Hierarchical Random Walk Inference in Knowledge Graphs

Qiao Liu
qliu@uestc.edu.cn

Liuyi Jiang
harperliuyi@gmail.com

Minghao Han
hanmhao@gmail.com

Yao Liu
liuyao@uestc.edu.cn

Zhiguang Qin
qinzg@uestc.edu.cn

School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu 610054, China

ABSTRACT

Relational inference is a crucial technique for knowledge base population. The central problem in the study of relational inference is to infer unknown relations between entities from the facts given in the knowledge bases. Two popular models have been put forth recently to solve this problem, which are the latent factor models and the random-walk models, respectively. However, each of them has their pros and cons, depending on their computational efficiency and inference accuracy. In this paper, we propose a hierarchical random-walk inference algorithm for relational learning in large scale graph-structured knowledge bases, which not only maintains the computational simplicity of the random-walk models, but also provides better inference accuracy than related works. The improvements come from two basic assumptions we proposed in this paper. Firstly, we assume that although a relation between two entities is syntactically directional, the information conveyed by this relation is equally shared between the connected entities, thus all of the relations are semantically bidirectional. Secondly, we assume that the topology structures of the relation-specific subgraphs in knowledge bases can be exploited to improve the performance of the random-walk based relational inference algorithms. The proposed algorithm and ideas are validated with numerical results on experimental data sampled from practical knowledge bases, and the results are compared to state-of-the-art approaches.

CCS Concepts

•**Information systems** → **Retrieval tasks and goals;** •**Computing methodologies** → **Knowledge representation and reasoning;** *Machine learning algorithms;*

Keywords

Relational inference; Random walk model; Statistical relational learning; Knowledge base; Knowledge graphs

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2911451.2911509

1. INTRODUCTION

The goal of relational inference research is to infer new knowledge (facts) from the existed knowledge bases[6]. This paper considers the problem of relational inference on large-scale *graph-structured knowledge bases* (**GKBs**, a.k.a. *knowledge graphs*), such as Freebase, YAGO and DBpedia, which stores factual information in the form of <entity-relation-entity> triplets. Currently, relational inference remains a challenge as the knowledge bases are plagued with incompleteness and ambiguity. For example, most of the basic information about the public figures that one would think typically available, such as the "Place of Birth"and the "Parents" attributes, are still missing in the latest version of Freebase and alike knowledge bases at this time[14].

However, research on statistical relational learning (SRL) reveals that the existing knowledge in GKBs contain useful information about the latent relations between entities, which can be effectively explored by using statistical learning methods[18]. Current research efforts are largely focused on developing relational learning models that can scale to massive GKBs, among which the latent factor model (**LFM**) and the random-walk model (**RWM**) are the most heavily studied, and some related algorithms have been introduced (instantly) into practical usage[14]. For instance, the *Never-Ending Language Learning* (NELL) project takes the RWM based *path ranking algorithm* (PRA) as their relation reasoning module. While in the Google's *Knowledge Vault* project, a hybrid solution that combines the LFM and the RWM is implemented for knowledge evaluation tasks[20].

Each of these two models has its benefits and limitations. Generally speaking, the RWMs are more computationally efficient than the LFMs, because the LFMs inherently involve a matrix factorization operation over the GKBs, which is difficult to be parallelized. In contrast the RWMs are naturally parallelized, which makes it more scalable for large-scale GKBs. However, according to our empirical studies (see Sec. 4.3), the performance of the RWM solutions are not as competitive as those of some LFM solutions, in terms of inferential accuracy and recall rate. Since both of the efficiency and the accuracy are crucial to the success of relational learning tasks in practice, we present in this study for the first time a comprehensive investigation of the potential benefits of the random-walk model, with the purpose of determining whether it can outperform the most competitive LFM solutions, thus providing new insights into the mechanisms that underlie graph based relational inference.

The principle contribution of this paper is the development of a new random-walk based learning algorithm, called the ***Hi**erarchical **R**andom-walk **i**nference* (**HiRi**) algorithm, for relational inference on GKBs. Specifically, we describe a two-tier random-walk mechanism for relational retrieval, wherein the upper-tier of the model corresponds to the *relation sequence* pattern recognition and learning process in a global perspective, the lower-tier is designed to capture useful information from inside the relation-specific subgraphs in GKBs (which means that each subgraph only represents one specific type of relation). The proposed HiRi algorithm outperforms widely used PRA on two benchmark data sets sampled from real GKBs, achieving an improvement in MRR score of up to 79.5%, and in Hits@1 score of up to 79.2%. The proposed algorithm also outperforms some state-of-the-art LFM based algorithms on all of the data sets.

Another contribution of this paper is that we propose two basic assumptions for relational inference model building:

(1) Since the relations between entities are semantically bidirectional, the knowledge graphs can be modeled with undirected graphs in relational learning tasks.

(2) The reciprocal information transfer between the head and tail entities within the relation-specific cliques of a GKB is of special value in relational learning tasks.

The first assumption is in accordance with intuition, but is in direct conflict with the basic assumption universally accepted by scholars, which represent the GKBs as directed graphs so as to make it consistent with the logical structure of the facts stored in GKBs. The second assumption explains why the latent factor models consistently outperform the random walk models with respect to recall and accuracy, thus it also helps explaining why a hierarchical random-walk mechanism is necessary for RWM-alike solutions. Experimental results are shown to agree with our assumptions, we hope this may shed some light on better understanding the existing methods, and on further study of this problem.

The rest of this paper is organized as follows. We first provide a brief literature review of the related work in Section 2, and then we present a detailed description of our methodology in Section 3. Experiments and discussions are provided in Section 4, and Section 5 concludes this paper.

2. RELATED WORK

Statistical relational learning (SRL) has received a lot of attention in information retrieval and artificial intelligence communities[6]. Methods from SRL research have also been applied to develop link prediction models and context-aware recommender systems[3, 13]. There is an extensive amount of literature available, hence we will only give a very brief overview of the closely related work in this section.

For relational learning in graph-structured knowledge bases, the most commonly-used research method in the past two decades was to develop probabilistic inference models, such as the Markov logic networks[18] and the Bayesian networks[11], based on the first-order logic rules, to infer new facts from existing facts[5]. However, such approaches suffer from the **scalability** problem (due to the computational expense associated with the rule learning process) and the **generalization** problem (caused by the brittleness of the logical rules), which limits their usefulness for relational learning on large-scale GKBs[17]. To catch up with the rapid expansion of the industrial GKBs, several new approaches have been devised, among which the latent factor models (LFMs) and the random-walk models (RWMs) have been in the forefront of academic efforts in recent years[16, 7].

The basic idea of the latent factor models is to obtain a vectorized *representation* for each of the entities and/or relations stored in the GKBs, by transforming it into a low-dimensional subspace, and then infer the missing facts from such *representations*[14]. Depending on the different approximate factorization schemes adopted, the LFMs are also called *tensor decomposition* models[16] or *structured embedding* models[2]. For instance, the RESCAL algorithm tries to represent the relation-specific subgraph of a GKB with a third-order tensor model, in which entities and their relation are mapped into different tensor spaces[15]. While the TransE algorithm treats a relation as a *translating operation* between the corresponding *head* and *tail* entities, thus in TransE, both of the relation and the associated entities are mapped into the same embedding space[1].

The TransH algorithm further improves the TransE by representing relations as hyperplanes, rather than vectors, in an embedding space, which enables an entity to have distinct representations when involved in different type of relations[21]. Both of the TransE and the TransH algorithms were trying to mapping the entities and relations into the same embedding concept space, however, Lin et. al. proposed in TransR model that the entities and relations should be treated differently, because they are conceptually different types of objects, thus should be mapped into different concept spaces, and they claimed that the (inferential) translation should only be performed in the relation space[9].

Another line of research that has also drawn increasing attention are the random-walk based relational learning models. Studies of the RWMs were originally inspired by the idea of the *First Order Inductive Learner* (FOIL), which is a supervised relational learning algorithm based on the Horn clause rules extracted from the GKBs[19]. The seminal work of the RWM for relational learning is the *Path Ranking Algorithm* (PRA), which extends the idea of the FOIL by searching through the GKB for *path features* instead of Horn clause rules[8]. The merits of the PRA algorithm are that its inference results are easily interpretable, and it is inherently parallelizable. In contrast, the meanings of the latent factor models (i.e. the embedding spaces) are hard to be explained. Furthermore, since the LFM solutions require a computational extensive matrix factorization process during the model training stage, it is difficult to be parallelized efficiently. For more details about the recent advances in theories and applications of the LFMs and the RWMs in this area, a good review can be found in [14].

The merits of PRA make it a promising candidate not only for research but also for industrial applications as well. In fact, it has already been used successfully in some large-scale GKB projects, such as the *Knowledge Vault* project of Google, and the *Never-Ending Language Learning* (NELL) project of Carnegie Mellon University[10]. However, according to the comparative study made in this paper, the random-walk based PRA algorithm is obviously at a disadvantage compared with the best performed LFM solutions, in terms of both the inferential accuracy and the recall rate.

Therefore the motivation of this work was to extend the existing studies by addressing the following question: is it possible to design a relational learning model that not only

Table 1: Meanings of notations.

def.	Meanings of notations
$\boldsymbol{G}$	symbol of the knowledge graph
N	number of the SPO triplets in $\boldsymbol{G}$
$\boldsymbol{R}$	set $\boldsymbol{R}$ contains all of the relation types in $\boldsymbol{G}$
n	number of the relation types in $\boldsymbol{R}$
$\boldsymbol{E}$	set $\boldsymbol{E}$ contains all of the unique entities in $\boldsymbol{G}$
m	number of the entities in $\boldsymbol{E}$
r_i	the i-th element in $\boldsymbol{R}$
$\boldsymbol{G}_i$	subgraph of $\boldsymbol{G}$ that only contains relation r_i
$\boldsymbol{H}_i$	set $\boldsymbol{H}_i$ contains all of the *head* entities in $\boldsymbol{G}_i$
$\boldsymbol{T}_i$	set $\boldsymbol{T}_i$ contains all of the *tail* entities in $\boldsymbol{G}_i$
$<h, r, t>$	SPO triplet, the unit of knowledge (fact)
$<h, t>$	ordered entity pair extracted from $<h, r, t>$

maintains the efficiency of the RWMs, but also keeps the accuracy of the LFMs? The proposed HiRi algorithm is a pure random-walk based relational learning algorithm, this is markedly distinct from the efforts of developing hybrid models (the blending of RWMs and LFMs), which has become a major focus of the research community[4, 12].

The most related work to ours is the PRA algorithm, both of them rely on the same pattern discovery strategy and path feature learning framework for relational retrieval modeling from the global perspective. However, there are two significant differences: firstly, the path feature discovery process of HiRi is based on the undirected graph representation of the GKB, although the PRA algorithm also does allow the inverse of a relation to be considered in its path feature discovery process, but only limited to some of the *non-functional* predicates (relations). Secondly, HiRi contains a local inference mechanism, which makes it capable of utilizing the inferable information conveyed in the relation-specific cliques to enhance the performance of the global inference procedure, which is not considered by the PRA algorithm. Experimental results demonstrate the effectiveness of HiRi and the validity of the underlying assumptions, which also provides a positive answer to the above question.

3. THE HIRI ALGORITHM

In this section, we start by introducing some preliminary background and the symbol system used in this paper, then we discuss the intuition and algorithmic framework of the proposed HiRi algorithm. The detailed implementation of the HiRi algorithm is divided into three parts, which will be described in Section 3.3 to 3.5, respectively.

3.1 Preliminary And Notation

Knowledge graphs (KGs) represent facts in the form of binary relations, in particular *<subject, predicate, object>* triplets, where *subject* and *object* are entities and *predicate* is the type of a relation. For simplicity and in accordance with previous studies[9, 20], we will use the notation $<h, r, t>$ to represent the SPO triplets, in which h and t represent the *head* (subject) and *tail* (object) entities and r is the relation between them. The meanings of the major notations used in this paper are given in Table 1. According to previous research, the relation types in a knowledge graph can be artificially classified into four categories by means of the

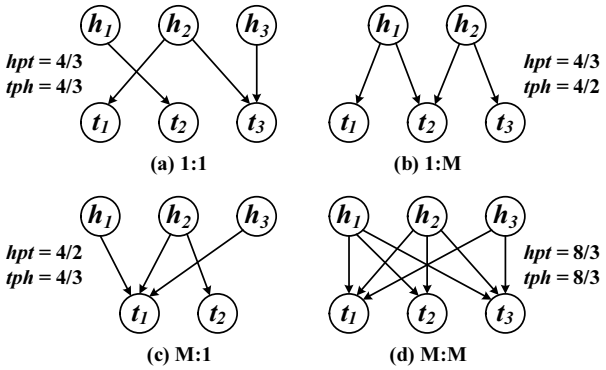

Figure 1: Illustration of the four relation categories.

heads-per-tail ratio (*hpt*) and *tails-per-head* ratio (*tph*):

$$hpt_i = \frac{\#\ triplets_i}{\#\ tails_i}; \quad tph_i = \frac{\#\ triplets_i}{\#\ heads_i}. \tag{1}$$

where $\#\ triplets_i$ denotes the number of SPO triplets in subgraph $\boldsymbol{G}_i$, $\#\ heads_i$ and $\#\ tails_i$ represents the number of *head* and *tail* entities in $\boldsymbol{G}_i$. Then a particular type of relation can be assigned to one of the four categories: one-to-one (1:1), one-to-many (1:M), many-to-one (M:1) and many-to-many (M:M), according to the following criteria.

$$\begin{cases} hpt_i < \delta & \text{and} & tph_i < \delta & \Rightarrow & \text{1:1} & \text{relations} \\ hpt_i < \delta & \text{and} & tph_i \geq \delta & \Rightarrow & \text{1:M} & \text{relations} \\ hpt_i \geq \delta & \text{and} & tph_i < \delta & \Rightarrow & \text{M:1} & \text{relations} \\ hpt_i \geq \delta & \text{and} & tph_i \geq \delta & \Rightarrow & \text{M:M} & \text{relations} \end{cases} \tag{2}$$

where $\delta \geq 1$ is an empirical parameter proposed in [1]. In this paper we follow the same approach as in [1] and [21], and choose $\delta = 1.5$ as the classification criteria. Figure 1 depicts four simple examples in an illustrative manner to facilitate intuitive understanding of this classification method.

3.2 Assumptions And Framework

The basic idea of the HiRi algorithm is simple: figure out why exactly the RWM solutions systematically perform worse than the LFM solutions, and try to fix it. For this purpose, we begin by comparing the performance of two typical algorithms that have attracted considerable attentions recently, which are the RWM based PRA algorithm[8], and the LFM based TransE algorithm[1]. Table 2 provides the comparison of the performance between PRA and TransE by use of the FB15k data sets, the test sets were split into four parts according to Eq. (2). For more details about the experimental settings and protocols, please refer to Sec. 4.

Roughly speaking, the Hits@10 score represents the averaged *hits* ratio on the top 10 prediction results of the corresponding algorithms. From Table 2, one can see that the TransE algorithm clearly performs better than the PRA algorithm on 1:M and M:M relations. Since similar situations are also observed in comparing PRA with other LFM solutions (see Section 4.5), we believe that the presence of such a recurring pattern worth further investigation.

Firstly, notice that the PRA performs comparable to the TransE algorithm on M:1 relations, which is in contrast with the results observed upon 1:M relations. Since if we ignore the directionality of the edges, the topological structures of the entity-relation graphs corresponding to 1:M and M:1 re-

Table 2: Comparison test: Hits@10 on FB15k

Algorithm	1:1	1: M	M:1	M:M
TransE	71.5%	49.0%	85.0%	72.9%
PRA	63.3%	20.4%	81.4%	32.6%

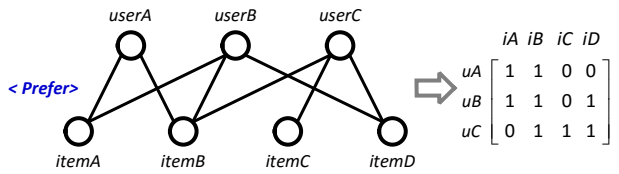

Figure 2: Inference on relation-specific cliques.

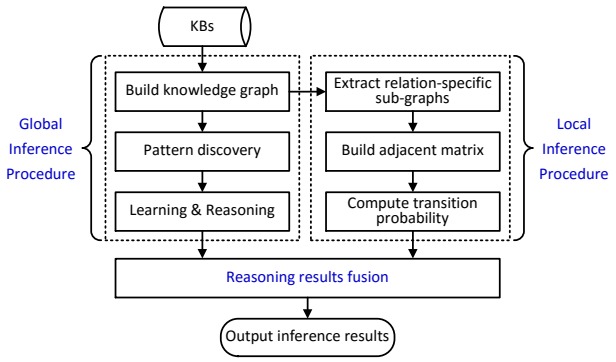

Figure 3: Flow chart of the HiRi algorithm.

lations are very similar to each other (see Figure 1(b) and 1(c)), we suggest that the reason for the different behavior of PRA on these relations is due to the "*directed relation*" assumption made by PRA. PRA models the KBs with directed graphs, and its *path feature* discovery process is a mimic of the logical inference process based on first order Horn clause rules[8]. This assumption is reasonable to some extent, in that it is in accordance with the logical and syntactic constraints of the natural languages. However, a potential problem with this assumption is that it probably underestimates the diversity of the syntax patterns used for relation expressions, and the incompleteness and imbalance of the facts stored in GKBs. This leads to our first assumption:

Assumption I. The semantic information of a relation is reciprocally shared between the connected entities, so it is reasonable to model the knowledge graphs with undirected graphs for relational learning tasks.

Secondly, based on the comparison results on M:M relations, we suggest that the *path feature* discovery process used in PRA is not efficient in utilizing the inferrable information contained in richly connected relation-specific cliques to make inference. Take Figure 2 for example, a collection of the <user-*prefer*-item> triplets is represented by a bipartite graph, the relation "*prefer*" is a typical M:M relation. One can see that since *userA* prefers *itemB*, and *itemB* is preferred by *userB* and *userC*, so it is reasonable to anticipate that *userA* may also be interested in *itemD*, because both of *userB* and *userC* prefer it. For the same reason, we could also anticipate that *userA* might prefer *itemC*, but with less confidence, since we have only one piece of evidence <*userC-prefer-itemC*> to support this inference. From the perspective of PRA, in both of above situations, the inference rule used are the same *path feature*: "$r_i \rightarrow r_i^{-1} \rightarrow r_i$", in which r_i^{-1} denotes the inverse of the relation $r_i \in \boldsymbol{R}$.

The problem with this *path feature* is that it cannot tell the difference between these two situations, and such information should not be ignored. In contrast, the latent factor models (such as TransE) can make full use of such information, by decomposing it into the vector representation of the relations and entities. This explains why TransE performs significantly better than PRA on M:M relations in FB15k data set. From this observation we conclude that some relations are more *reciprocal* and *transitive* than other relations. From the perspective of RWMs, *reciprocal* means that the inverse of the relation is effective in bridging the inference path from the *tail* entities back to the related *head* entities. *Transitive* means that the specified relation is inferable from the *path feature* of the form "$r_i \rightarrow r_i^{-1} \rightarrow \cdots \rightarrow r_i$", which can be modeled with an odd-hop random-walk model. This leads us to the second assumption:

Assumption II. The topological structure of the relation-specific cliques may convey useful information for relation inference, which can be employed to enhance the performance of the random-walk based relational learning algorithms.

In designing of the problem-solving algorithm, we propose a hierarchical structured random-walk model based on above assumptions, a schematic flow chart of the proposed HiRi algorithm is illustrated in Figure 3. Basically, the inference process of HiRi consists of the following steps, greater details of the implementation are given in subsequent sections.

(1) **Global inference procedure**: For each $r_i \in \boldsymbol{G}$, we learn an global inference model $f(h, r_i, t)$ with a revised PRA algorithm from the global knowledge graph $\boldsymbol{G}$, in which $f(h, r_i, t)$ denotes the probability of the random event that there exists a relation r_i between the entity pair $<h, t>$ from the global perspective.

(2) **Local inference procedure**: For each $r_i \in \boldsymbol{G}$, we calculate a 3-hop transition probability matrix from the relation-specific clique $\boldsymbol{G}_i$, so that for any given $<h, r_i, t>$, we could easily compute the probability (denoted by $g(h, r_i, t)$) of the random event that there exists a relation r_i between $<h, t>$, from the local perspective.

(3) **Results fusion procedure**: We merge the results from the aforementioned two steps with a linear model. For any given triplet $<h, r_i, t>$, the final output is a relevance score, denoted by $score(h, r_i, t)$, which indicates the relevance of the entity pair $<h, t>$ with regard to r_i.

3.3 Learning Model For Global Inference

In this paper, we use a revised PRA algorithm for recognizing *path features* and for learning the global inference models for each r_i from $\boldsymbol{G}$. As mentioned above, the major difference between PRA and our implementation is that we use undirected graph for *path feature* discovery. The direct effect is to increase the chance of finding more plausible *features*. Another difference is that we abandon the *importance weight* parameter assigned to each facts to save labor costs.

For completeness, a brief review of the essentials of the PRA algorithm is given in the following, more details can be found in [7] and [8]. For any given relation r_i, PRA tries to find out all of the path sequences within 3-hop from each $h \in \boldsymbol{H}_i$ to its direct neighborhood t, subject to $t \in \boldsymbol{T}_i$. Some additional constraints are also imposed over the random-walkers, such as the first move can not choose r_i.

The qualified path sequences are called the *path features*, which will be used to build a learning model for relational inference on r_i. Let Π_i denotes the *path feature set* of r_i, $\mathbf{x}$ denotes the feature vector of entity pair $<h,t>$, which is an instance of Π_i, $x_j \in \mathbf{x}$ represents the j-th element of $\mathbf{x}$, which is defined as the probability of a random-walker started from entity h, after the path sequence $\pi_j \in \Pi^i$, reached entity t. Let Θ_i represent the corresponding coefficients vector to the feature vector Π_i, $\theta_j \in \Theta_i$ denotes the j-th element of Θ_i. Let $f(h, r_i, t)$ denotes the strength of the possibility that there exists relation r_i in between entity pair $<h,t>$. The global inference model can be represented as:

$$f(h, r_i, t) = \mathbf{x}^T \cdot \Theta_i = \sum_{j=1}^{|\Theta_i|} x_j \theta_j \tag{3}$$

where $|\Theta_i|$ denotes the number of elements in Θ_i. The parameter estimation process is described as follows. Firstly, for each relation $r_i \in \boldsymbol{R}$, we construct a training dataset $\mathbf{Dat}_i = \{(\boldsymbol{x}_k, y_k)\}$ from $\boldsymbol{G}$, where $\boldsymbol{x}_k$ is an instance of Π_i, each $\boldsymbol{x}_k$ is corresponding to a node pair $<h_k, t_k>$ in $\boldsymbol{G}$, $y_k = 1$ if $<h_k, r_i, t_k> \in G_i$, else $y_k = 0$. Secondly, the parameter vector Θ_i is estimated by fitting a penalized logistic regression model, the target function is defined as:

$$\Theta_i = \underset{\Theta_i}{\operatorname{argmax}}(\mathcal{L}(\mathbf{x}, \Theta_i) - \lambda_1 \|\Theta_i\|_1, -\lambda_2 \|\Theta_i\|_2). \tag{4}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are ℓ_1-norm and ℓ_2-norm penalty factors, the ℓ_1 penalty encourages sparsity in the coefficients of Θ_i, while the ℓ_2 penalty shrinks the coefficients to prevent over-fitting. $\mathcal{L}(\mathbf{x}, \Theta)$ is the likelihood function, defined as:

$$\mathcal{L}(\mathbf{x}, \Theta) = \sum_{k=0}^{|\mathbf{Dat}_i|} (y_k \ln q_k + (1 - y_k) \ln(1 - q_k)). \tag{5}$$

where q_k denotes the probability $p(y_k = 1 | \mathbf{x}_k; \Theta_i)$, which is defined with the following sigmoid function:

$$q_k = 1/(1 + \exp(-\mathbf{x}_k^T \cdot \Theta_i)) \tag{6}$$

After getting the coefficients vector Θ^i, for any given entity pairs $<h,t>$, we can compute the *score of global relevance* of $<h, r_i, t>$ with Eq. (3), by constructing a feature vector $\mathbf{x}$ for $<h,t>$ from searching through the global knowledge graph $\boldsymbol{G}$. Next, we are going to compute the *score of local relevance* of $<h, r_i, t>$, according to the second assumption.

3.4 Random Walk For Local Inference

In this section, the scope of inference is limited to the relation-specific clique $\boldsymbol{G}_i$. Our objective is to infer new beliefs from current beliefs by using of the first order Horn clause rule, defined as follows:

$$r_i(h, t') \wedge r_i^{-1}(t', h') \wedge r_i(h', t) \Rightarrow r_i(h, t) \tag{7}$$

where $r_i(h,t)$ denotes the triplet $<h, r_i, t>$ to be evaluated, $r_i(h,t')$, $r_i(h',t')$, and $r_i(h',t)$ are *triplets* existed in $\boldsymbol{G}_i$. Our idea is that, if the inference rule (7) is applicable to a relation r_i with respect to entity pair $<h,t>$, it should be reflected by the *relation pattern* of the facts existed in $\boldsymbol{G}_i$. The more evidence it provides, the more likely the inference results are valid. A 3-hop random-walk model would be sufficient to capture all such evidence in $\boldsymbol{G}_i$, and the transition probability of the random-walker from h to t can be taken as the measure of the strength of the feasibility of $r_i(h,t)$.

In order to compute the transition probability efficiently for all of the entity pairs in $\boldsymbol{G}_i$, we resort to the transition matrix representation of graph $\boldsymbol{G}_i$. Firstly, we construct an adjacent matrix $\mathbf{A}_i$ from $\boldsymbol{G}_i$ (as shown in right part of Fig. 2), in which each row is corresponding to a *head* entity in $\boldsymbol{G}_i$, and each column is corresponding to a *tail* entity in $\boldsymbol{G}_i$. Secondly, we construct two diagonal matrices $\mathbf{D}_h$ and $\mathbf{D}_t$ for the *head* and *tail* entities in $\boldsymbol{G}_i$, respectively. The diagonal elements of $\mathbf{D}_h$ and $\mathbf{D}_t$ are the degrees of the corresponding entities in $\boldsymbol{G}_i$. The 3-hop transition matrix between the *head* and *tail* entities of $\boldsymbol{G}_i$ can be computed as follows:

$$\mathbf{M}_i = (\mathbf{D}_h^{-1}\mathbf{A}_i)(\mathbf{D}_t^{-1}\mathbf{A}_i^T)(\mathbf{D}_h^{-1}\mathbf{A}_i) \tag{8}$$

in which $\mathbf{M}_i[h,t]$ represents the probability of a random-walker started from entity $h \in \boldsymbol{G}_i$, after three moves along the paths in $\boldsymbol{G}_i$, finally appeared in entity $t \in \boldsymbol{G}_i$. Based on previous discussion, the *score of local relevance* of entity pair $<h,t>$ with regard to relation r_i is defined as:

$$g(h, r_i, t) = \mathbf{M}_i[h, t] \tag{9}$$

Given an entity pair $<h,t>$ and a relation r_i, we can compute two relevant scores according to Eq. (3) and (9), in which $f(h, r_i, t)$ is derived from the global structure of the knowledge graph $\boldsymbol{G}$ based on our first assumption. $g(h, r_i, t)$ is derived from the local structure of the relation-specific clique $\boldsymbol{G}_i$ based on the second assumption. Next, we describe how to combine these estimators into a single measure of relevance for entity pairs (with respect to specific r_i).

3.5 Combine The Inference Results

In this section, we describe how to combine the strength of the two inferential systems proposed in previous sections. According to Eq. (3), $f(h, r_i, t)$ can be seen as a linear combination of probabilities (recall that each $x_j \in \mathbf{x}$ represents the probability of the random event that a random walker moves along the *path* π_j from h to t on graph $\boldsymbol{G}$). Since $g(h, r_i, t)$ is also a probability of the same type, so that they are additive. Which naturally leads to the following linear equation for results fusion :

$$score(h, r_i, t) = f(h, r_i, t) + \alpha \cdot g(h, r_i, t) \tag{10}$$

in which $\alpha > 0$ is a weighting factor that indicates the relative importance of the local inference results. Note that Eq. (3) is linear, thus the Eq. (10) can be rewritten as:

$$score(h, r_i, t) = \left(g(h, r_i, t), \mathbf{x}^T\right) \cdot \begin{pmatrix} \alpha \\ \Theta_i \end{pmatrix} \tag{11}$$

Eq. (11) should raise some doubt about the necessity of the local inference process, because under the *undirected graph assumption*, for most of the relation $r_i \in \boldsymbol{R}$ (especially the M:M relations), it is highly possible that the *path feature* of the form "$r_i \rightarrow r_i^{-1} \rightarrow r_i$" has already been included in the *path feature set* Π_i, which is exactly the same as the situation considered the local inference process.

However, as we found by experiments, excluding the local inference results from Eq. (10) will actually decrease the

performance of HiRi (see Section 4.3), which indicates that the effects of the information transitivity within a relation-specific clique should be taken into account individually when modeling the relation from GKBs. For ease of investigation and interpretation, we rewrite the Eq. (10) as:

$$score(h, r_i, t) = \frac{e^{f(h,r_i,t)}}{1 + e^{f(h,r_i,t)}} + \alpha \cdot g(h, r_i, t) \qquad (12)$$

In this equation, the contribution of all the global path features are considered as a whole by using sigmoid function, and the $f(h, r_i, t)$ score is mapped to range $(0, 1)$. Since now the value ranges of the transformed $f(h, r_i, t)$ scores are comparable to the $g(h, r_i, t)$ scores, it will be much easier to investigate the effects of the local inference results to the relational learning model proposed above, and to interpret the relative importance of the local and global inferential process by varying the value of the weighting factor α.

4. EXPERIMENTS AND DISCUSSION

To evaluate the performance of the proposed algorithm, we compared it with four representative algorithms proposed for the relational inference tasks on large-scale GKBs, namely PRA [8], RESCAL [15], TransE[1], and TransR[9].

The PRA algorithm is the most representative random-walk based relational learning algorithm, which has been successfully used in some large-scale GKB projects.

The other candidates belong to the latent factor models family, in which RESCAL is a classical tensor factorization algorithm which can be treated as a benchmark. TransE and TransR are two of the most competitive structured embedding algorithms at this time.

Besides, in order to test the validity of the assumptions involved in the HiRi algorithm, we take the global inference module individually as the **baseline** for comparison with the proposed hierarchical random-walk scheme.

4.1 Data Sets

The evaluation is performed on two data sets extracted from Wordnet and Freebase[1], which were created by A. Bordes et al.[1], and have been frequently used in recent research for performance comparison and evaluation[21, 9]. In order to be in accordance with related works, these data sets will be denoted as WN18 and FB15k in the rest of this section. The statistics of the data sets are summarised in Table 3.

The FB15k data sets are sampled from the Freebase (a practical large-scale knowledge base), which cover the facts from almost all aspects of the physical world. Comparing with WN18 data sets, which are sampled from a dictionary-alike GKB, the knowledge distributions and structures of FB15k are more close to reality and more comparable to other industrial GKB products. For this reason, our discussion will mainly focus on this data set. To further explore the relation structure of FB15k, we manually split all of the relations into four categories according to Eq. (2). Detailed statistics are given in Table 4 and Table 5, respectively.

From Table 3, 4 and 5 one can see that there are 1,345 types of relations contained in FB15k, which are evenly distributed across four categories, however the distribution of the number of triplets (i.e. edges of the knowledge graph) are extremely unbalanced. The triplets with one-to-many and many-to-many relations add up to 90% of the total number of triplets in FB15k, which indicates that effectively dealing with relation types belonging to these categories is critical to the overall performance of the algorithm.

[1]https://everest.hds.utc.fr/doku.php?id=en:transe

Table 3: Statistics of the data sets

Data set	WN18	FB15k
#entities	40,943	14,951
#relation types	18	1,345
#triplets in training set	141,442	483,142
#triplets in validation set	5,000	50,000
#triplets in test set	5,000	59,071

Table 4: Distribution of the relations of FB15k

Categories	1:1	1: M	M:1	M:M
Training set	27.36%	22.97%	29.29%	20.38%
Test set	25.50%	23.27%	28.92%	22.31%

Table 5: Distribution of the facts (triplets) of FB15k

Categories	1:1	1: M	M:1	M:M
Training set	1.57%	9.48%	15.88%	73.07%
Testing set	1.48%	9.54%	15.12%	73.86%

4.2 Experimental Setup

We follow the evaluation protocol used in [1] and [8]. Firstly, for each triplet $<h, r_i, t>$ in the test set, the *head* entity h is replaced by each of the entities in the training set, then we remove from this corrupted triplets set (denoted by C) of all the triplets that appear in the training, validation and test set, except the test triplet of interest. For each of the triplets in C, we compute its relevance score by using the algorithms on trial, after that the relevance scores are sorted by ascending order to form a recommendation list. The rank of the correct fact in this list is denoted by $Rank(?, r_i, t)$.

Secondly, repeat this procedure while this time removing the *tail* entity t instead of h, and record the rank of the correct fact in this new list (of the ordered corrupted triplets) as $Rank(h, r_i, ?)$. The rank of a test triplet $<h, r_i, t>$ reported in this paper is the average of above two rank numbers:

$$Rank(h, r_i, t) = \frac{Rank(?, r_i, t) + Rank(h, r_i, ?)}{2} \qquad (13)$$

Based on this definition, we report three measures of performance in the tests, namely *Hits@1*, *Hits@10*, and the *mean reciprocal rank* (MRR), explained as follows. The Hits@1 score denotes the proportion of correct facts which were ranked first by the algorithms on trial. Similarly, the Hits@10 score is the average proportion of correct facts ranked in the top 10 position. Empirically speaking, the Hits@1 score can be seen as the measure of *predictive accuracy* of the inferential algorithm, while the Hits@10 score reflects the *recall rate* of the algorithms (since in many real-world expert systems, top 10 recommendation is a psychological boundary of the acceptable length of the recommendation list for manual inspections). The MRR score is defined as:

Table 6: Experimental results on WN18 dataset

Algorithms	MRR	Hits@1	Hits@10
HiRi	**0.691**	**79.1%**	90.8%
Baseline	0.667	65.4%	67.9%
PRA	0.458	42.2%	48.1%
Rescal	0.431	10.2%	52.8%
TransE	0.495	11.3%	89.2%
TransR	0.605	33.5%	**91.7%**

Table 7: Experimental results on FB15k dataset

Algorithms	MRR	Hits@1	Hits@10
HiRi	**0.603**	**54.3%**	70.3%
Baseline	0.515	49.7%	54.3%
PRA	0.336	30.3%	39.2%
Rescal	0.354	23.5%	44.1%
TransE	0.463	29.7%	**73.4%**
TransR	0.346	21.8%	65.5%

$$MRR = \frac{1}{N} \sum_{<h,r_i,t> \in \boldsymbol{G}} \frac{1}{Rank(h, r_i, t)} \tag{14}$$

where $\boldsymbol{G}$ refers to the test set, N denotes the number of triplets in $\boldsymbol{G}$. The MRR is a normalized score of range $[0, 1]$, an increase in its value reflects that the majority "hits" will appear higher higher in the ranking order of the recommendation list, which indicates a better performance of the corresponding relational inference algorithm.

4.3 Overall Performance

For experiments with HiRi, we selected the optimal configurations by grid search, the penalty factors λ_1 and λ_2 were set to 0.001 and 0.001 for all of the tests in this section. For all data sets, the weighting factor α is varied from 0.0 to 12.0 with step size 0.1, the best models were selected by early stopping using the MRR score on the validation sets.

The test performance of the different algorithms on both data sets are reported in Table 6 and Table 7 for comparison purposes, in which the best results of each column are highlighted in boldface. We first compare our work with other related works (except the Baseline algorithm). The test results show that HiRi consistently outperforms PRA and those three LFM algorithms in terms of both MRR and Hits@1. However, there are two exceptions when considering the Hits@10 scores. In these cases, HiRi is found to perform comparable to TransR (on WN18) and TransE (on FB15k), respectively, but the difference is neither significant nor consistent (this issue will be further discussed later).

To sum up, the comparison between HiRi and other alternative algorithms indicates that the *recall rate* of HiRi is consistently in accordance with the best performed methods, while its MRR and Hits@1 scores are notably better than alternative methods. Since MRR represents the (inverse) average ranking position of the correct facts in the recommendation list, and Hits@1 means the *first round hit* probability, all these improvements can offer considerable benefits for practical applications and better user experiences.

Next, we further investigate the difference between HiRi and other two RWM algorithms by looking at their respective performance on each data set. Experimental results show that on both data sets, the Baseline algorithm consistently and significantly outperforms PRA. Meantime, it was outperformed consistently and significantly by HiRi. Since the Baseline algorithm can be seen as the *undirected graph* based PRA, a direct conclusion obtained from above results is that the *undirect graph* assumption made in this paper is effective in promoting the performance of the random-walk models, but this is not enough for a successful inference algorithm. The obtained solution can be further improved by imposing the local structure information on the algorithm.

4.4 Fine-Grained Analysis

In this section, we take a closer look at the impacts of the localized inference to the overall performance of HiRi, by tuning the weighting factor α and inspecting respectively the test results on each category of relations. Since there are only 18 relations contained in the WN18, the FB15k data set was used for all the tests in the following discussion.

We first test the impact of α on each of the three performance measures used in this paper, the numeric results are depicted in Figure 4. Simulations are done for α varying in the range $[0, 12]$ with step size 0.1. The dash lines represent the performance of the Baseline algorithm, which is equivalent to setting $\alpha = 0$ in the HiRi algorithm.

As shown in Figure 4(a), 4(b), and 4(c), almost the same behavior can be observed for all of the three measures, an inflection point appears at $\alpha = 0.5$ for all observations. In comparing with the Baseline, the value of MRR, Hits@1, and Hits@10 with respect to the peak point achieves an improvement of approximately 17.09%, 10.06% and 29.47%, respectively. These results reveal that the reasoning ability of the random-walk model can be further enhanced by introducing a localized inferential mechanism into the system. More than that, the numerical results also indicate that in practice, there exists an appreciable amount of relations that can be inferred from their local structures.

Figure 4 also shows that, after reaching the peak value at $\alpha = 0.5$, all of the three performance indicators started to decrease as the weighting factor α increased. Among which the MRR and Hits@1 score seemed to be affected more seriously, and eventually at $\alpha \geq 9.0$, the Hits@1 score even dropped below the Baseline level of performance. This indicates that the results of the global and local inference procedures are complementary to each other, both of them are necessary to the random-walk inference models.

We also notice that in Figure 4, the Hits@10 score decreases slightly with the increase of α, and its value remained above 70% for all $\alpha \in [0.5, 12]$. This is in contrast to the rapid decrease behavior observed in MRR, which indicates that the increment of α will likely result in a substantial downgrading of the ranks of the correct results, in which about 30% of the relations will be significantly affected in FB15k test set. In order to understand the reason of this discrepancy in between the recall rate of different relations, we manually partitioned the 1,345 relations in FB15k test set into four categories, then perform tests on each of them with the HiRi algorithm, the results are depicted in Figure 5. Some conclusions can be drawn from Figure 5 as follows.

Firstly, The MRR scores of HiRi are highly and positively

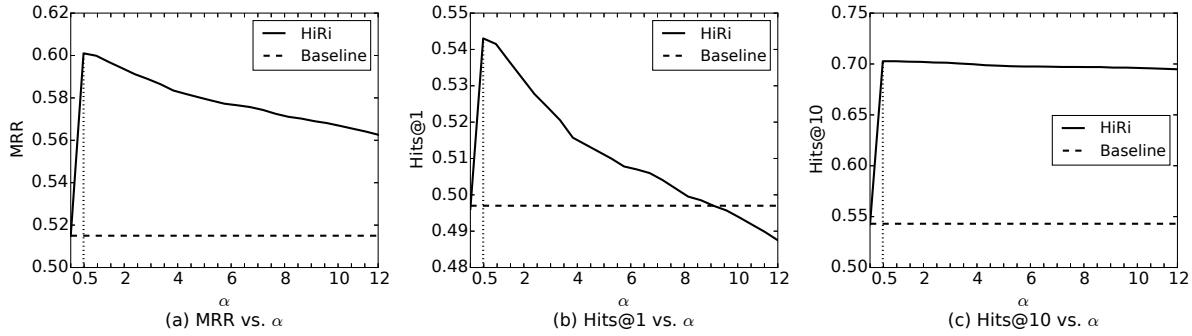

Figure 4: Evaluation of the impacts of α in HiRi on FB15k test set

correlated with its *predictive accuracy* (measured by Hits@1 scores) on all the four category of relations, and both of them declined steadily after the peak value has been reached, which indicates that the inference accuracy of HiRi depends on both of the global and local inference results. However, the Hits@10 recall rate of HiRi are found to be quite insensitive to the variation of α on all of the relation categories after the peak value had been reached, except on M:1 relations it shows a clear downward trend in the recall rate at $\alpha \geq 3.5$. The stability of the Hits@10 scores indicates that the inference results provided by both of the two inference procedures of HiRi are largely consistent with each other, while the variation of MRR and Hits@1 (and Hits@10 on M:1 relations) reflects that there exists some differences between the resulting list of the local and global inference procedures, indicating that both of them provide a one-sided view of the general patterns of the relationships between entities, and can be fine-tuned by varying the value of α to provide better performance than either method alone. These results clearly support our second assumption about the effectiveness of making use of the connection structures in relation-specific cliques to improve and to enhance the performance of the relational retrieval models.

Secondly, each of the performance curves displayed in Figure 5(b) and Figure 5(d) possesses a clear flex point at approximately $\alpha = 0.5$, which suggests that the incorporated local inference procedure is especially helpful in cases of reasoning on 1:M and M:M relations. This also reveals that the global random-walk inference model can not make full use of the information available in these relation-specific cliques. However, as can be seen from the variation of the MRR and Hits@1 scores in Figure 5(a) and Figure 5(c), incorporating a local inference procedure into the HiRi algorithm may only cause a slight decrease in the predictive accuracy. The most plausible explanation of this phenomenon is that the first order Horn clause rule of the specific form "$r_i \rightarrow r_i^{-1} \rightarrow r_i$" might not be suitable for relational inference in such cases.

Thirdly, we notice that both of the MRR and the Hits@1 scores of HiRi are less sensitive to the variance of α on 1:1 and 1:M relations than on other two category of relations, which indicates that the inference results of the global and local inference procedures of HiRi are relatively more consistent with each other on 1:1 and 1:M relations than on other relations. This observation suggests that the reasoning power of random-walk models on 1:1 and 1:M relations is mainly coming from the internal structures of the relation-specific cliques, which can be modeled effectively by use of the proposed local inference method. However, since the interconnections between relation-specific cliques are deliberately neglected by the local inference procedure, it can not capture the inferrable relation sequence patterns other than of the form "$r_i \rightarrow r_i^{-1} \rightarrow r_i$", thus a global inference mechanism is a necessary requirement for capturing such information from the data, this again verifies the validity of the proposed hierarchical random-walk inference scheme.

Finally, as can be seen from Figure 5(b) and Figure 5(c), the performance of HiRi exhibits clear differences with the increase in α on 1:M and M:1 relations. Perhaps the only explanation that can be put forward for this observation is that, the directionality of the relations might play a role in affecting the performance of HiRi, which seems in conflict with our first assumption that the GKBs should be modeled with undirected graphs. In order to justify the validity of this assumption, and to seek a deeper understanding of the reasoning power of the random-walk models, we perform tests on each category of relations on FB15k respectively. Results and discussions are presented in the next section.

4.5 Further Investigation

In this section, we provide more evidence on the validity of our solution by taking into consideration the directionality of the relations in the experiments. The tests are performed on each category of relations respectively with four selected algorithms. Since our HiRi algorithm significantly outperforms other related approaches in terms of Hit@1 and MRR, hence for clarity of presentation in the paper, we only focus our discussion on Hit@10 scores of the tests in this section. The numerical results are reported in Table 8.

Firstly, test results show that the Baseline algorithm performs better on three category of relations (1:1, 1:M, and M:1) compared with PRA. Further, comparing the Baseline algorithm with HiRi, one could see that the differences of their performance are trivial, which suggests that the superior performance of HiRi with regard to PRA is mainly resulting from our first assumption. It is worth noting that the most significant improvement was found in dealing with 1:M relations, both of the *undirected* RWM algorithms outperform PRA with improvements of approximately 200%, which provides a solid evidence for our first assumption.

Secondly, by comparing the performance of HiRi with PRA and Baseline algorithms on M:M relations in Table 8, it is clear to see that through adopting the *undirected graph* as-

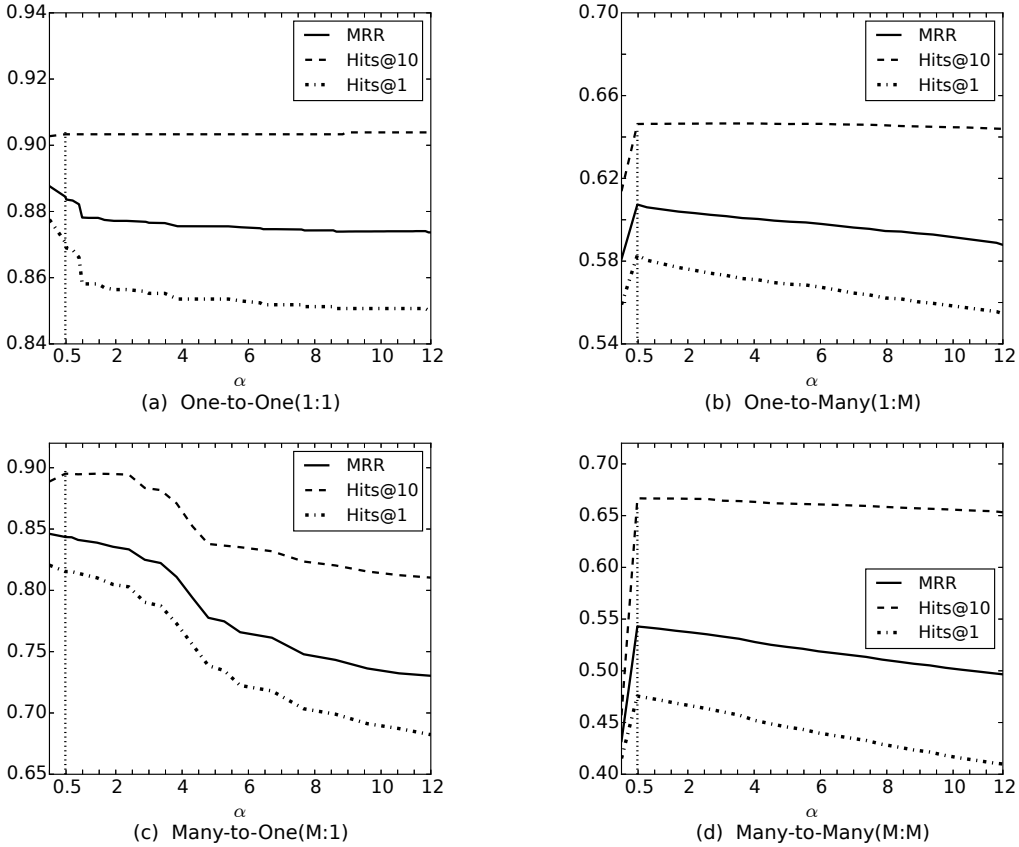

Figure 5: Evaluation of the impacts of α on Hits@10 of HiRi, on four category of relations in FB15k test set

sumption, the Hits@10 score was improved by 40.49% (Baseline vs. PRA), while after incorporating the local inference results into the inference model, the Hits@10 score was improved again by 54.15% (HiRi vs. Baseline). This helps in verifying the validity of our second assumption, which claims that some type of relations are more inferrable than other relations, such information can be modeled and learned from the structure of the relation-specific cliques, to enhance the inference ability of the random-walk based models.

Lastly, Table 8 shows that HiRi outperforms TranseE on three of the four category of relations, except on M:M relations (with a 3.26% discrepancy), combining with the observation that the local inference mechanism only helps in improving the recall rate of HiRi on M:M relations, which suggests that the major advantage of the latent factor models is that they can make full use of the structure information of GKBs by using of the matrix factorization techniques, while such information can not be utilized sufficiently in PRA through learning from the *path features* extracted from the global knowledge graph. However, on the other hand, this "advantage" can also become a disadvantage of the latent factor model, in that it tends to fit the data too closely, and thus resulted in a loss of generalization ability. In fact, this may seriously affect the accuracy of the inference results provided by such approaches, especially in dealing with incomplete and unbalanced knowledge graphs. In contrast, the random-walk models are less affected by this problem, because the focus of RWMs is to discover the inferrable relation patterns to build inference rules. We also notice that the parameter estimation process of the RWMs are based on supervised learning techniques, which means that their generalization ability will largely depend on the learning model and the training data.

However, it is still reasonable to expect that the RWM solutions are more flexible than the LFM solutions, in that they are not restricted to *perfectly fit* of all of the facts existed in the training data (as required in LFMs). As can be seen from Table 8, the generalization ability of HiRi is confirmed by the fact that it achieves better *recall rate* than TransE on most of the relations (more than 78%, according to Table 4). Combining with the results reported in Table 7, we could conclude that by adopting the undirected graph assumption, and by introducing a local inference mechanism into the model, the random-walk models can be superior to the latent factor models in terms of both the accuracy and the recall rate, which makes it a promising candidate for further investigation and application.

5. CONCLUSIONS

We propose in this paper a novel hierarchical random-walk inference algorithm (HiRi) based on two assumptions drawn from our empirical studies. There were two important findings in this study. First, we found that since the relations

Table 8: Evaluation of Hits@10 on FB15k test set

Categories	1:1	1: M	M:1	M:M
HiRi	**89.5%**	**60.5%**	**92.3%**	70.6%
Baseline	89.5%	60.0%	91.4%	45.8%
PRA	63.3%	20.4%	81.4%	32.6%
TransE	71.5%	49.0%	85.0%	**72.9%**

are semantically bidirectional, the undirected graph representation of the knowledge graph can be effectively used for learning the first order Horn clause rules from the GKBs, which we believe to be essential for better understanding the existing methods and for designing new models. Second, we found that the path feature of the form "$r \rightarrow r^{-1} \rightarrow r$" is of special importance to relational learning for some specific type of relations, and we work out a simple but very effective solution for utilizing such information in a RWM framework. Regarding future work, our plan is to further explore the performance of the HiRi algorithm on more large-scale data sets, to examine the practical consistency of the proposed inference model, and to further improve the algorithm performance, then make it publicly available to the community.

6. ACKNOWLEDGMENTS

This work was supported by NSFC under grant 61133016, 61502087, and U1401257, and by the Fundamental research funds for the central universities under grant ZYGX2014J066.

7. REFERENCES

[1] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Proc. of the 27th NIPS*, pages 2787–2795, 2013.

[2] A. Bordes, J. Weston, R. Collobert, and Y. Bengio. Learning structured embeddings of knowledge bases. In *Proc. of the 25th AAAI Conference on Artificial Intelligence*, pages 301–306, 2011.

[3] J. Cheng, T. Yuan, J. Wang, and H. Lu. Group latent factor model for recommendation with multiple user behaviors. In *Proc. of the 37th ACM SIGIR*, pages 995–998, 2014.

[4] M. Gardner, P. Talukdar, J. Krishnamurthy, and T. Mitchell. Incorporating vector space similarity in random walk inference over knowledge bases. In *Proc. of the 2014 EMNLP*, pages 397–406, 2014.

[5] L. Getoor and L. Mihalkova. Learning statistical models from relational data. In *Proc. of the 2011 ACM SIGMOD*, pages 1195–1198, 2011.

[6] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. The MIT Press, 2007.

[7] N. Lao and W. W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, 81(1):53–67, 2010.

[8] N. Lao, T. Mitchell, and W. W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Proc. of the 2011 EMNLP*, pages 529–539, 2011.

[9] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proc. of the 29th AAAI Conference on Artificial Intelligence*, pages 2181–2187, 2015.

[10] T. M. Mitchell, W. W. Cohen, E. R. H. Jr., P. P. Talukdar, J. Betteridge, A. Carlson, B. D. Mishra, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. A. Platanios, A. Ritter, M. Samadi, B. Settles, R. C. Wang, D. T. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proc. of the 29th AAAI Conference on Artificial Intelligence*, pages 2302–2310, 2015.

[11] S. Natarajan, T. Khot, K. Kersting, B. Gutmann, and J. Shavlik. Gradient-based boosting for statistical relational learning: The relational dependency network case. *Machine Learning*, 86(1):25–56, 2011.

[12] A. Neelakantan, B. Roth, and A. McCallum. Compositional vector space models for knowledge base inference. In *Proc. of the 53rd ACL-IJCNLP 2015*, pages 156–166, 2015.

[13] T. V. Nguyen, A. Karatzoglou, and L. Baltrunas. Gaussian process factorization machines for context aware recommendations. In *Proc. of the 37th ACM SIGIR*, pages 63–72, 2014.

[14] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

[15] M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In *Proc. of the 28th ICML*, pages 809–816, 2011.

[16] M. Nickel, V. Tresp, and H.-P. Kriegel. Factorizing yago: Scalable machine learning for linked data. In *Proc. of the 21st International Conference on World Wide Web*, pages 271–280, 2012.

[17] F. Niu, C. Zhang, C. Re, and J. Shavlik. Scaling inference for markov logic via dual decomposition. In *Proc. of the 12th IEEE ICDM*, pages 1032–1037, 2012.

[18] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1):107–136, 2006.

[19] S. Schoenmackers, O. Etzioni, D. S. Weld, and J. Davis. Learning first-order horn clauses from web text. In *Proc. of the EMNLP*, pages 1088–1098, 2010.

[20] C. Wang, Y. Song, A. El-Kishky, D. Roth, M. Zhang, and J. Han. Incorporating world knowledge to document clustering via heterogeneous information networks. In *Proc. of the 21th ACM SIGKDD*, pages 1215–1224, 2015.

[21] Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge graph embedding by translating on hyperplanes. In *Proc. of the 28th AAAI Conference on Artificial Intelligence*, pages 1112–1119, 2014.