Research Statement

Ni Lao

I feel very excited to work at the intersection of machine learning, information retrieval, and natural language processing. Specifically, I am working on prediction and recommendation tasks on heterogeneous graphs. Before this, I have worked on a wide range of projects including syntactic and sematic parsing, question answering, utility based retrieval evaluation, learning to rank, contextual search, modeling species distributions, automatic system diagnosis, and robotic soccer players.

1 Research Problems and Challenges

Current information extraction, integration and retrieval research have entered the stage of relational learning. Diverse data types such as users, documents, locations, and parsed text can be modeled as connected heterogeneous graphs. The fact that a node in the graph is connected to other nodes through different types of edges is important to how we understand text, how we organize human knowledge, and how to satisfy users' information needs. For example, intelligent systems may infer users' intentions by considering the context of their past behavior (e.g. click data on page links and ads, posts on other users' Blogs) and their relation to other users. These systems can therefore send the right information to the right people at the right time. Although inductive logic programming is a general framework for a wide range of applications, it cannot be applied to most realistic problems due to its inefficiency in learning. Developing general purpose yet practical models faces many challenges:

- **Complex feature generation and selection**. The attributes of a node can be correlated to attributes of other nodes connected by a sequence of relations. A major challenge here is feature selection. the number of possible ways in which two entities are connected is exponential in the length of the connection. Therefore, it is important to constrain the family of paths which we consider. Criteria like the gradient of a loss function are effective ways to pick out the most useful features.
- **Efficient inference**. Not only do we need to consider models with complex features, but we also need to consider datasets of large sizes. Graphical models are very good for joint inferences, but they are not applicable to large scale problems unless some approximate inference is applied. On the other hand, random walk based inference scale well with large data sets.
- **Hidden concept discovery**. For relational learning, hidden variables reduce the need to model long range dependencies, which can be computationally prohibitive. For example, the feature IsMotherOf(A,B) $\rightarrow \neg$ IsFatherOf(A,C) connects nodes B,C which are two relations (IsMotherOf, IsFatherOf) apart from each other. However, inference and learning for a model with all such second order features might be very expensive. On the other hand, if the model is able to discover the concept of gender, then the same knowledge can be captured by a pair of first order features (IsMotherOf(A,B) $\rightarrow$Female(A), Female(A) $\rightarrow \neg$ IsFatherOf(A,C)), which belong to a family of simpler and more efficient models. An effective way of introducing hidden variables to CRF models is to measure the improvement of log-likelihood.

My research strives to develop statistically sound approaches in face of these challenges, and apply them in real world applications.

2 Research Accomplishments

My thesis develops prediction and recommendation algorithms that leverage complex relation patterns. In one project we help scientists find information in a network of social relationships (e.g. co-authorship) as well as semantic annotations (e.g. topics, named entity annotations). The information-access tasks is modeled as proximity queries-- with typed nodes representing documents, terms, and metadata, and labeled edges representing the relationships between them. We propose a novel method for learning a weighted combination of Path-Constrained Random Walkers, which is able to discover and leverage complex path features (Lao & Cohen, Machine Learning 2010). Our experiments show that the proposed method gives 10% to 40% improvement in mean average precision (MAP) over other random walk based models. We also experiment with different approximation (e.g. truncation, sampling, particle filtering) to the random walk process. The speedup techniques give up to a 100-fold query execution speedup with little loss in MAP (Lao & Cohen, KDD 2010).

I am also working on the problem of performing learning and inference in large scale knowledge bases containing imperfect knowledge with incomplete coverage. We show that a soft inference procedure based on Path-Constrained Random Walkers through the knowledge base graph can be used to reliably infer new beliefs for the knowledge base (Lao, Mitchell, Cohen, EMNLP 2011). The new system gives much higher recall than FOIL (First Order Inductive Learner) based approach.

Another topic I am working on is feature selection and Hidden Concept Discovery (HCD) for relational CRFs (Lao, et al., NIPS 2010). First, based on the concept of relation trees, we introduce a way to automatically construct arbitrarily complex relational CRFs from the schema of a domain. Second, we give an efficient feature selection method to handle the large number of automatically generated features, which is an approximation to Grafting (Zhu et al. KDD 2010). Third, we develop a novel HCD algorithm, which introduces a new hidden variable whenever it is estimated to improve the Mean Field Contrastive Divergence objective function. We conducted experiments on datasets involving entities like persons, countries, and biomedical concepts, and infer their attributes (e.g. gender, age) and relations (e.g. MilitaryAlliance(UK,USA)). Experiment results show that the proposed method can achieve similar prediction quality as the state-of-the-art approaches, but is significantly faster in training.

3 Future Directions

I will continue to pursue information retrieval and machine learning from the perspective of relational learning, and here are some future directions:

- **Efficient inductive logical programing**. First order logic is very expressive in describing knowledge, but has proved to be hard to learn. Based on sound statistic principles, we can develop random walk models that is almost as expressive as first order logic but much more efficient to learn.
- **Learning long relation paths**. We can efficiently explore a large feature space by combining existing relation paths. A sparse approximation of the candidate paths' gains can be estimated by combining forward random walk and backward propagation of prediction errors.
- **Graph structure learning**. Modifying an existing graph structure--i.e. introducing new nodes or edges -- can potentially simplify model complexity and improve performance. Candidate structures can be evaluated by expected gains of the objective function on training data.

In summary, my research aims to answer the fundamental question of how complex relational models can be efficiently learnt from data. This research brings together topics like information retrieval, statistical learning, logics, and can impact many applications in related fields.