so, a real world dataset was used and our algorithm was combined with state-of-the-art measures on a classification model for link prediction. This section reports on steps involved in this process.

3.1 Scenario Description

The Lattes Platform is the public repository of Brazilian scientific curriculum which is comprised by approximately a million of registered researchers. Infor-

mation is given in HTML format, and ranges from personal information such as name and address to a list of publications, examination board participations, research areas, research projects and advising/advisor information. There is implicit relational information in these web pages, for instance collaboration networks are given by advising/adviser links, shared publications and so on. We have randomly selected a set of 1100 researchers from engineering and math backgrounds and based on assertional data about these researchers a probabilistic ontology has been learned. To perform link prediction, this ontology has also been extended with some probabilistic roles — learning is mainly addressed to probabilistic inclusions and concepts. Part of the revised ontology is as follows.

$$
\begin{array}{ll}
 & P(\mathsf{Publication}) = 0.3 \\
 & P(\mathsf{Board}) = 0.33 \\
 & P(\mathsf{sharePublication}) = 0.22 \\
 & P(\mathsf{wasAdvised}) = 0.05 \\
 & P(\mathsf{hasSameInstitution}) = 0.14 \\
 & P(\mathsf{sameExaminationBoard}) = 0.31 \\
\mathsf{ResearcherLattes} \equiv & \mathsf{Person} \\
 & \sqcap(\exists \mathsf{hasPublication.Publication} \\
 & \sqcap \exists \mathsf{advises.Person} \sqcap \exists \mathsf{participate.Board}) \\
P(\mathsf{PublicationCollaborator} & \mid \mathsf{Researcher} \sqcap \exists \mathsf{sharePublication.Researcher}) = 0.91 \\
P(\mathsf{SupervisionCollaborator} & \mid \mathsf{Researcher} \sqcap \exists \mathsf{wasAdvised.Researcher}) = 0.94 \\
P(\mathsf{SameInstitution} & \mid \mathsf{Researcher} \sqcap \exists \mathsf{hasSameInstitution.Researcher}) = 0.92 \\
P(\mathsf{SameBoard} & \mid \mathsf{Researcher} \sqcap \\
 & \exists \mathsf{sameExaminationBoard.Researcher}) = 0.95 \\
P(\mathsf{NearCollaborator} & \mid \mathsf{Researcher} \sqcap \exists \mathsf{sharePublication.} \exists \mathsf{hasSameInstitution.} \\
 & \exists \mathsf{sharePublication.Researcher}) = 0.95 \\
\mathsf{FacultyNearCollaborator} \equiv & \mathsf{NearCollaborator} \\
 & \sqcap \exists \mathsf{sameExaminationBoard.Researcher} \\
P(\mathsf{NullMobilityResearcher} & \mid \mathsf{Researcher} \sqcap \exists \mathsf{wasAdvised.} \\
 & \exists \mathsf{hasSameInstitution.Researcher}) = 0.98 \\
\mathsf{StrongRelatedResearcher} \equiv & \mathsf{Researcher} \\
 & \sqcap (\exists \mathsf{sharePublication.Researcher} \sqcap \\
 & \exists \mathsf{wasAdvised.Researcher}) \\
\mathsf{InheritedResearcher} \equiv & \mathsf{Researcher} \\
 & \sqcap (\exists \mathsf{sameExaminationBoard.Researcher} \sqcap \\
 & \exists \mathsf{wasAdvised.Researcher})
\end{array}
$$

In this probabilistic ontology concepts and probabilistic inclusions denote mutual research interests. For instance, a PublicationCollaborator inclusion refers to Researchers who shares a Publication, thus relates two nodes (instances of concept Researcher) in a collaboration graph. Therefore, the concept Researcher and the role sharePublication are inputs to the algorithm we proposed in Algorithm 1. Moreover, their instances were used to define a collaboration network, which was also provided to the algorithm. Topological graph information was computed accordingly. Figure 2 depicts a subset of collaborations among researchers. To perform inferences and therefore to obtain link predictions we resort to the lifted algorithm in CR$\mathcal{ALC}$.

If we carefully inspect this collaboration graph we could be interested, for instance, in predicting links among researchers from different groups. Since filling form is prone to errors, there is uncertainty regarding real collaborations. Thus, in Figure 2 one could further investigate whether a link between researcher R (red octagon node) and the researcher B (blue polygon node) is suitable.

In order to infer this, the probability of a possible link between R and B is calculated, $P(link(R, B)|E)$, where E denotes evidence about researchers such as

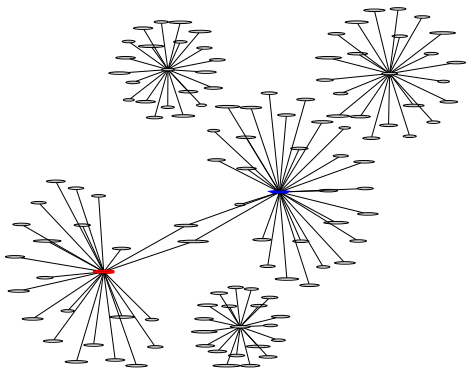

Fig. 2. Lattes collaboration network.

publications, institution, examination board participations and so on. Since the role sharePublication defines the semantics of the links in the graph, it is through it that we must calculate $P(link(R,B)|E)$. Concept PublicationCollaborator is defined by the role sharePublication and considering as evidence $\mathsf{Researcher}(R) \sqcap \exists \mathsf{hasSameInstitution.Researcher}(B)$ one can infer $P(link(R,B)|E)$ through:

$$P(\mathsf{PublicationCollaborator(R)} \mid \mathsf{Researcher(R)} \sqcap \exists \mathsf{hasSameInstitution.Researcher}(B)) = 0.57.$$

If we took a threshold of 0.60, the link between R and B would not be included.

One could gain more evidence, such as information about nodes that indirectly connect these two groups (Figure 2), denoted by I_1, I_2. The inference would be

$$P(\mathsf{PublicationCollaborator(R)} \mid \mathsf{Researcher(R)} \sqcap \exists \mathsf{sharePublication}(I_1).\exists \mathsf{sharePublication(B)} \sqcap \exists \mathsf{sharePublication}(I_2).\exists \mathsf{sharePublication(B)}) = 0.65.$$

Because more information was provided the probability inferred was different. The same threshold now would preserve the link.

In order to compare with existing baseline algorithms, topological and semantic features have also been defined. Further details are given as follows.

3.2 Methodology

In this section we describe our main design choices to run experiments. According to cross validation principles, our dataset (1100 researchers which give rise to

1400 true co-authoring links) has been divided in training and validation sets. To avoid skeweness (due to unbalanced classes), every fold is comprised by balanced negative and positive instances, where positive instances correspond to a link between two nodes while negative instance means that there is not a link between these two nodes.

In order to classify possible co-authoring links and therefore to perform comparisons with previous approaches we resort to the Logistic regression classification algorithm.

In a classification approach for link prediction, features are commonly extracted from topological graph properties such as neighborhood and paths between nodes. In addition, numerical features are also computed from joint probability distributions and semantics.

Two baseline graph-based numerical features have been used in our experiments. First, the Katz measure [10] is a weighted sum of the number of paths in the graph that connect two nodes, with higher weight for shorter paths. This leads to the following formula:

$$Katz(x, y) = \sum_{i=1}^{\infty} \beta^i p_i$$

where p_i is the number of paths of length i connecting x and y, while β (≤ 1) parameter is used to regularize this feature. A small value of β considers only the shorter paths.

Since computing all paths (∞) is expensive we only consider paths of length at most four ($i \leq 4$).

The second numerical feature is the Adamic-Adar measure [1] which computes the similarity between two nodes in a graph. Let $\Gamma(x)$ be the set of all neighbors of node x. Then the similarity between two nodes x, y is given by

$$\text{Adamic-Adar}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

The intuition behind the score is that instead of simply counting the number of neighbors shared by two nodes, we should weight the hub nodes less and rarer nodes more. In this way, Adamic-Adar weighs the common neighbors with smaller degree more heavily.

We have also considered semantic features. The degree of semantic similarity among entities is something that can be useful to predict links that might not be captured by either topological or frequency-based features [20]. In this work, for each author a document with the words appearing in the title of his publications (removing stop words) is considered. Thus, an author is represented as a set of words, which allow us to compute two features based on semantic similarity:

i The keyword match count between two authors [6].
ii The cosine between the TFIDF features vectors of two authors [20].

To compute (ii), we derive a bag of words representation for each author, weighting each word by its TFIDF (Term Frequency - Inverse Document Frequency) measure. The TFIDF weighting scheme assigns to term t a weight in document d given by

$$\text{TFIDF}_{t,d} = \text{TF}_{t,d} \times \text{IDF}_t$$

$\text{TF}_{t,d}$ is the term frequency in d, and IDF_t denotes the inverse document frequency of t which is given by $\text{IDF}_t = \log \frac{\text{N}}{\text{DF}_t}$, where N is the total number of documents and DF_t is the number of documents containing the term.

The standard way of quantifying the similarity between two documents d_1 and d_2 is to compute the cosine similarity of their vector representations $\overrightarrow{V}(d_1)$ and $\overrightarrow{V}(d_2)$

$$\text{cosine}(d_1, d_2) = \frac{\overrightarrow{V}(d_1) \cdot \overrightarrow{V}(d_2)}{|\overrightarrow{V}(d_1)||\overrightarrow{V}(d_2)|}$$

where the numerator represents the dot product (also known as the inner product) of the vectors $\overrightarrow{V}(d_1)$ and $\overrightarrow{V}(d_2)$, while the denominator is the product of their Euclidean lengths.

Finally, we also use the probability, $P(r(x, y)|evidence)$, given by our probabilistic description logic model, as a numerical feature in the classification model. We wish to investigate whether this probabilistic logic measure can improve the classification approach for link prediction.

3.3 Results

In order to evaluate suitability of our approach in predicting co-authorships in the Lattes dataset, three experiments were run. In the first experiment two baseline scores, Katz and Adamic-Adar, have been used as features in the logistic regression algorithm. After a ten-fold cross validation process the classification algorithm yielded results on accuracy which are depicted in Table 1.

Given the Lattes dataset, one can see that the Katz feature yields the best accuracy (75.49%) when the two topological features are used in isolation. Katz has been shown to be among the most effective topological measures for the link prediction task [10]. Furthermore, when we combine the Katz and the Adamic-Adar features, we improve the accuracy to 76.44%.

Table 1. Classification results on accuracy (%) for baseline features: Adamic-Adar (Adamic), Katz and a combined one (Adamic+Katz)

	Adamic	Katz	Adamic+Katz
Lattes dataset	72.75 ± 1.87	75.49 ± 2.07	**76.44 ± 2.03**

In the second experiment, we evalute two features based on semantic similarity and their combination with topological features. Results on accuracy for

these semantic features are depicted in Table 2. The cosine similarity feature performs better than matching keyword feature and outerperforms the two former topological features. This feature alone yields 82.45% on accuracy. When we combine all the four features together, there is an improvement in accuracy to 85.63%.

Table 2. Classification results on accuracy(%) for semantic similarity features: matching keyword (match) and cosine similarity (cosine) and topological features.

	match	cosine	Adamic+Katz+match+cosine
Lattes dataset	69.42 ± 2.66	82.45 ± 1.37	**85.63 ± 1.23**

In the third experiment, a probabilistic feature based on our probabilistic description logic approach was introduced into the model. Results on accuracy for this feature are depicted in Table 3. The probabilistic description logic feature performs better than the other features. This feature yields 87.72% on accuracy. When we combine all the five features together, there is an improvement in accuracy to 89.48%.

Table 3. Classification results on accuracy(%) for probabilistic description logics and baseline features: CR$\mathcal{ALC}$ based (cralc) and Adamic-Adar, Katz, match, cosine, CR$\mathcal{ALC}$ (Adamic+Katz+match+cosine+cralc).

	cralc	Adamic+Katz+match+cosine+cralc
Lattes dataset	87.72 ± 0.52	**89.48 ± 0.96**

It is worth noting that the probabilistic logic feature probability outer performs all other features and allow us to improve the classification model for link prediction on accuracy.

Nothing prevent us to define ad-hoc probabilistic networks to estimate link probabilities. However, by doing so we are expected to define a large propositionalized network (a relational Bayesian network) [15] or estimate local probabilistic networks [20]. These approaches do not scale well since computing probabilistic inference for large networks is expensive.

To overcome these performance and scalability issues, we resort to lifted inference in CR$\mathcal{ALC}$ which is based on variational methods — tunned by evidence defined according nodes's neighborhood. Thus, for a ten thousand network, if evidence is given for 5 nodes, then there is only 6 slices which have messages interchanged.

In our experiments, the average runtime for inference in CR$\mathcal{ALC}$ (1100 nodes network) was 135 milliseconds. Table 4 depicts some runtime results for larger networks which demonstrates the scalability of our approach.

Table 4. Average runtime for inference.

nodes	runtime(milliseconds)
1100	135
10000	168
100000	175
1000000	185

On the other hand, a propositionalized relational Bayesian network fails to run inference due to out of memory issues.

4 Conclusion

In [15, 14] we have presented an approach for predicting links that resorts to both graph-based and ontological information. Given a collaborative network, we encode interests and graph features through a CR$\mathcal{ALC}$ probabilistic ontology. In order to predict links we resort to probabilistic inference, where only information about two nodes being analyzed and the nodes in their path are used as evidence. Thus, making the proposal scalable. In this paper, we evaluated our proposal focused on an academic domain, and we aimed at predicting links among researchers. The approach was successfully compared with graph-based and semantic-based features. As future work we intend to consider other datasets.

Previous combined approaches for link prediction [4, 2] have focused on machine learning algorithms [12]. In such schemes, numerical graph-based features and ontology-based features are computed; then both features are input into a machine learning setting where prediction is performed. Unless from such approaches, in our work we adopt a generic ontology (instead of a hierarchical ontology, expressing only is-a relationships among interests). Therefore, our approach uses more information about the domain to help the prediction. Moreover, in [19], a Probabilistic Relational Model is used for link prediction task. This is one of the approaches more closed to ours, since uses semantic features considering a probabilistic graphical model. However, inference is done in a propositionalized network that can not scale for large networks.

Acknowledgements

The third author is partially supported by CNPq. The work reported here has received substantial support by FAPESP grant 2008/03995-5 and FAPERJ grant E-26/111484/2010.

References

1. L.A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2001.

2. W. Aljandal, V. Bahirwani, D. Caragea, and H.W. Hsu. Ontology-aware classification and association rule mining for interest and link prediction in social networks. In *AAAI 2009 Spring Symposium on Social Semantic Web: Where Web 2.0 Meets Web 3.0*, Standford,CA, 2009.
3. F. Baader and W. Nutt. Basic description logics. In *Description Logic Handbook*, pages 47–100. Cambridge University Press, 2007.
4. D. Caragea, V. Bahirwani, W. Aljandal, and W.H. Hsu. Ontology-based link prediction in the livejournal social network. In *SARA'09*, pages 1–1, 2009.
5. F. G. Cozman and R. B. Polastro. Complexity analysis and variational inference for interpretation-based probabilistic description logics. In *Conference on Uncertainty in Artificial Intelligence*, 2009.
6. Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.
7. J. Heinsohn. Probabilistic description logics. In *International Conf. on Uncertainty in Artificial Intelligence*, pages 311–318, 1994.
8. M. Jaeger. Relational Bayesian networks: a survey. *Linkoping Electronic Articles in Computer and Information Science*, 6, 2002.
9. J. Kunegis and A. Lommatzsch. Learning spectral graph transformations for link prediction. In *Proceedings of the ICML*, pages 561–568, 2009.
10. D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. *American Society for Information Science and Technology*, 7(58):1019–1031, 2007.
11. T. Lukasiewicz and U. Straccia. Managing uncertainty and vagueness in description logics for the semantic web. *Semantic Web Journal*, 6(4):291–308, 2008.
12. T. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
13. A.H. Mohammad and J.Z. Mohammed. A survey of link prediction in social networks. In *Social Network Data Analytics*, pages 243–275. 2011.
14. J. Ochoa-Luna, K. Revoredo, and F.G. Cozman. A scalable semantic link prediction approach through probabilistic description logics. In *In proceeding of 9th Artificial Intelligence National Meeting (ENIA) - to appear*, 2012.
15. K. Revoredo, J. Ochoa-Luna, and F.G. Cozman. Semantic link prediction through probabilistic description logics. In *Bobillo, F., et al. (eds.) Proceedings of the 7th International Workshop on URSW*, volume 778, pages 87–97.
16. K. Revoredo, J. Ochoa-Luna, and F.G. Cozman. Learning probabilistic description logics: A framework and algorithms. In *In proceedings of the MICAI*, volume 7094 of *LNCS*, pages 28–39. Springer, 2011.
17. M. Sachan and R. Ichise. Using semantic information to improve link prediction results in network datasets. *International Journal of COmputer Theory and Engeneering*, 3:71–76, 2011.
18. F. Sebastiani. A probabilistic terminological logic for modelling information retrieval. In *ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 122–130, 1994.
19. B. Taskar, M. faiWong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Proceedings of the 17th NIPS*, 2003.
20. C. Wang, V. Satuluri, and S. Parthasarathy. Local probabilistic models for link prediction. In *Proceedings of the 2007 Seventh IEEE ICDM*, pages 322–331, Washington, DC, USA, 2007. IEEE Computer Society.
21. T. Wohlfarth and R. Ichise. Semantic and event-based approach for link prediction. In *Proceedings of the 7th International Conference on Practical Aspects of Knowledge Management*, 2008.