Content-Structural Relation Inference in Knowledge Base

Zeya Zhao[1,2], Yantao Jia[1], Yuanzhuo Wang[1], Xueqi Cheng[1]

[1]CAS Key lab of Network Data Science & Technology, Institute of Computing Technology , CAS, Beijing, 100190, China

[2]National Digital Switching System Engineering and Technological Research Center, Zhengzhou, P. R. China

zhaozeya@software.ict.ac.cn, jiayantao@ict.ac.cn, wangyuanzhuo@ict.ac.cn, cxq@ict.ac.cn

Abstract

Relation inference between concepts in knowledge base has been extensively studied in recent years. Previous methods mostly apply the relations in the knowledge base, without fully utilizing the contents, i.e., the attributes of concepts in knowledge base. In this paper, we propose a content-structural relation inference method (CSRI) which integrates the content and structural information between concepts for relation inference. Experiments on data sets show that CSRI obtains 15% improvement compared with the state-of-the-art methods.

Introduction

Several large scale Knowledge Bases (KBs) have developed in recent years, such as Freebase and YAGO. However, none of the KBs has a satisfactory coverage of knowledge and relation inference is an indispensable task to extend the coverage. In fact, relation inference has been paid much attention due to the successful applications in many areas such as semantic search, machine question and answering, description of logic concepts (Minervini and d'Amato 2012), and learning on the Semantic Web (Huang and Tresp, 2011) etc. There are a lot of related techniques, such as, FOIL (Quinlan and Cameron 1993), PRA (Lao and Subramanya 2012), etc. FOIL makes inferences mainly by learning the first-order Horn clauses from text corpora. PRA deals with the inference problem by using the structural information, i.e., the relation paths between concepts. However, when the concept pairs have no relation paths between them, PRA fails to make any inference. For example, in Freebase there are 13% such concept pairs which cannot be inferred effectively. To cover the shortage of the structural inference methods, we propose the Content-Structural Relation Inference method (CSRI) combining the structural information and the content information of concepts into a unified setting. The contribution of CSRI is to take full advantage of the rich attributes of concepts in the mainstream knowledge bases, so as to get a better inference result.

Content-Structural Relation Inference

CSRI can be formalized as follows. Given a concept s and relation R, we aim to find the concept nodes which have relation R with s. Firstly, find the nodes reached from s by using the breadth-first search up to the breadth of three and the nodes identical with s in attributes to form the candidate node set T . Secondly, for node $t \in T$, we calculate the probability $P(R_{s,t} = R)$ of the occurrence of relation R between s and t. At last, we rank those concept nodes in terms of $P(R_{s,t} = R)$ and select the top N nodes as the inference result. As is known to all, KB consists of concepts, and two mutually exclusive parts, i.e., the structural or relation information, and the content or attributes information of concepts. Therefore, $P(R_{s,t} = R)$ can be partitioned according to the law of total probability:

$$P(R_{s,t} = R) = P(R \mid A_{s,t}) \cdot P(A_{s,t}) + P(R \mid P_{s,t}) \cdot P(P_{s,t}) \quad (1)$$

where $A_{s,t}$ denotes attribute information of s and t, $P_{s,t}$ denotes relation information between s and t, $P(R \mid A_{s,t})$ and $P(R \mid P_{s,t})$ are two conditional probabilities, $P(A_{s,t})$, $P(P_{s,t})$ are two prior probabilities such that $P(A_{s,t}) + P(P_{s,t}) = 1$ and they can be inferred from the knowledge base itself. So we proceed by computing these probabilities in the following three steps. Step 1), we elaborate the procedure to calculate $P(R \mid A_{s,t})$ in (1). The idea is to decompose it into two measures S_1 and S_2 , and set $P(R \mid A_{s,t}) = \max[S_1, S_2]$ where $S_1 = Sim(s,t)$ is the attributes similarity between s and t, $S_2 = \max [Sim(s, s_i) \cdot Sim(t, t_i)]$ is the attributes similarity between node pair (s,t) and $(s_i, t_i) \in P_R$ where P_R is the node pair set in which each pair have the relation R between nodes. Next, we illustrate how to compute the attribute similarity between concept nodes, e.g. $Sim(s,t)$. Since there are many noisy attributes of concept nodes in the KB, firstly we have to select some significant or discriminative

attributes for inference. Suppose that the attributes of *s* and *t* constitute the set A and B, respectively. We employ the classic induction decision tree method to select the discriminative attributes with respect to R to form set C, the reason we use the induction decision tree is to regard the attributes as features of the classification problem with respect to different relation R. Secondly, we compute the common elements among the three sets A, B and C to form set D. After these two steps, the attribute similarity $Sim(s,t)$ is equal to the number of identical attribute values of *s* and *t* whose attributes belongs to D, divided by the number of elements of C. For example, for relation R, suppose that $A=\{a_1,a_2,a_3,a_4\}$, $B=\{a_1,a_3,a_4\}$, and assume that we get the significant attribute set $C=\{a_1,a_2,a_3\}$. By computing the common elements among A, B, and C, we get $D=\{a_1,a_3\}$. Furthermore, for concept *s*, if $a_1=\alpha$, $a_3=\beta$ and for concept *t*, $a_1=\alpha$, $a_3=\gamma$, then we deduce $Sim(s,t)$ is equal to $1/3$. Step 2), we elaborate the procedure to calculate $P(R|P_{s,t})$. Here we adopt the idea of PRA and define the value of $P(R|P_{s,t})$ by combing the results of different random walks through relation graph. More detail can be referred to (Lao and Subramanya 2012). Step 3), it remains to determine two prior probabilities $P(A_{s,t})$ and $P(P_{s,t})$ which can be estimated by the maximum likelihood estimation technique on the training set. Moreover, they satisfy the equation $P(A_{s,t})+P(P_{s,t})=1$.

Experiments

The experiments are carried out via cross-validation on two public data sets, Freebase and Wikipedia. Both of the two data sets are representative KBs. In Freebase, the data is organized in the forms RDF-liked of triples (subject, predicate, object) which are further classified into a series of domains. We select those 1,300,000 triples in the domain of person. In Wikipedia, we selected 1,100,000 triples in the domain of person. For one relation R we remove 20% of their relations in the KB, and then use the pruned KB to infer the removed relations. We also use the measure MAP (mean average precision) to evaluate the inference performance, because the inference result was given as a rank list, and MAP evaluates the overall quality of the rank list. Our comparison baselines are two typical relation inferences method PRA and FOIL. The experiment results are carried out on six typical types of relations between people. By observation, we find that when $P(A_{s,t})=0.01$ and $P(P_{s,t})=0.99$ the performance of CSRI obtains the best. The results are depicted in Table 1.

From Table 1, it is obvious that CSRI obtains the highest MAP values on both data sets. This is unsurprising since the mixture of content information of concepts tackles the structural sparsity and the CSRI leverages the content information and structural information of KB. Moreover, CSRI can infer the relation between the node pairs with no relation paths but PRA fails to do so effectively. Furthermore, we compute the increase between PRA and CSRI. On Freebase the average increase between PRA and CSRI is equal 22.2% and the average increase on Wikipedia is equal 7.9%. Therefore, the average increase on these two data sets by CSRI is equal to 15%. Notice that, CSRI outperforms not much on Wikipedia is because Wikipedia has poorer attributes of concepts than Freebase. It should be mentioned that we also examine the performance of CSRI for other relations, such as location-person relations, and the results of CSRI still competitive.

Table 1: Comparison of different inference methods

dataset	**MAP**			
	relation	**FOIL**	**PRA**	**CSRI**
Freebase	**Children**	0.208	0.501	**0.702**
	Spouse	0.148	0.478	**0.720**
	Parent	0.246	0.413	**0.637**
Wikipedia	**Colleague**	0.191	0.612	**0.688**
	Spouse	0.183	0.603	**0.692**
	Brother	0.213	0.513	**0.584**

Conclusion

In this paper, we propose the CSRI method for relation inference which integrates the attributes information and the relation information of concepts in the knowledge base and experiments demonstrate the effectiveness of it.

Acknowledgments

This work is supported by National Grand Fundamental Research 973 Program of China (No. 2014CB340405), National Natural Science Foundation of China (No. 61173008, 61232010, 61303244), and Beijing nova program (No.Z121101002512063.

References

Minervini P, d'Amato C, Fanizzi N. Learning probabilistic Description logic concepts: under different Assumptions on missing knowledge. In *Proc*. ACM. 2012: 378-383.

Huang Y, Tresp V, Bundschus M, et al. Multivariate prediction for learning on the semantic web. Inductive Logic Programming. Springer Berlin Heidelberg, 2011: 92-104.

Quinlan J.R; Cameron-Jones R.M. 1993. FOIL: a midterm report. In *Proc*. *ECML*. page 3-20.

Lao N, Subramanya A, Pereira F, et al. Reading the web with learned syntactic-semantic inference rules. In *Proc*. *EMNLP*. pages 1017-1026.