symmetry

MDPI

Article

ELPKG: A High-Accuracy Link Prediction Approach for Knowledge Graph Completion

Jiangtao Ma [1,2], Yaqiong Qiao [3], Guangwu Hu [4,*], Yanjun Wang [1,3], Chaoqin Zhang [1,2], Yongzhong Huang [5], Arun Kumar Sangaiah [6], Huaiguang Wu [1], Hongpo Zhang [3,7] and Kai Ren [7]

[1] School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[2] National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China
[3] State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China
[4] School of Computer Science, Shenzhen Institute of Information Technology, Shenzhen 518172, China
[5] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
[6] School of Computer Science and Engineering, VIT University, Vellore 632014, India
[7] Cooperative Innovation Center of Internet Healthcare, Zhengzhou University, Zhengzhou 450000, China
* Correspondence: hugw@sziit.edu.cn; Tel.: +86-755-89226342

Received: 31 July 2019; Accepted: 27 August 2019; Published: 2 September 2019

check for updates

Abstract: Link prediction in knowledge graph is the task of utilizing the existing relations to infer new relations so as to build a more complete knowledge graph. The inferred new relations plus original knowledge graph is the symmetry of completion knowledge graph. Previous research on link predication only focuses on path or semantic-based features, which can hardly have a full insight of features between entities and may result in a certain ratio of false inference results. To improve the accuracy of link predication, we propose a novel approach named Entity Link Prediction for Knowledge Graph (ELPKG), which can achieve a high accuracy on large-scale knowledge graphs while keeping desirable efficiency. ELPKG first combines path and semantic-based features together to represent the relationships between entities. Then it adopts a probabilistic soft logic-based reasoning method that effectively solves the problem of non-deterministic knowledge reasoning. Finally, the relation between entities is completed based on the entity link prediction algorithm. Extensive experiments on real dataset show that ELPKG outperforms baseline methods on hits@1, hits@10, and MRR.

Keywords: relation completion; knowledge graph completion; link prediction; probabilistic soft logic

1. Introduction

Knowledge graph (KG) refers to a network that contains specific topic related entities (i.e., nodes) and the related information (i.e., relation or predicate) between entities. Then a fact in a knowledge graph can be represented by the tuple relationship <entity1, predicate, entity2>. Note that the entity in KG could be a specific object or an abstract concept, such as people, organization, dataset, and related documentation. As a promising artificial intelligence technique, KG has been widely adopted in many scenarios, e.g., question answering [1], recommendation system [2], co-reference resolution [3], information retrieval [4], and cross-language plagiarism detection [5]. However, the big issue in KG is that some link information is incomplete. For example, in the Google knowledge vault project [6], 71% of the personal information lacks "place of birth", while 75% lacks "nationality" information in Freebase [7].

In order to complement the missing information between entities in KG, the knowledge graph completion (KGC) solution utilizes the existing knowledge to infer latent ones. In other words, KGC uses the existing facts to predict potential relations between entities in knowledge graphs. KGC utilize the symmetry of complete knowledge graph and real knowledge graph to inference the relations between entities. To some degree, KGC is similar to link prediction in complex networks. However, it is much more complex than that because it not only predicts possible link relationships between nodes, but also infers the diversified information contained in these link relations, i.e., tags and other property information. This process of complementing links contained in KG is also known as the KGC link prediction procedure, which uses relational links and semantic reasoning to enable knowledge graph inference and complement missing knowledge. Thus, KGC can generate new facts based on the existing ones through edge number increase and graph enlarging.

The existing methods to solve the link prediction problem can be divided into three types: Tensor decomposition-based methods [8–12] entity vector embedding representation [13–17], and path-based reasoning methods [18,19] showing their merits from the angle of multidimensional array decomposition, semantic relationship reasoning, and path-based relationship reasoning. Li et al. [11] found that in many practical applications, users only care about a part of the domain data, and the domain-specific knowledge in the context of the knowledge graph is highly correlated, so they propose a personalized tensor decomposition method (personalized tensor factorization). This approach allows the user to customize specific domain knowledge to speed up the CP process. Duan et al. [12] proposed an integrated matrix decomposition method to divide the traditional link prediction problem into several small problems, so as to quickly complete the link prediction problem. TransE [13] is a pioneer in the embedding-based method, which represents each link prediction as a translation vector from the subject to the object. Although TransE achieves good accuracy, it has poor interpretability and is more difficult to implement parallel computing. TransR [14] and TransH [15] are extensions of TransE that can handle more complex relational data at the expense of efficiency. Lao et al. [18] proposed a path ranking algorithm (PRA) reasoning method-based on random walk. PRA involves learning specific relation path features and classifying them using the logistic regression method. However, this method of relationship-based co-occurrence statistics faces serious data sparseness problems. To address this problem, Neelakantan et al. [19] designed a knowledge-based completion model that supports any path length reasoning using a distributed vector space model. They use the semantic vector of the binary relation of the path to represent the distributed vector of the constructed entity and reason in the vector space. This method can be extended to data that does not appear in the training set, and can predict relations that are not present in the supervised training set. Such methods have good interpretability, and can automatically discover rules from data. The accuracy is often higher than the embedding-based method, but it is difficult to deal with sparse data. It is ineffective when dealing with low-connectivity graphs. And the efficiency of feature extraction is also low. Although these three kinds of methods have merits to address the knowledge graph completion problem, their results involve a certain ratio of false positive/negative inference or unpredicted information in different levels, since these methods only take one single factor into consideration. Actually, one-dimensional features cannot represent the entities' whole relations, resulting in that link prediction accuracy cannot satisfy real demands.

To improve the link prediction accuracy of KGC, in this article, we propose a novel approach named Entity Link Prediction for Knowledge Graph (ELPKG), which can offer a high accuracy link predication rate while keeping efficiency. ELPKG firstly combines the path-based entity relation representation method and the vector space-based embedding representation method together, so as to solve the relation representation issue between entities. After that, ELPKG utilizes the probabilistic soft logic method to conduct knowledge reasoning and solve the knowledge conflict or the inconsistency issue. Finally, ELPKG completes the relation between KG entities based on a novel link prediction algorithm. Compared with the existing work, the main contributions of our scheme are:

(1) We propose a novel model ELPKG that employs both semantic relation and path relation to complete knowledge graphs. Based on entity vectors and path features, ELPKG invents a novel link prediction algorithm to complete knowledge graphs. This algorithm first trains the triple relationship the fact represented with entity vector data, and then it finds the path between nodes through the breadth-first search method. To the best of our knowledge, this is the first attempt to predict link relationship for completing knowledge graphs.

(2) To achieve high accuracy during KGC, ELPKG adopts a reasoning mechanism based on probabilistic soft logic, which effectively solves the problem of non-deterministic knowledge reasoning and improves the effect and efficiency of reasoning on large-scale KG.

(3) We conduct a large number of experiments on the real dataset YAGO [20] and NELL [21]. It shows that our approach achieves a significant improvement in prediction accuracy, which are 35%, 24%, and 17% higher than the baseline method on hits@1, hits@10 and MRR on the YAGO dataset, and 34%, 21%, 16% on the NELL dataset, respectively.

The rest of this article is organized as follows: we first formulate the KGC link prediction problem and share our insights in Section 2. Then in Section 3, we elaborate our approach in detail, and Section 4 evaluates our scheme and compares it with some classic schemes with different datasets. Finally, we conclude the whole article and discuss our further research plan in the last section.

2. Problem Statement

To better describe the KGC problem, we first try to formulate it to present our insight. As Figure 1 illustrated, taking the simple movie knowledge graph as an example, KG can denote a directed graph G (V, E, T), where V is the set of entity nodes, E is the collection of directed edges between nodes, and T is the set of tuples between entities. In this graph, one node denotes an entity and a directed edge links two entities. Each relation links to a fact that describes the relationship between two entities. Thus, two nodes and their links could form a piece of knowledge. The tuple <subject, relation, object> in the knowledge graph is abbreviated as <*s*, *r*, *o*>, where subject *s* and object *o* are entities, and relation *r* is the relationship between entities. In a knowledge graph, if it contains the tuple <*s*, *r*, *o*>, it may also contain <s', *r*, *o*>, <*s*, r', *o*>, <*s*, *r*, o'>. As shown in Figure 1, <Leonard Nimoy, starred, Star Trek> denotes Leonard Nimoy starred in the role of Star Trek, <Alec Guinness, starred, Star Wars> denotes Alec Guinness starred in the role of Star Wars, and <Star Trek, Science Fiction, Star Wars> denotes that Star Trek and Star Wars both belong to the Science Fiction *genre* of film. So, there will be an edge between Leonard Nimoy and Alec Guinness <Leonard Nimoy, acted in the same *genre* of movie, Alec Guinness>, i.e., they are all starred in the same *genre* of movies.

The link prediction in KG is to inference the potential edges (i.e., the potential relations between entities) in the existing knowledge graph, so as to complete the knowledge graph. Given a knowledge graph KG contains a collection of triples T = {<*s*, *r*, *o*>}, each triple is composed of two entities $s, o \in \boldsymbol{E}$ (entity set) and relation $r \in R$ (relation set). Given a relation *r*, through entity pair <*s*, *o*> such that <*s*, *r*, *o*> $\notin T$, we can predict <s', *r*, o'>$\in T'$, where T' is the triple of the knowledge graph after completion.

$$L = \sum_{(s,r,o)\in T} \sum_{(s',r,o')\in T'} [d(s+r,o) - d(s'+r,o')] \tag{1}$$

where $d(s+r,o)$ denotes the distance between ($s+r$) and *o* and the optimization of MinL is our research goal. To calculate the similarity of tuples <*s*, *r*, *o*> and <s', *r*, o'>, the Euclidean distance is employed to measure the distance of two tuples in the vector space, i.e., the shorter the two tuples, the more similar they are. Thus, we can predict whether there is a relation *r* between s' and o'.

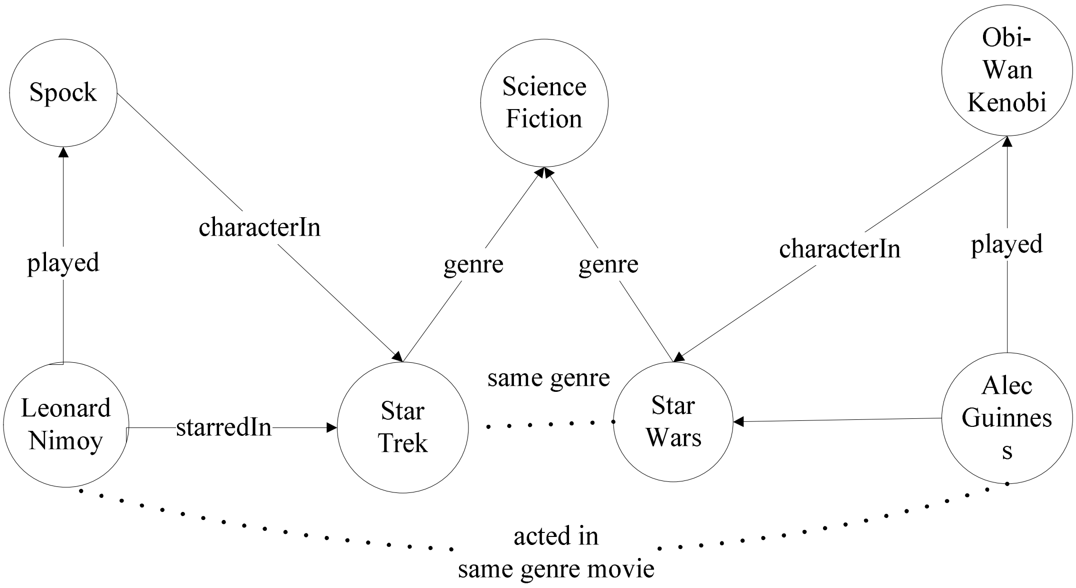

Figure 1. A simple movie knowledge graph adapted from [22]. The entity nodes represent movie or artist objects, the directed edges represent relations between nodes, and the edge labels explain types of relations.

3. Solution Description

To achieve entity link prediction for knowledge graphs, ELPKG firstly combines the vector space-based embedding representation methods and path-based entity relation representation methods to solve the relation representation issue between the entities in a knowledge graph. After that, it uses the probabilistic soft logic method to conduct knowledge reasoning in the knowledge graph to solve the knowledge conflict or the inconsistency problem. Finally, it completes the relation between entities based on the link prediction algorithm between entities in the knowledge graph.

3.1. Entity Relation Representation Based on Vector Embedding

The semantic relation between entities in knowledge graphs can be represented by a word embedding vector space model. Therefore, vector space representations of entities and relations can be learned in the knowledge base, and these representations can be used to predict missing facts. Many methods based on vector space representation treat knowledge base as a tensor and reconstruct it by explicit tensor decomposition. The neural network-based method obtains the vector space representation from the network and uses this to train model. ELKPK considers the tensor K of the knowledge graph modeling as the product of the entity matrix, the relation matrix, and the entity matrix transpose:

$$K_r \approx VR_rV^T \tag{2}$$

where Kr is the $E{\times}E$ knowledge graph tensor corresponding to relation r, V is the $E \times d$ matrix, each entity vector's length is d, and Rr is a $d \times d$ matrix containing a latent relation r. The parameters in V and Rr are optimized according to Equation (3), and the optimization results are used to reconstruct the initial tensor:

$$\min_{V,R} \frac{1}{2}(\sum_r \|K_r - VR_rV^T\|_F^2) + \lambda_V\|V\|_F^2 + \lambda_R\sum_r \|R_r\|_F^2 \tag{3}$$

It can be seen that the semantic association relation between entities can be inferred by using the embedding vector method, and the reasoning method of path relation in Section 3.2 can be used to infer entity relations more accurately.

3.2. Path-Based Entity Relation Representation

Lao and Cohen proposed the path ranking algorithm (PRA) [18] for knowledge-based reasoning. PRA uses the structure features of the knowledge graph to perform random walks to find the sequence (or path) of the relation and predict new entity relations. PRA uses these paths as features of a logistic regression model to reason about missing relations in the knowledge graph. This method uses the structure between entity relations as a feature to predict new relations, which has a stronger representation than other inference methods. Another advantage of this method is that the random walk method can be easily extended to large graphs, which facilitates real-time inference requirements, and the number of steps of random walks can be adjusted according to the range of inference time. The feature vector of PRA is large and sparse. It lacks a mechanism to explore the similarity between knowledge graph predicates to effectively reduce the size of feature vectors. The vector space representation model has rich semantic representation ability, so ELPKG uses path feature and vector space representation to represent the relation between entities.

If relation path $P = R_1R_2 \ldots R_l$ is not empty, let $P' = R_1R_2 \ldots R_{l-1}$, and $h_{s,P}(e)$ is a path constrained random walk distribution, which represents the path feature of node e:

$$h_{s,P}(e) = \sum_{e' \in range(P')} h_{s,P'}(e') \cdot P(e|e'; R_l) \tag{4}$$

$P(e|e'; R_l) = \frac{R_l(e',e)}{|R_l(e',\cdot)|}$ represents the probability of node e' random walks to another node along with edge type R_l. $R(e', e)$ represents whether there is an edge between e' and e, and the edge's type is R.

Given path $P_1P_2 \ldots P_n$, each $h_{s,P_i}(e)$ is node e's path feature, the sequence of nodes is ranked through linear model:

$$\theta_1 h_{s,P_1}(e) + \theta_2 h_{s,P_2}(e) + \ldots \theta_n h_{s,P_n}(e) \tag{5}$$

where θ_i ($i \in [1,n]$) is the weight of the path, the order of candidate node e related to the query node s can be given by the following scoring function:

$$score(e; s) = \sum_{P \in P_l} h_{s,P(e)} \theta_P \tag{6}$$

P_l represents a set of relation paths whose path length is less than or equal to l.

Given relation R and the set of node pairs $\{(s_i, t_i)\}$, a training set D = $\{(X_i, r_i)\}$can be constructed, i.e., X_i represents all the path vectors of node pairs (s_i, t_i), $h_{s_i,P_j}(t_i)$ is the jth component of X_i, and r_i represents whether $R(s_i, t_i)$ is true or not. The parameter θ is estimated by maximizing the following objective function:

$$O(\theta) = \sum_i o_i(\theta) - \lambda_1|\theta|_1 - \lambda_2|\theta|_2 \tag{7}$$

Among them, λ_1 and λ_2 are parameters of controlling L_1-regularization and L_2-regularization for feature selection respectively, and $o_i(\theta)$ is an objective function for each instance. Therefore, we use the random walk method to find the paths between entities and then speculates the potential entity relations in the knowledge graph. This method can infer entity relations through path relations in the knowledge graph. However, there is non-deterministic knowledge in the real world, so non-deterministic knowledge often appears in the corresponding knowledge graphs (i.e., the probability of facts is between 0 and 1). In this article, the reasoning method of probabilistic soft logic is used to solve the problem of non-deterministic reasoning issue in the knowledge graph.

3.3. Probabilistic Soft Logic-Based Reasoning Method

In the process of reasoning relationship between entities, probabilistic soft logic is used to solve the problem of knowledge inconsistency and knowledge conflict, which makes link prediction more accurately and realizes knowledge graph completion. Since the knowledge graph is a collection of knowledge constituting an indeterminate data source, a method based on a probabilistic graphical model can be used to represent the knowledge graph. Although the knowledge graph can be reasoned using the Markov logic network, since the logical value of the Markov logic network is only 0 and 1, when there is a probability multiplication relation, the existence of zero probability will obtain a multiplication result of 0. In reality, some facts are non-deterministic, and the probability of existence is [0,1]. Therefore, to accurately describe the uncertain knowledge in the real world, we employ probabilistic soft logic [23] (PSL) to reason them in the knowledge graph, which can be represented by the Lukasiewicz t-norm:

$$I(v_1 \wedge v_2) = \max\{0, I(v_1) + I(v_2) - 1\} \tag{8}$$

$$I(v_1 \vee v_2) = \min\{I(v_1) + I(v_2), 1\} \tag{9}$$

$$I(\neg l_1) = 1 - I(v_1) \tag{10}$$

where $I(v)$ represents the probability that the fact v is true. The distance between each entity relation representation and the true value of knowledge is expressed as:

$$d_r(I) = \max\{0, I(body) - I(head)\} \tag{11}$$

Among them, *body* is the reasoning result, *head* is the corresponding fact, and the probability distribution represented by the Markov random field [24] containing hinge loss is used:

$$P(I) = \frac{1}{Z}\exp[-\sum_{r \in R} w_r (d_r(I))^{p_r}] \tag{12}$$

Among them, $P(I)$ is the probability distribution of knowledge rule I, Z is the normalization constant, r is the reference fact, w_r is the fact's weight, p_r is the distance metric constant, and $d_r(I)$ is the distance between the fact and the inference prediction result. There is a certain distance between each rule and the true value of the fact. PSL is a probabilistic logical framework with efficient reasoning, which can be interpreted as the joint probability distribution of various variables on the knowledge graph. Finding the knowledge graph with the most consistent knowledge on this probabilistic graph model helps to infer the accurate knowledge graph to the greatest extent possible.

3.4. Entity Linking Prediction Algorithm

To predict the potential entity relations, we propose a link predication method named ELPKG, which first obtains the vector space representation between entities by training the triple in the knowledge graph and then finds the path of the node to be queried by the breadth-first search method according to the entity nodes involved in the query. Finally, the relation probability of the predicted entities is inferred by the vector representation and the path representation between the entities using the probabilistic soft logic. The prediction results greater than a certain threshold are put into the set of prediction results. The ELPKG details are shown in Algorithm 1.

Algorithm 1. Entity Linking Prediction in Knowledge Graph (ELPKG)

```
Input:  knowledge graph G = (V,E,T), set of query tuples Q, set of training tuples T = {(s,r,o)}, set of entities E
        and set of relation R.
Output: The prediction set P corresponding to the query tuple set Q.
P ←∅ // Initialize the link prediction set P to empty
△← train({(s,r,o)}); // Train tuple to get the embedding vector of the tuple
while (Q)
  L← searchBFS($q_i$,G);// Use the BFS method to get the path in the knowledge graph
  while (l∈L) do
        t←PSL(l, △,T) //Obtaining entity pairs at both ends of path l through probabilistic soft logic
        P←P∪{ o}// Add entity pairs to the link prediction set
      L←L-l
    end while
  Q← Q.remove($q_i$)
end while
return P
```

4. Evaluation

4.1. Dataset

We extract the character entities, facts, predicates, and tuples from the YAGO (https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads) and NELL (http://rtw.ml.cmu.edu/rtw/resources) datasets as experimental data. Based on these data, we establish the corresponding human relation knowledge graph. The statistics of the datasets used in this experiment are shown in Table 1.

Table 1. Dataset statistics.

Dataset	Number of Entities	Number of Relation Type	Number of Tuples
YAGO	192628	51	192900
NELL	2156462	50	2465372
YAGO-50	192628	50	100774
YAGO-rest	192628	41	92126

4.2. Evaluation Criteria

ELPKG adopts the same evaluation criteria as the baseline method. The ranking of the entity candidate results generated by the inference method is used to evaluate the algorithm. The first N hit rate hits@N and the average reciprocal rank (MRR) are used to evaluate the proposed method.

The first N hit rate hits@N is defined as:

$$\text{hits@N} = \frac{1}{|T|} \sum_{(s,r,o)\in T} \text{ind}(\text{rank}(s,r,o) \le N) \tag{13}$$

where T is the test set, $|T|$ is the number of triples in the test set, and $ind(\cdot)$ is the indicator function defined as:

$$\text{ind}(x) = \begin{cases} 1, \ x = \text{True} \\ 0, \ x = \text{False} \end{cases} \tag{14}$$

hits@N indicates the probability that the correct reasoning result appears in the first N results, which is similar to the recall rate of the knowledge reasoning algorithm. In the experiment, N takes 1 and 10, i.e., the first hit rate hits@1 and the top 10 hit rate hits@10, respectively. MRR is the average sum of the reciprocals of all the facts in the test set. The higher the fact is ranked in the inference result, the larger the MRR value indicates the reasoning results. Therefore, MRR can more comprehensively evaluate the comprehensive effect of the inference algorithm, which is defined as:

$$\mathrm{MRR} = \frac{1}{|T|} \sum_{(s,r,o) \in T} \frac{1}{\mathrm{rank}(s,r,o)} \tag{15}$$

Filtered mean rank (FMR) is the average position of the test sample in the predicted results. We utilize an iterative setup in which samples existed in the training set are removed from the prediction list.

4.3. Comparison Methods

The mainstream knowledge graph completion methods Rescal [8], TransE [13], HolE [16], and PRA [18] are employed as baseline methods. Rescal is a model based on latent relation features. It expresses the entity relations in the triple through the pairwise interaction of latent features. The shared entity representation in Rescal also captures the similarity of entities in the relation domain, i.e., the entities associated with similar relations are similar. TransE represents the semantic relations through latent feature representation. It models them as a translation between two entities and uses the distance between the entities embedding vectors to measure their similarities. HolE is also a knowledge graph link prediction method based on vector space representation. It uses the whole embedding method to learn the combined vector space representation features on the whole knowledge graph. This method is based on the holographic embedding model of entities. HolE uses the loop association method to create the combination of vector space representation. PRA uses a finite-step random walk method to perform link prediction in the knowledge graph, taking path probability as a feature to predict the potential relations between entities.

4.4. Experimental Results

The proposed algorithm is implemented with python, and the proposed solution is conducted on a sever with Intel Xeon E5-2620 V3 CPU, NVIDIA Tesla K80 GPU,128G memory, and Centos 6.4. In the experiment, the scalability of the method is tested by the learning efficiency of new relations. A total of 50 relations are randomly selected from the YAGO data and the data is divided into the YAGO-50 and YAGO-rest datasets, where the former contains tuples of these 50 relation types, and the later contains 41 relation types, ensuring that both datasets contain all entities. YAGO-50 is divided into a training set containing 386,636 tuples and a test set of 62,138 tuples. YAGO-rest contains a training set of 50,000 tuples (1000 tuples per relation) and 42,126 tuples. Groups of test sets use these data to do the following experiments: First, they use the YAGO-rest training set and test set to train and select the model, learning only 50 related parameters on the two sets. Second, they conduct a test of link prediction. Third, they repeat experiments using 10, 100, 1000, and 10,000 tuples for each relation.

Figure 2 shows the comparison of ELPKG with the baseline method on FMR. ELPKG can learn new relations with fewer samples. When the number of new relation tuples is increased from 10 to 100, ELPKG learned fastest. FMR drops rapidly from 6378 to 996 and keeps the best level. Figure 3 shows the comparison between ELPKG and the baseline method on hits@10. ELPKG has the fastest learning speed. When there are only 100 new relations, hits@10 reaches 21%, and the result will increase with the labeled samples. Therefore, ELPKG has good versatility and does not need to modify any trained word embedding vectors

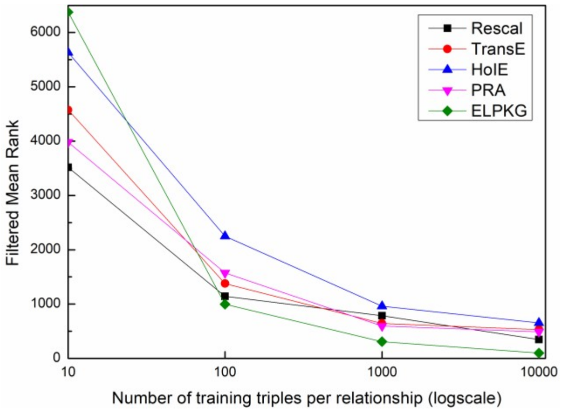

Figure 2. The results of the FMR comparison between ELPKG and the baseline methods.

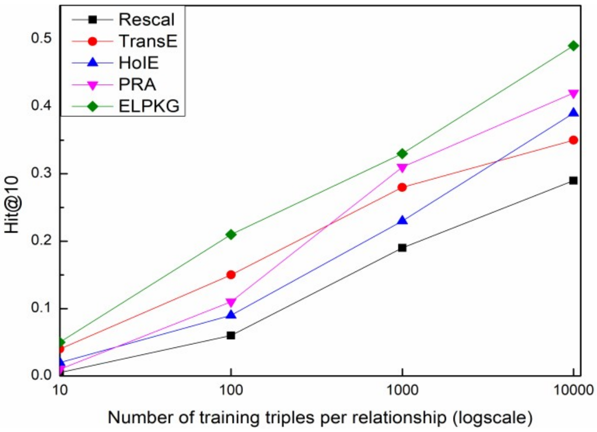

Figure 3. Comparison of ELPKG and baseline methods on hits@10.

Figure 4 shows the comparison of ELPKG with the baseline methods on YAGO. It can be seen that ELPKG achieves the best results. ELPKG is 23%, 14%, and 11% higher than PRA, 46%, 39%, 24% higher than Rescal on hits@1, hits@10, and MRR, respectively. Its average value of hits@1, hits@10, and MRR is 35%, 24%, and 17% higher than the others, respectively.

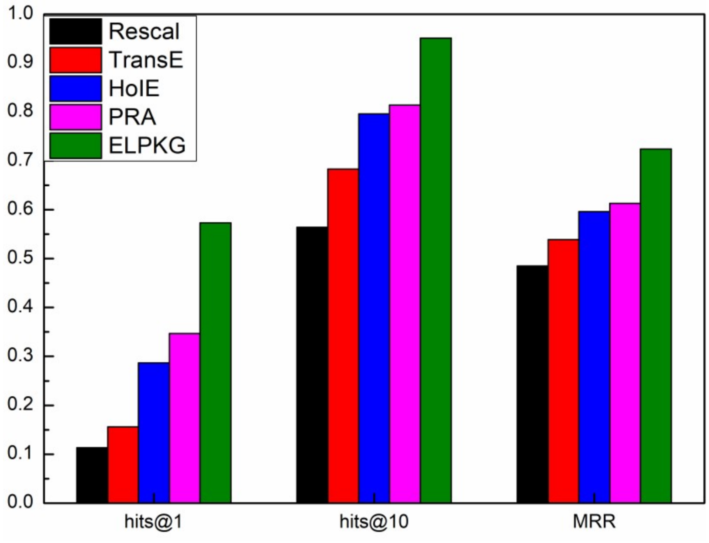

Figure 4. Experimental result comparison of ELPKG with the baseline methods on YAGO.

Figure 5 shows the comparison of ELPKG with the baseline methods on the NELL dataset. ELPKG has achieved the best results in hits@1, hits@10, and MRR. ELPKG is 22%, 11%, and 11% higher than PRA on hits@1, hits@10, and MRR, and it is 45%, 36%, 23% higher than Rescal on hits@1, hits@10, and MRR, respectively. ELPKG is 34%, 21%, and 16% higher than the average of the baseline methods on hits@1, hits@10, and MRR, respectively.

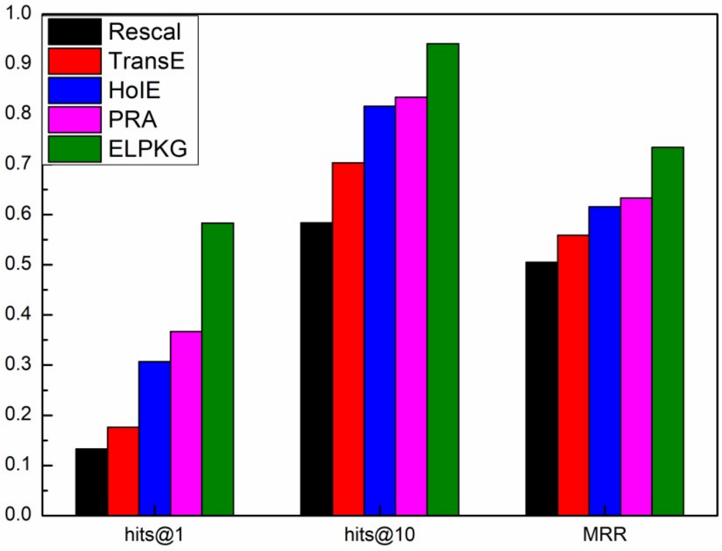

Figure 5. Experimental result comparison of ELPKG and the baseline methods on the NELL dataset.

Figure 6 shows the comparison of ELPKG with the baseline methods on the YAGO-50 dataset. It can be seen that ELPKG has achieved the best results in hits@1, hits@10, and MRR. ELPKG is 21%, 9%, and 10% higher than PRA on hits@1, hits@10, and MRR, and 45%, 34%, 21% higher than Rescal on hits@1, hits@10, and MRR, respectively. ELPKG is 33%, 18%, and 15% higher than the average of the baseline methods on hits@1, hits@10, and MRR, respectively.

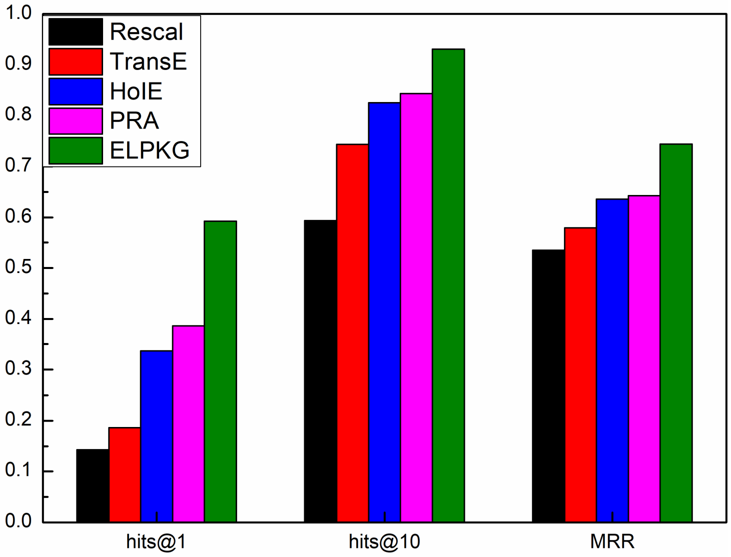

Figure 6. Experimental result comparison of ELPKG and the baseline methods on the YAGO-50 dataset.

Figure 7 shows the comparison of ELPKG with the baseline methods on the YAGO-rest dataset. ELPKG has achieved the best results in hits@1, hits@10, and MRR. ELPKG is 16%, 8%, and 8% higher than PRA on hits@hits@1, hits@hits@10, and MRR, and 46%, 34%, 21% higher than Rescal on hits@1, hits@10, and MRR, respectively. ELPKG is 30%, 18%, and 14% higher than the average of the baseline methods on hits@1, hits@10, and MRR, respectively.

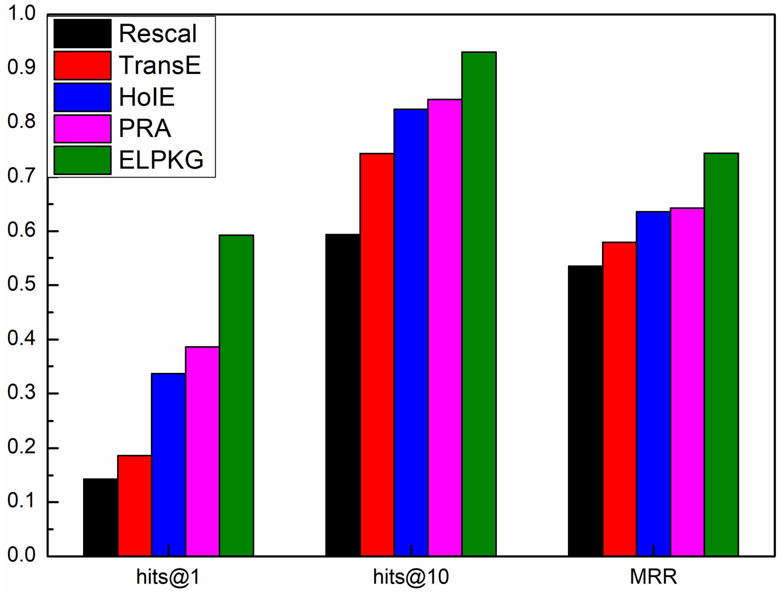

Figure 7. Experimental result comparison of ELPKG and the baseline methods on the YAGO-rest dataset.

The experimental results on the YAGO, NELL, YAGO-50, and YAGO-rest datasets show that ELPKG has significant advantages over the baseline methods in every aspect, which indicates the effectiveness of the proposed method.

Figure 8 shows the comparison of efficiency between Rescal, TransE, HOlE, PRA, and ELPKG on the YAGO and NELL datasets. We can get the conclusion that ELPKG has the highest efficiency. On the YAGO dataset, ELKG's running time is only 10.4% of HOlE, and ELPKG's running time is only 53.2% of PRA. On the NELL dataset, ELPKG's running time is only 12.3% of HOlE, and ELPKG's running time is only 65.6% of PRA.

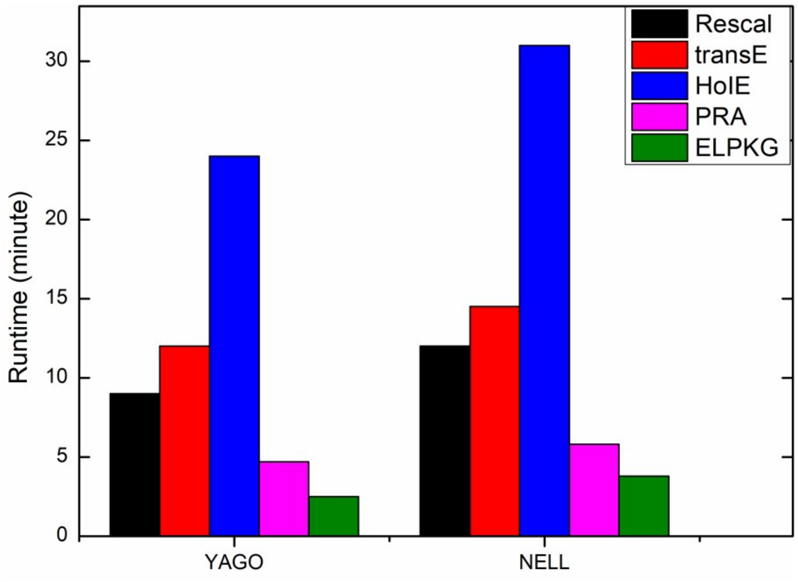

Figure 8. Comparison of running time between ELPKG and baseline methods.

5. Conclusions and Future Work

In this article, we propose a knowledge link prediction method named ELPKG to address the problem of missing relations in the knowledge graph. ELPKG is based on the combination of entity semantics and path structure of the knowledge graph to complement the relation between entities. ELPKG utilizes semantic-based word vector features to embody semantic relations between entities while path features represent graph structure relations between entities. Therefore, it can mine potential relations between entities more accurately. In the process of knowledge graph completion, ELPKG employs probabilistic soft logic to solve the inference issue between non-deterministic knowledge in link prediction, improving the accuracy and efficiency of link prediction. Furthermore, ELPKG does not require external knowledge in the process of entity link prediction. The entity relations can be predicted merely by the existing knowledge in the knowledge graph. Experimental results have shown that ELPKG outperforms baseline methods on YAGO and NELL datasets. In future work, we will extend this method to also solve the problem of knowledge-based question-and-answer and recommendation systems.

Author Contributions: J.M. and Y.Q. conceived and designed the experiments. G.H. and Y.W. performed the experiments. C.Z., K.R. and H.Z. analyzed the data. J.M. and Y.W. wrote the article. G.H., A.K.S. and H.W. reviewed and edited the manuscript. All the authors read and approved the manuscript.

Funding: This research was funded by National Nature Science Foundation of China (No. 61702462, No.61672470, No. 61802352, No.61866008), the Doctoral Research Fund of Zhengzhou University of Light Industry (2017BSJJ046, 2018BSJJ039, 13501050045), Scientific & Technological Project of Henan Province (No.182102210607), the Second Education Fund for Industry and Education project “Digital Science and Technology, Wisdom for the Future”(2018A01094), Foundation of Henan Province Educational Committee (No.17A520064) and Fundamental Research Project of Shenzhen Municipality (No.JCYJ20170817115335418).

Acknowledgments: The authors would like to thank the editors and the reviewers for their comments on an earlier draft of this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zheng, W.; Yu, J.X.; Zou, L.; Cheng, H. Question Answering Over Knowledge Graphs: Question Understanding Via Template Decomposition. *Proc. VLDB Endow.* **2018**, *11*, 1373–1386. [CrossRef]
2. Ayala-Gómez, F.; Daróczy, B.; Benczúr, A.A.; Mathioudakis, M.; Gionis, A. Global citation recommendation using knowledge graphs. *J. Intell. Fuzzy Syst.* **2018**, *34*, 3089–3100. [CrossRef]
3. Dou, J.; Qin, J.; Jin, Z.; Li, Z. Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage. *J. Vis. Lang. Comput.* **2018**, *48*, 19–28. [CrossRef]
4. Sun, M.; Liu, Z.; Xiong, C.; Liu, Z.-H. Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2395–2405.
5. Franco-Salvador, M.; Rosso, P.; Montes-y-Gómez, M. A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Inf. Process. Manag.* **2016**, *52*, 550–570. [CrossRef]
6. Dong, X.; Gabrilovich, E.; Heitz, G.; Horn, W.; Lao, N.; Murphy, K.; Strohmann, T.; Sun, S.; Zhang, W. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM: New York, NY, USA, 2014; pp. 601–610.
7. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 9–12 June 2008; ACM: New York, NY, USA, 2008; pp. 1247–1250.
8. Nickel, M.; Tresp, V.; Kriegel, H.-P. A Three-way Model for Collective Learning on Multi-relational Data. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July 2011; pp. 809–816.

9. Chang, K.-W.; Yih, S.W.; Yang, B.; Meek, C. Typed Tensor Decomposition of Knowledge Bases for Relation Extraction. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1568–1579.
10. Franz, T.; Schultz, A.; Sizov, S.; Staab, S. TripleRank: Ranking Semantic Web Data by Tensor Decomposition. In *Proceedings of the Semantic Web—ISWC 2009: 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, 25–29 October 2009*; Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 213–228. ISBN 978-3-642-04930-9.
11. Li, X.; Huang, S.; Candan, K.S.; Sapino, M.L. Focusing Decomposition Accuracy by Personalizing Tensor Decomposition (PTD). In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management; ACM: New York, NY, USA, 2014; pp. 689–698.
12. Duan, L.; Aggarwal, C.; Ma, S.; Hu, R.; Huai, J. Scaling Up Link Prediction with Ensembles. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining; ACM: New York, NY, USA, 2016; pp. 367–376.
13. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2013; pp. 2787–2795.
14. Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In Proceedings of the Twenty-Ninth AAAI conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 2181–2187.
15. Feng, J. Knowledge Graph Embedding by Translating on Hyperplanes. In Proceedings of the AAAI, Québec City, QC, Canada, 27–31 July 2014.
16. Nickel, M.; Rosasco, L.; Poggio, T. Holographic Embeddings of Knowledge Graphs. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 1955–1961.
17. Socher, R.; Chen, D.; Manning, C.D.; Ng, A. Reasoning with Neural Tensor Networks for Knowledge Base Completion. In *Advances in Neural Information Processing Systems 26*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2013; pp. 926–934.
18. Lao, N.; Mitchell, T.; Cohen, W.W. Random Walk Inference and Learning in a Large Scale Knowledge Base. In Proceedings of the Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 529–539.
19. Das, R.; Neelakantan, A.; Belanger, D.; Mccallum, A. Chains of Reasoning over Entities, Relations, and Text using Recurrent Neural Networks. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; pp. 132–141.
20. Rebele, T.; Suchanek, F.; Hoffart, J.; Biega, J.; Kuzey, E.; Weikum, G. YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. In *Proceedings of the Semantic Web—ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, 17–21 October 2016, Part II*; Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 177–185. ISBN 978-3-319-46547-0.
21. Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; Krishnamurthy, J.; et al. Never-Ending Learning. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15), Austin, TX, USA, 25–30 January 2015.
22. Nickel, M.; Murphy, K.; Tresp, V.; Gabrilovich, E. A Review of Relational Machine Learning for Knowledge Graphs. *Proc. IEEE* **2016**, *104*, 11–33. [CrossRef]
23. Bach, S.H.; Broecheler, M.; Huang, B.; Getoor, L. Hinge-Loss Markov Random Fields and Probabilistic Soft Logic. *J. Mach. Learn. Res.* **2017**, *18*, 1–67.
24. Bach, S.H.; Huang, B.; London, B.; Getoor, L. Hinge-loss Markov Random Fields: Convex Inference for Structured Prediction. In Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, Bellevue, WA, USA, 11–15 August 2013; pp. 32–41.