KEIPD: Knowledge Extraction and Inference System for Personal Documents

Zhaoyang Lv[1], Yuanyuan Liu[1], and Xiaohui Yu[1,2](✉)

[1] School of Computer Science and Technology, Shandong University, Jinan, China
{zhylv,yuanyliu}@mail.sdu.edu.cn
[2] School of Information Technology, York University, Toronto, ON, Canada
xyu@sdu.edu.cn

Abstract. Public personal documents on the Internet, such as resumes and personal homepages, may imply social relationships among people, which is of great value in various applications. This paper presents KEIPD, a system to extract and infer knowledge from personal documents. KEIPD employs a tree-similarity based approach to extract information from personal documents to obtain a relational network of entities. Then the inference of social relationships can be transformed into a link prediction problem. KEIPD implements some popular unsupervised predictors for link prediction and prune the candidate entity pairs based on the domain-dependent constraint.

1 Introduction

There is plentiful public personal information of celebrities on the Internet, e.g. resumes, personal profiles and personal homepages, which may imply social relationships among the celebrities. For example, two people may be schoolmates if they have been studied in the same university during an overlapped time period. This information can be organized as a social network to support community discovery, most influential nodes discovery and other researches. Compared with traditional social networks, it has some distinguished characteristics: First, links in this network may represent various types of relationships (e.g. schoolmates and colleagues) rather than homogeneous relationships. Second, the network is more realistic where the links are deduced from the factual experiences of people instead of the interaction data of users via a social application. Third, the formation of a link is sensitive to time as the example described above.

The construction of such a social network can be viewed as a two-step process. We can build a relational network by extracting events from personal documents where nodes represent main entities in the documents, including the person, the organizations he belongs to, the locations of these organizations, etc. Then the social network can be regarded as a view of the relational network after predicting the link between arbitrary person-person pair. Challenges to implement such a system can be concluded as follows: the information unit in a personal document is an event rather than a binary relation, which is more complicated to extract; how to infer knowledge properly on a network embedded with heterogeneous

F. Li et al. (Eds.): APWeb 2016, Part II, LNCS 9932, pp. 592–595, 2016.
DOI: 10.1007/978-3-319-45817-5_72

nodes and links. KEIPD employs a tree-similarity based approach to extract events. For link prediction, unsupervised predictors. The system is based on a considerably mature graph database.

2 System Overview

Figure 1 shows the overview of KEIPD. We will introduce the details of the information extraction module and knowledge inference module in this section.

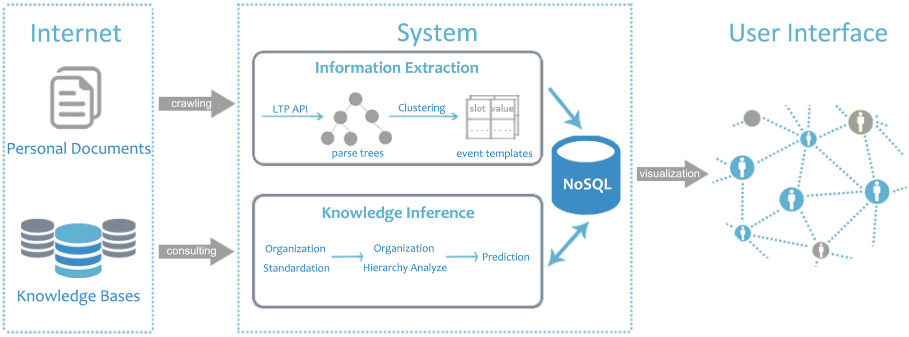

Fig. 1. System overview

2.1 Information Extraction

According to the classical entity types proposed in the Message Understanding Conference (MUC) and the characteristics of personal documents, we consider three entity types here: *Person*, *Organization* and *Location*.

Event Template. Take resume documents for example. A resume displays some fixed classes of events by time order, among which the most typical event is membership, as shown in Table 1. Events in the same class correspond to a common predefined template.

Table 1. The template of membership events

Slot	Note
Type	Type of membership: employment/education/member
Person	Participant
Organization	Organization name
Role	The role of the participant in the organization. For employment, role is the position; for education, role is "student"
Time	Start time and end time of this membership

Tree-similarity based method. We adapt the method from [3] to perform event extraction. It is assumed that sentences describing the same type of events own similar parse tree structures. First, we refer to a integrated natural language processing tool, LTP [1], to preprocess the text through Named Entity Recognition (NER) and Dependency Parsing tasks. The results of these two tasks are merged to constitute a NE-tagged parse tree where key attributes of the nodes include the word, part-of-speech tagging, the results of NER and Dependency Parsing.

Then, the parse trees are clustered with the tree-similarity function:

$$K(T_1, T_2) = m(r_1, r_2) * s(r_1, r_2) + K_c(r_1[\mathbf{c}], r_2[\mathbf{c}]) \tag{1}$$

where

$$K_c(p_1[\mathbf{c}], p_2[\mathbf{c}]) = \arg\max_{\mathbf{a},\mathbf{b}} K(p_1[\mathbf{a}], p_2[\mathbf{b}]) \tag{2}$$

$$K(p_1[\mathbf{a}], p_2[\mathbf{b}]) = \sum_{i=1}^{l} K(p_1[\mathbf{a}_i], p_2[\mathbf{b}_i]) \tag{3}$$

Here, T_1 and T_2 are two trees where r_1 and r_2 are their root nodes. Equation (3) is the similarity function over two arbitrary children node sequences $p_1[\mathbf{a}]$ and $p_2[\mathbf{b}]$. Due to space limitations, see [3] for more details.

We adjust the match function $m(r_1, r_2)$ and the node-similarity function $s(r_1, r_2)$ to suit our data:

$$m(p_i, p_j) = \begin{cases} 0 & p_i.relate = p_j.relate \\ 1 & otherwise \end{cases} \tag{4}$$

$$s(p_i, p_j) = \begin{cases} 0.2 & p_i.ne \neq p_j.ne \\ 0.5 & p_i = O, p_j = O, p_i.pos \neq p_j.pos \\ 0.8 & p_i = O, p_i.pos = p_j.pos \\ 1.0 & p_i \neq O, p_i.ne = p_j.ne \end{cases} \tag{5}$$

The weight in Eq. (5) is assigned empirically according to the discriminative ability of the feature types.

The calculation of similarity starts from leaf nodes and goes up to the root employing a dynamic programming algorithm. We summarize syntactic rules manually for different clusters to fill the corresponding event template.

2.2 Knowledge Inference

Online Knowledge Bases. Considering the complexity of natural language, we process the relational network with the assistance of some online knowledge bases. The hierarchical characteristics of entities belonging to *Location* and *Organization* are key factors for link prediction. Therefore, we crawl an external knowledge base about fine-grained regionalism in China which contains more

Table 2. Unsupervised predictors for link prediction

Predictor	Explanation
Preferential attachment	$\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert$
Common neighbors	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$
Adamic/Adar	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log\lvert\Gamma(z)\rvert}$
Shortest path	The length of the shortest path between x and y
Katz	$\sum_{l=1}^{L} \beta^l \cdot \lvert paths_{x,y}^{(l)}\rvert$

than 700K locations. For organizations, we refer to an online encyclopedia[1] to normalize their names and design a simple algorithm to infer the hierarchy by analyzing prefix relations.

Link Prediction. Besides the predictors shown in Table 2, we also experiment with *rooted PageRank* and*PropFlow*, see [2] for more details. As demonstrated in Sect. 1, the formation of a link is strongly dependent on the time attributes, so we prune the candidate entity pairs before prediction using the time constraint.

3 Demonstration Scenarios

There are about 15K personal documents of politicians crawled from *People*[2] as source data. The system will be demonstrated via two types of query operations:

(1) Point query. Given a specific person as a query condition, the system will return related people with corresponding relationships.
(2) Path query. Given two specific people as query conditions, the system will return all the paths/the shortest path between them.

References

1. Che, W., Li, Z., Liu, T.: LTP: a Chinese language technology platform. In: Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. pp. 13–16. Association for Computational Linguistics (2010)
2. Davis, D., Lichtenwalter, R., Chawla, N.V.: Multi-relational link prediction in heterogeneous information networks. In: 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 281–288. IEEE (2011)
3. Zhang, M., Su, J., Wang, D., Zhou, G., Tan, C.-L.: Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In: Dale, R., Wong, K.-F., Su, J., Kwong, O.Y. (eds.) IJCNLP 2005. LNCS (LNAI), vol. 3651, pp. 378–389. Springer, Heidelberg (2005)

[1] http://baike.baidu.com.

[2] http://www.people.com.cn.