Entity Alignment Method for Power Data Knowledge Graph of Semantic and Structural Information

IOP Conf. Series: Materials Science and Engineering **569** (2019) 052103 doi:10.1088/1757-899X/569/5/052103

Entity Alignment Method for Power Data Knowledge Graph of Semantic and Structural Information

Wang Zhiqiang[1], Wang Yuan[2,3*], Zhao Kang[2,3], Wang Xin[4] and Hu Hui[5]

[1]State Grid Zhejiang Information & Telecommunication Co., LTD, Hangzhou, Zhejiang, 310020, China

[2]NARI Group Corporation/State Grid Electric Power Research Institute, Nanjing, Jiangsu, 211000, China

[3]China Realtime Database Co., Ltd, Nanjing, Jiangsu, 211000, China

[4]State Grid Anhui Electric Power Co., LTD, Hefei, Anhui, 230061, China

[5]Xi'an Jiaotong University, Xi'an, Shanxi, 710049, China

[*] 406839653@qq.com

Abstract. With the continuous deepening of information construction of State Grid Corporation of China, organizing and utilizing the accumulated mass run data effectively and intelligently has become an urgent problem to solve. Knowledge graph has become an increasingly important hot technology for establishing sematic connection network for power data in full-service unified data centre. Entity alignment is one of the key steps for constructing high-quality power knowledge graph to solve the problem of a large number of entity heterogeneity and redundancy existing between different business systems. This paper proposes an entity alignment method for power data with sematic and structural information with a co-training framework. The semantic and structural models are complemented from the other after they are trained under their perspectives separately. The experiment shows the model achieves satisfactory results with higher accuracy and F1.

1. Introduction

STATE GRID Corporation of China has accumulated a huge amount of power data during the construction smart grid. With the continuous development of big data technology, the potential value of these data is being more and more valued by the corporation and academia. At present, SG Corporation has constructed a full-service unified data centre of electric power system to achieve unified storage and shared use of data, then the Knowledge Graph established in the data centre can establish semantic connections among the data and provide unified semantic-level data service. However, there are lots of heterogeneous and redundant entities and relationships extracted in the knowledge graph because the data in the data centre come from different business systems where the objects are defined and described according to their own business rules. Therefore, entities alignment that can clean and merge entities that point to the same objects and solve the problem of entity reuse has become one of the key processes of building a high-quality power knowledge graph.

Entity alignment technology is aimed at achieving high-quality links between data sources by recognizing those entities in different datasets that point to the same object and linking the entities to a unique one with unified global identity by constructing co-reference links such as *owl:SameAs*. The

inference is usually based on two kinds of methods: traditional and representing learning. The former one mainly depends on attribute similarity.

Based on the string similarity, ontologies and entities can be aligned by matching the literal quantities of the attributes[1-2]. Entity alignment problem is transformed into classification problem and established probability models[3], the weight coefficient is considered in [4] to promote alignment accuracy rate. The method based on attribute information is simple and easy to use and can usually achieves satisfactory results to construct general knowledge graphs, but in the face of heterogeneous power data, it becomes inefficient or even no longer applicable. Sarawagi proposed an active learning method, which requires manual labeling of some entities which are difficult to identify, and then uses artificial feedback to improve the system effect. Then Arasu added filtering operation to reduce the workload of manual data annotation by picking pairs of entities with large amount of information [6]. Song proposed Histsim and DisNGram algorithms for heterogeneous data based on PARIS[7-8], the former uses historical records of entity alignment to calculate entity similarity and the later calculates character set similarity to inference entity alignment. Similarity calculation functions need to be specially designed for different types of attributes, so it causes large amount of manual work. This kind of alignment is also based on attribute similarity, without the consideration of semantic similarity, the effect of entity alignment is limited.

In recent years, deep learning technology has developed rapidly and knowledge representation learning is proposed. It calculates and deduces in a low-dimensional dense vector space where the entities and relationships in knowledge graph are modeled and mapped. TransE[9] is the earliest knowledge representation learning model. It is simple and can achieve satisfactory result on large-scale data sets for one-to-one relationships. Then some extended models are proposed to compute more complex entity relationships on the basis of TransE, such as TransH, TransR and TranSparse, [10-12].

The data of knowledge graph of power grid full-service data comes from different service systems, so the entity alignment task is actually a cross-network relational inference process which is difficult to achieve satisfactory results only based on attribute or semantic information. Cross-KG method can be used for joint learning of two knowledge graphs because of the contribution of complementary information of two data sets in relationship inference[13]. This method works well in the construction of general knowledge graph, but the requirement of large number of labeled alignment data which in practice means a large amount of workload of business experts in power system is unrealistic.

Aiming at this problem, this paper proposes an entity alignment method for power full-service data. The model is trained under a co-training framework. We divide the data into two independent perspectives: semantic and structural. Select the reliable results from one perspective to assist the training of another one. Experiments show that this method can achieve higher accuracy and *F1* value for power full-service data entity alignment.

2. Description of entity alignment for full-service power data

The task of entity alignment for full-service power data is to find record pairs in different service systems that point to the same physical object, that is, the entity alignment between data sets. Firstly, we give the formal definition of knowledge graph and knowledge graph entity alignment.

Knowledge graph. The knowledge graph is composed of triples as the following way: KG=(E，R，F), E=$\{e_1,e_2\ldots,e_{Ne}\}$ represents collection of entities. It includes instance and attribute values. R=$\{r_1,r_2,\ldots,r_{Nr}\}$ represents a set of binary relations. It describes the relationship between entities. $F \subseteq E \times R \times E$ represents fact triple set.

Knowledge graph entity alignment. Supposing KG_1 and KG_2 are two given knowledge graphs, we find out all entities aligned to KG_2 (or KG_1) in KG_1 (or KG_2) respectively. $Align_{entity}$(KG_1，KG_2)={(e，e'|，$e \in E_1, e' \in E_2$)}

3. Entity alignment method with semantic and structural information

3.1. Knowledge graph entity alignment of representation learning

Entity alignment using representation learning is mainly divided into two steps. Firstly, map KG_1 and KG_2 into a low-dimensional vector space and obtain corresponding knowledge representations $\mathbf{KG_1}$ and $\mathbf{KG_2}$. Then train the model to learn alignment relationship φ between entities based on the labeled entity alignment data set N. (e_1,e_2) represents a pair of entities in data set N, which means (e_1,e_2)$\in$N and $e_1 \in E_1$, $e_2 \in E_2$. The alignment of entities in two knowledge graphs can be considered as a special relationship *r*=SameAs* and it can form a triple (e_1,*SameAs*,e_2). Before the training, the embeddings of entities and relationships of KG_1 and KG_2 should be initialized by uniform distribution. Then the loss function of vector space representation is defined as:

$$L=\sum_{(h,r,t)\in\Delta}\sum_{(h',r',t')\in\Delta'}[\varphi(h,t)+\gamma-\varphi(h',t')]_+ \tag{1}$$

$(h,r,t)\in\Delta$ represents the collection of positive triples. It means those existing triples in the real knowledge graph. $(h',r',t')\in\Delta'$ represents the collection of negative triples. It means the triples don't exist in the real knowledge graph. The negative triples are generated by replacing the head or tail vector of positive triples. When constructing negative triples from positive triples with *r*=SameAs* relationship, the replaced head or tail entity should have the same type with the original entity in the data set. In another word, e_1 is replaced by another entity e_1' with the same data type in KG_1 or e_2 is replaced by another entity e_2' with the same data type in KG_2. $\varphi=||h+r-t||$, it measures the degree of matching between two entities by regarding tail entity *t* as the translation process of head entity *h* through *r* according to the idea of translation model. γ is used for separating positive and negative entity pairs, $\gamma>0$

When optimizing the loss function, constraint the relation vector *r*=SameAs* to be zero vector and set the maximum number of iterations of loss function or the threshold of stopping iteration. Each iteration updates the head entity vector, tail entity vector and relation vector. Vector gradient method can be used in updating as follows:

$$\forall\ i\in\{0,1,2\ldots\dim\} \tag{2}$$

$$\mathbf{h}_i=\mathbf{h}_i-\mu*2*|\mathbf{t}_i-\mathbf{h}_i-\mathbf{r}_i|$$
$$\mathbf{r}_i=\mathbf{r}_i-\mu*2*|\mathbf{t}_i-\mathbf{h}_i-\mathbf{r}_i|$$
$$\mathbf{t}_i=\mathbf{t}_i-\mu*2*|\mathbf{t}_i-\mathbf{h}_i-\mathbf{r}_i|$$
$$\mathbf{h}'_i=\mathbf{h}'_i-\mu*2*|\mathbf{t}'_i-\mathbf{h}'_i-\mathbf{r}'_i|$$
$$\mathbf{r}_i'=\mathbf{r}'_i-\mu*2*|\mathbf{t}'_i-\mathbf{h}'_i-\mathbf{r}'_i|$$
$$\mathbf{t}'_i=\mathbf{t}'_i-\mu*2*|\mathbf{t}'_i-\mathbf{h}'_i-\mathbf{r}'_i|$$

dim represents the dimension of space vector. $\boldsymbol{h}_i$ represents i_{st} dimensional vector of $\boldsymbol{h}$. μ represents the learning rate.

It's worth noting that some attributes can be aligned automatically by eliminating the ambiguity of attribute names and some entities with similar semantics are close in the semantic space for example "AC bus" and "AC line"

3.2. Co-training with semantic and structural perspectives

Models of semantic and structural features infer entity alignment relations from their respective perspectives. Both aspects of information are useful for inference in the construction of knowledge graph of full-service power data knowledge graph. Using Co-training framework to complement the inference results from two perspectives can improve the performance of entity alignment.

Specifically, we divide the training data into two independent perspectives, semantic perspective and structural perspective. Generate training data of the semantic perspective according to a small

labeled align data set L_{se}. Then train and obtain the semantic perspective model m_{se}. Predict the entity alignment relation for unlabeled data set X_{se} and pick out reliable entity pairs L'_{se}. Put L'_{se} into labeled data of the structural perspective and obtain a new labeled data set X_{st}. Then repeat a similar process, generate training data of the structural perspective according to X_{st}. Then train and obtain the semantic perspective model m_{st}. Predict the entity alignment relation for unlabeled data set X_{st} and pick out best results L'_{st}. Put L'_{st} into labeled data of the structural perspective and obtain a new labeled data set X_{se}. The two models are iterated until convergence.

Algorithm Co-training with semantic and structural perspectives

Input: KG$_1$ triples T_1={(h，r，t)}
Aligned Data Source 1 Entity Set L_1，Data Source 1 Entity Set to Align U_1
KG$_2$ triples T_2={(h，r，t)}
Aligned Data Source 2 Entity Set L_2，Data Source 2 Entity Set to Align U_2
Labeled aligned entity pairs L={(e_1,*SameAs*,e_2)},$e_1 \in L_1$,$e_2 \in L_2$

Output：Embedding vectors of entities and relationships after training

construct a joint knowledge map $T=T_1 \cup T_2 \cup L$
Divide the training ternary triples into two perspectives，$X_1=T_{se} \cup L$，$X_2=T_{st} \cup L$
Loop for k iterations：
  Train entity alignment model m_1 under the first perspective according to X_1
  Train entity alignment model m_2 under the second perspective according to X_2
  Infer alignment entity pairs by m_1 and pick out reliable entity pairs L'_1
  Infer alignment entity pairs by m_1 and pick out reliable entity pairs L'_1
  $X_1 \leftarrow X_1 \cup L'_2$
  $X_2 \leftarrow X_2 \cup L'_1$
End Loop

3.3. Entity Similarity Computing Based on Attribute Matching

Although the power data may be describe in different formats in different business systems, they have common parts inevitably. Similarity calculation based on common attributes of different source entities can also provide some reference results. *up*(e_1,e_2) is defined as common set of attributes for entities e_1 and e_2,:

$$up(e_1, e_2) = property1 \cap property2 \tag{3}$$

property1 represents the attribute set of e_1. *Property2* represents the attribute set of e_2.

The similarity *sim*(p_i) of the common attribute p_i of two entities can be calculated by:

$$Sim(p_i) = \frac{lcs(v1x, v2y)}{\max(len(v1x), len(v2y))} \tag{4}$$

p_i corresponds to the x_{st} attribute of e_1 and the y_{st} attribute of e_2, *p1x* and *p2y*. the values are *v1x* and *v2y* respectively. *lcs(v1x，v2y)* represents their longest common subsequence of attribute values.

The similarity of e_1 and e_2 is the average of similarity of their common attributes:

$$Sim(e_1, e_2) = average(Sim(p_i)) \tag{5}$$

When inferring the entity h that align to the given entity t^* in another data set, we score all entity relationships (h', *SameAs*, t^*) based on scoring function and choose the highest marked one as result. The scoring function is defined with vector representation similarity and attribute similarity.

$$f_{predict}(\mathbf{h}, \mathbf{r}, \mathbf{t}^*) = (1 + w \times Sim(\mathbf{h}, \mathbf{t}^*)) \| \mathbf{h} - \mathbf{t}^* \| \tag{6}$$

$\|\boldsymbol{h}\text{-}\boldsymbol{t}^*\|$ measures vector representation. $Sim(\mathbf{h}, \mathbf{t}^*)$ represents attribute similarity. w is penalty coefficient ranging from 0 to 1. The value is determined by the reliability of attributes of data set.

4. experiment and result analysis

The experimental data set comes from full-service unified data centre in SG Zhejiang Electric Power Corporation. The running data of a region for a week is extracted from the buffer layer where the data of all the business systems are access directly and stored. There are 1,032 equipment entities recorded in Operation and Maintenance System and 876 in Material System. Altogether there are 1,448 different equipment entities recorded in the two systems. So, 460 equipment entities can be aligned. We select 160 entities as a training set, 150 as a validation set and 150 as a test set. There are altogether 178 supplier entities in the Marketing System and Material System. Among them, 104 are involved in the marketing system and 161 are involved in the material system. So, 87 supplier entities can be aligned. We select 47 entities as a training set, 20 as a validation set and 20 as a test set.

The experiment compares entity alignment method with attribute similarity LCS and the method with knowledge representation learning cross-KG and SEEA. The evaluation index are precision *P*, Recall *R* and *F1*.

Accuracy reflects the rate of extraction results. It is defined as:

$$P = N_{success} / N_{total}$$

N_{total} represents the total number of relationships inferred. $N_{success}$ represents the total number of correct inferences.

Recall rate reflects the proportion of correct inferences to all existing alignment relationships. It is defined as:

$$R = R_{success} / R_{total}$$

$R_{success}$ represents the correct inferences. R_{total} represents the number of all existing alignment relationships.

F1 is an evaluation indicator of comprehensive accuracy and recall rate. It can reflect the overall effect.

$$F1 = 2 \cdot R \cdot P / (R + P)$$

The experimental results are shown in Table 1.

Table 1 The methods and experimental comparison involved in this paper

method	**Alignment rate for equipment entities(%)**			**Alignment rate for supplier entities(%)**		
	P	*R*	*F1*	*P*	*R*	*F1*
LCS	80.37	86.67	83.40	77.78	70.00	73.69
Cross KG	92.74	90.00	91.35	80.00	80.00	80.00
SEEA	90.66	89.33	89.99	78.95	75.00	76.92
this paper	98.68	97.33	98.00	95.00	95.00	95.00

The results of the experiment show that the method of entity alignment with semantic and structural information has achieved ideal results in entity alignment tasks of full-service power data. There is a comprehensive enhance in precision, recall and F1 compared to method with semantic or structural information. The reason for this is that both the semantic information and structural information are effective in entity alignment for power data. On one side, some physical concepts in the electrical industry are very similar in name but quite different in nature. They can be easily distinguished from their attribute structure. On the other side, entities pointing to the same areas or in hierarchical relationships may have common attributes, so they are easier to be distinguished from semantic perspective.

5. conclusion

The experiment shows that the method of entity alignment with semantic and structural information can work well in entity alignment for power data. The models are divided into two perspectives: semantics and structure. Training the two models separately from their perspectives, then select the best results from one perspective to supply another and iterate. Finally, the model is promoted with higher P, R and F1. In the construction of a knowledge graph in a specific field, both the semantic and structural information are important in entity alignment. Fully use of these two aspects of information

can theoretically achieve satisfactory effect. Therefore, this method has strong generality for the construction of a industry knowledge graph.

Acknowledgements
This work was financially supported by State Grid science and technology project: Research on key technologies and application of unified data model of power grid based on Knowledge Graph. (Project number: 5211XT180045) fund.

references

[1] Lacoste-Julien S, Palla K, Davies A, Kasneci G, Graepel T, Ghahramani Z. (2012). Sigma: simple greedy matching for aligning large knowledge bases. In proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago. pp. 572-580.
[2] Sun Y, Ma L, Wang S. (2015) A comparative evaluation of string similarity metrics for ontology alignment. J. Journal of Information & Computational Science. 12(3):957-964.
[3] Newcombe H B, Kennedy J M, Axford S J, et al. (1959) Automatic linkage of vital records. J. Science. 130(3381):954-959.
[4] Herzog, Thomas N, Scheuren Fritz, J, Winkler, William E. (2007). Data quality and record linkage techniques. Springer Science & Business Media. Berlin.
[5] Sarawagi S, Bhamidipaty A. (2002) Interactive Deduplication using Active Learning. In proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[6] Arasu A, Gotz M, Kaushik R. (2010) On active learning of record matching packages. In: ACM SIGMOD International Conference on Management of Data. Indianapolis. pp. 783.
[7] Suchanek F M, Abiteboul S, Senellart P. (2011). PARIS: probabilistic alignment of relations, instances, and schema. J. Vldb Endowment. 5(3), 157-168.
[8] Song D, Luo Y, Heflin J. (2016). Linking heterogeneous data in the semantic web using scalable and domain-independent candidate selection. J. IEEE Transactions on Knowledge & Data Engineering, 29(1), 143-156.
[9] Bordes A, Usunier N, Garicia-Duran A, et al. (2013) Translating Embeddings for Modeling Multi-Relational Data. In proceedings of the 26th International Conference on Neural Information Proceeding Systems. Stateline. pp. 2787-2795.
[10] Wang Z, Zhang J, Feng J, Chen Z. (2014) Knowledge Graph Embedding by Translating on Hyperplanes. J. AAAI. (16):1112-1119.
[11] Lin Y, Liu Z, Sun M, et al. (2017) Learning Entity and Relation Embeddings for Knowledge Graph Completion. J. Procedia Computer Science. 108:345-354.
[12] Ji G, Liu K, He S, Zhao J. (2016) Knowledge Graph Completion with Adaptive Sparse Transfer Matrix. J. AAAI. 985-991.
[13] Cai P, Li W, Feng Y, et al. (2017) Learning Knowledge Representation Across Knowledge Graphs. In AAAI 2017 Workshop on Knowledge-Based Techniques for Problem Solving and Reasoning. San Francisco.