Semantic Link Prediction through Probabilistic Description Logics

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

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

Abstract. Predicting potential links between nodes in a network is a problem of great practical interest. Link prediction is mostly based on graph-based features and, recently, on approaches that consider the semantics of the domain. However, there is uncertainty in these predictions; by modeling it, one can improve prediction results. In this paper, we propose an algorithm for link prediction that uses a probabilistic ontology described through the probabilistic description logic CR$\mathcal{ALC}$. We use an academic domain in order to evaluate this proposal.

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 recently 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 [12, 20] aims at predicting whether two nodes (i.e. people) should be connected given that we know previous information about their relationships or interests. A common approach is to exploit the network structure, where numerical information about nodes is analyzed [12, 20, 9]. However, knowledge about the objects represented in the nodes can improve prediction results. For instance consider that the researchers *Joe* and *Mike* do not have a publication in common, thus they do not share a link in a collaboration network. Moreover, graph features do not indicate a potential link between them. However, they have published in the same journal and they both teach the same course in their respectively universities. This information can be an indication of a potential collaboration between them. Given this, approaches that are based on the semantics related to the domain of the objects represented by the nodes [21, 18] have been proposed. In some of them, an ontology modeling the domain and the object interests were used in the prediction task.

However, there is uncertainty in such predictions. Often, it is not possible to guarantee the relationship between two objects (nodes). This is maybe due to the fact that information about the domain is incomplete. Thus, it would be interesting if link prediction approaches could handle the *probability* of a link conditioned on the information about the domain. In our example, knowing that the probability of the relationship between *Joe* and *Mike* conditioned on the knowledge of them publishing in the same journal and teaching the same course is high implies a link between them in the network; otherwise, a link is not suggested. In graph-based approaches, probabilistic models learned through machine learning algorithms were used for link prediction. Some examples of probabilistic models are Probabilistic Relational Model (PRM) [6], Probabilistic Entity Relationship Model (PERM) [7] and Stochastic Relational Model (SRM) [22]. On approaches based on semantic we claim that ontologies must be used to model the domain. Therefore, to model uncertainty, probabilistic approaches, such as probabilistic ontologies, must be considered.

An ontology can be represented through a description logic [2], which is typically a decidable fragment of first-order logic that tries to reach a practical balance between expressivity and complexity. To encode uncertainty, a probabilistic description logic (PDL) must be contemplated. The literature contains a number of proposals for PDLs [8, 10, 19, 13]. In this paper we adopt a recently proposed PDL, called Credal $\mathcal{ALC}$ (CR$\mathcal{ALC}$) [4, 16, 5], that extends the popular logic $\mathcal{ALC}$ [2]. In CR$\mathcal{ALC}$ one can specify 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. These sentences are called *probabilistic inclusions*. Exact and approximate inference algorithms that deal with probabilistic inclusions have been proposed [4, 5], using ideas inherited from the theory of Relational Bayesian Networks (RBN)[11].

In this paper, we propose to use a probabilistic ontology defined with the PDL CR$\mathcal{ALC}$ for semantic link prediction.

The paper is organized as follows. Section 2 reviews basic concepts of PDLs and CR$\mathcal{ALC}$. Section 3 presents our algorithm for semantic link prediction through the PDL CR$\mathcal{ALC}$. Experiments are discussed in Section 4, and Section 5 concludes the paper.

2 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) [2]. Knowledge is expressed in terms of *individuals*, *concepts*, and *roles*. The semantic 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 (PDLs) have appeared in the literature. Heinsohn [8], Jaeger [10] and Sebastiani [19] 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 $\mathsf{Professor}$. This characterizes an *interpretation-based* semantics.

The PDL $\text{CR}\mathcal{ALC}$ 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 $\text{CR}\mathcal{ALC}$ 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.

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 depicts $\mathcal{G}(\mathcal{T})$.

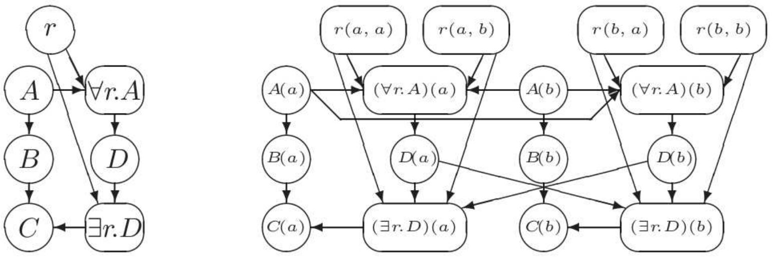

Fig. 1. $\mathcal{G}(\mathcal{T})$ for terminology $\mathcal{T}$ in Example 1 and its grounding for domain $\mathcal{D} = \{a, b\}$.

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 and probabilistic inference (for exact calculations) or by a first order loopy propagation algorithm (for approximate calculations) [5].

3 Link Prediction by using CR$\mathcal{ALC}$

In this section we describe how to apply the PDL CR$\mathcal{ALC}$ for semantic link prediction. We borrowed some syntax from the graph-based approach where each node (a person in a social network) is represented by A, B, C, and we are interested in defining whether a link between A and B is suitable given there is no link between these nodes. Interests between the nodes are modeled through a probabilistic ontology represented by the PDL CR$\mathcal{ALC}$. The prediction link task can be described as:

Given:

- a network defining relationship between objects;
- an ontology represented by CR$\mathcal{ALC}$ describing the domain of the objects;
- the ontology role that defines the semantic of the relationship between objects;
- the ontology concept that describes the network objects.

Find:

- a revised network defining relationship between objects.

The proposed algorithm for link prediction receives a network of a specific domain. For instance, in a collaboration network the nodes represent researchers and the relationship can have the semantic "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, Publication, StrongRelatedResearcher and NearCollaborator and roles like hasPublication, hasSameInstitution and sharePublication. This ontology can be learned automatically through a learning algorithm as the ones proposed in [15, 17]. Thus, the nodes represent instances of one of the concepts described in the PDL CR$\mathcal{ALC}$ and the semantic of the links is described by one of the roles in the PDL 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 C defined as the concept that provides the semantic for the network nodes. For each pair it checks whether the corresponding nodes exist in the network (this can be improved by exploring graph-based properties). If not the probability of the link is calculated through the probability of the defined role conditioned on evidence. The evidence is provided by the instances of the ontology. As many instances the ontology have the better is the inference performed. The inference is performed through the Relational Bayesian network build from ontology O. If the probability inferred is greater than a threshold then the corresponding link

```
Require: a network $N$, an ontology $O$, the role $r(\_,\_)$ representing the semantic of
    the network link, the concept $C$ describing the objects of the network and a
    $threshold$.
Ensure: a revised network $N_f$
define $N_f$ as $N$
for all pair of instances $(a, b)$ of concept $C$ do
    if does not exist a link between nodes $a$ and $b$ in the network $N$ then
        infer probability $P(r(a, b)|evidences)$ using the RBN created through the
        ontology $O$
        if $P(r(a, b)|evidences) > threshold$ then
            add a link between $a$ and $b$ in network $N_f$
        end if
    end if
end for
```

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

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.

4 Preliminary Results

Experiments were run over a collaborative network of researchers. Data was gathered from the Lattes curriculum platform [3], the public repository for Brazilian curriculum researchers. In this platform, every researcher has a unique Lattes code that allows one to link to other researchers according to: shared publications, advising tasks, and examination board participations. Given this collaborative network we are interested in predicting further links among researchers in order to either promote further collaborations (suitable co-workers to research tasks would be suggested) or gather information about research groups. Due to form-filling errors there are many missing links among researchers; thus, we are unable to completely state co-working relationships using only the Lattes platform.

To tackle link prediction we firstly have collected information about 1200 researchers and learned a probabilistic ontology [15, 17], represented by the PDL cr$\mathcal{ALC}$, for modeling their research interests. A simplified probabilistic ontology

[3] http://lattes.cnpq.br/

is given by:

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

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 (Researcher) in a collaboration graph. Therefore, the concept Researcher and the role sharePublication are inputs to the algorithm we proposed in Algorithm 1.

To perform inferences and therefore to obtain link predictions, a propositionalization step (a resulting relational Bayesian network) is required.

In addition, a collaboration graph, based on shared publications, was also defined. Statistical information was computed accordingly. Figure 2 depicts collaborations among 303 researchers. Several relationships and clusterings can also be observed.

If we carefully inspect this collaboration graph (Figure 3 shows a subgraph obtained from Figure 2) we could be interested, for instance, in predicting links among researchers from different groups.

Thus, in Figure 3 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

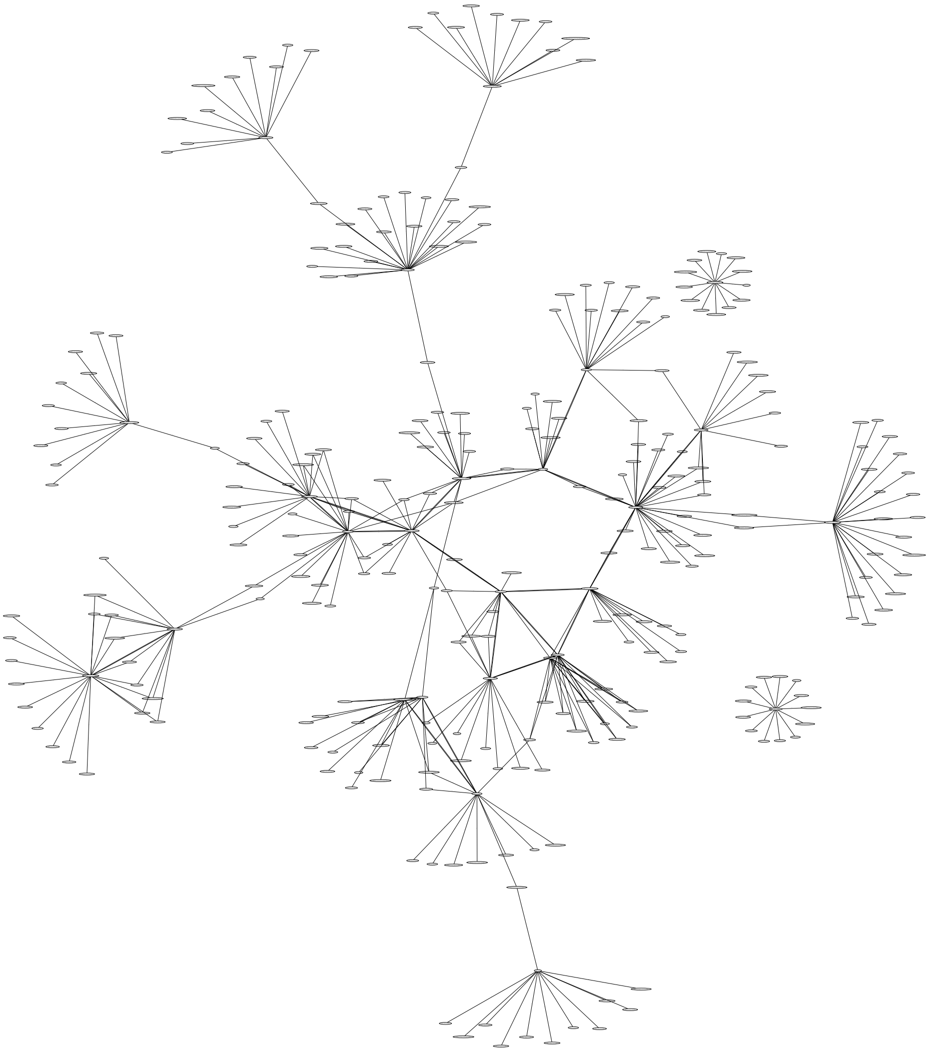

Fig. 2. Collaboration graph among researchers.

B is calculated, $P(link(R,B)|E)$, where E denotes evidence about researchers such as publications, institution, examination board participations and so on. The role $\mathsf{sharePublication}$ is the one defining the semantic of the links in the graph. Therefore, it is through it that we must calculate $P(link(R,B)|E)$. Since the concept $\mathsf{PublicationCollaborator}$ is defined by the role $\mathsf{sharePublication}$ and considering as evidence $\mathsf{Researcher}(R) \sqcap \exists \mathsf{hasSameInstitution}.\mathsf{Researcher}(B)$ one can infer $P(link(R,B)|E)$ through:

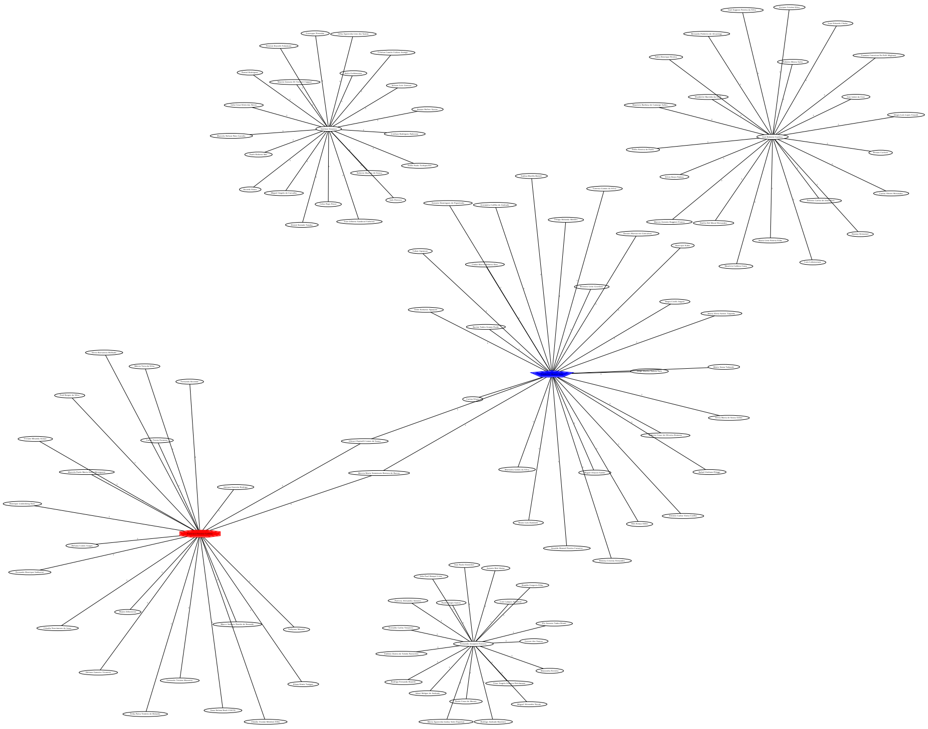

Fig. 3. Collaboration subgraph.

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

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 3), denoted by I_1, I_2. The inference would be

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

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

Other inferences are possible by considering the suggestion of links between surrounding nodes, i.e. nodes directly linked to the two nodes R and B , denoted by $R_1, \ldots, R_k$, and $B_1, \ldots, B_n$ respectively. For each $i = 1, ..., k$ and $j = 1, ..., n$, calculates $P(link(R_i, R_j)|E)$ and $P(link(B_i, B_j)|E)$.

As a rule, if we are interested in discovering whether A and B could be linked, probabilistic inference $P(link(A,B))$ should be performed.

In a more general framework, graph information could be useful to deal with a large number of link predictions. Note that graph adjacency allow us to address probabilistic inference for promising nodes. In a naive approach, each pair of nodes in the collaboration graph would be evaluated so, multiple probabilistic relational Bayesian inference calls would be required.

On the other hand, if graph-based information is used, such naive scheme could be improved. In our approach, two nodes are probabilistically evaluated if there is a path between them (number of incoming/outgoing edges, number of mutual friends, node distances are also considered). Thus, numerical graph-based information guides the inference process in the relational Bayesian network (linked to the probabilistic ontology). In addition, other candidates sharing any kind of evidence are also evaluated, i.e., interests based features (linked to ontological knowledge) allow us to further explore link prediction.

Alternatively, by completing an overall link predicting task we can devise further functionalities to the resulting collaboration network. The resulting graph can be considered as being a probabilistic network, i.e., probabilities inferred for each link could be denote strenght of the relationship.

5 Conclusion

We have presented an approach for predicting links that resorts both to graph-based and ontological information. Given a collaborative network, e.g., a social network, we encode interests and graph features through a CR$\mathcal{ALC}$ probabilistic ontology. In order to predict links we resort to probabilistic inference. Preliminary results focused on an academic domain, and we aimed at predicting links among researchers. These preliminary results showed the potential of the idea.

Previous combined approaches for link prediction [3, 1] have focused on machine learning algorithms [14]. 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. Differently 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.

Acknowledgements

The third author is partially supported by CNPq. The work reported here has received substantial support by FAPESP grant 2008/03995-5.

References

1. 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.

2. F. Baader and W. Nutt. Basic description logics. In *Description Logic Handbook*, pages 47–100. Cambridge University Press, 2002.
3. D. Caragea, V. Bahirwani, W. Aljandal, and W. H. Hsu. Ontology-based link prediction in the livejournal social network. In *Symposium on Abstraction, Reformulation and Approximation*, 2009.
4. F. G. Cozman and R. B. Polastro. Loopy propagation in a probabilistic description logic. In Sergio Greco and Thomas Lukasiewicz, editors, *Second International Conference on Scalable Uncertainty Management*, Lecture Notes in Artificial Intelligence (LNAI 5291), pages 120–133. Springer, 2008.
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. N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proc. 16th Int. Joint Conference on Artificial Intelligence*, pages 1300–1309, 1999.
7. D. Heckerman, C. Meek, and D. Koller. Probabilistic entity-relationship models, prms, and plate models. In *In Proceedings of the 21st International Conference on Machine Learning*, 2004.
8. J. Heinsohn. Probabilistic description logics. In *International Conf. on Uncertainty in Artificial Intelligence*, pages 311–318, 1994.
9. W.H. Hsu, A.L. King, M.S.R. Paradesi, T. Pydimarri, and T. Wneinger. Collaborative and structural recommendation of friends using weblog-based social network analysis. In *Proceedings of Computational Approaches to Analysing WebLogs (AAAI)*, 2006.
10. M. Jaeger. Probabilistic reasoning in terminological logics. In *Principals of Knowledge Representation (KR)*, pages 461–472, 1994.
11. M. Jaeger. Relational Bayesian networks: a survey. *Linkoping Electronic Articles in Computer and Information Science*, 6, 2002.
12. D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and Knowledge Management*, pages 556–559, New York, NY, USA, 2003. ACM.
13. T. Lukasiewicz and U. Straccia. Managing uncertainty and vagueness in description logics for the semantic web. *Web Semant.*, 6:291–308, November 2008.
14. T. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
15. J. Ochoa-Luna, K. Revoredo, and F.G. Cozman. Learning sentences and assessments in probabilistic description logics. In *Bobillo, F., et al. (eds.) Proceedings of the 6th International Workshop on Uncertainty Reasoning for the Semantic Web*, volume 654, pages 85–96, Shangai, China, 2010. CEUR-WS.org.
16. R. B. Polastro and F. G. Cozman. Inference in probabilistic ontologies with attributive concept descriptions and nominals. In *4th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW) at the 7th International Semantic Web Conference (ISWC)*, Karlsruhe, Germany, 2008.
17. K. Revoredo, J. Ochoa-Luna, and F. Cozman. Learning terminologies in probabilistic description logics. In Antônio da Rocha Costa, Rosa Vicari, and Flavio Tonidandel, editors, *Advances in Artificial Intelligence SBIA 2010*, volume 6404 of *Lecture Notes in Computer Science*, pages 41–50. Springer / Heidelberg, Berlin, 2010.

18. 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.
19. 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.
20. B. Taskar, M. FaiWong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Proceedings of the 17th Neural Information Processing Systems (NIPS)*, 2003.
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.
22. K. Yu, W. Chu, S. Yu, V. Tresp, and Z. Xu. Stochastic relational models for discriminative link prediction. In *Proceedings of the Neural Information Processing Systems (NIPS)*, 2006.