A Novel Method for Mining Relationships of Entities on Web

Xinyan Huang 1,3, Xinjun Wang*1,2, Hui Li1, Yongqing Zheng1

1Shandong University
Num 1500, SunHua Road in High Tech Industrial Development Zone, Ji’nan, China
2dareway software co.,ltd
Num 1500, SunHua Road in High Tech Industrial Development Zone, Ji’nan, China
3Shandong University of Finance and Economics
Num 7366, East Erhuan Road in Lixia Zone, Ji’nan, China
20063462@sdufe.edu.cn, corresponding author: wxj@sdu.edu.cn, lih@sdu.edu.cn

Abstract. In this paper, we address an important and difficult problem: mining relationships of entities on web. We propose a novel and efficient framework to solve the aforementioned problem. Different from all the existing methods of mining relationships of entities, our method makes full use of the events that have been extracted already to mine relationships between entities. Pairs (E, T) are employed to represent the close relationships. In addition, our method could mine the deep semantic meaning through analyzing the events that the relationships occur in. In the future, multidimensional search and deep analysis are to be performed on these pairs (E, T) and events.

Keywords: relationship; event; entity

1 Introduction

Mining semantic relationships between entities has received much attention lately. The purpose of mining entity relations is to identify the semantic relations between entities, which is an important issue in the information extraction and data mining [3].

Existing methods mainly depend on the relations templates and the context that entities extracted from. Recently, Luo [7] proposed a method for mining relations between entities based on web. The top ranked Web page pairs are likely to contain the relationships between the two entities. StatSnowball [6] uses the discriminative Markov logic networks and softens hard rules by learning their weights in a maximum likelihood estimate sense. Renlifang based on the StatSnowball can mine the relation between people.

However, definition of templates for comprehensive relationships of entities is difficult. Given the dynamics, huge scale, and unstructured feature of web, mining the relationships between entities is still a challenge [4] [1]. In this paper, we study the problem of automatically generating semantic relations for entities through events. An event is something that happens at some specific time. We want to provide co-
occurrence relations in an event between entities, which indicate the semantics and hidden association of entities. We can find in the different time interval, a pair of entities might have different semantic relations.

We propose that if at least mine entities that are in the same event for at least mint timestamp [5], and then there are relations between them. Then we denote this group of entities as E and the set of these timestamps as T which are not required to be consecutive, entities in E is represented as E1,E2… etc and timestamps is represented as t1, t2…etc. If we set mine = 2 and mint = 3, we can find that entities {E2, E4} always occurs together at \{t1, t2, t3, t4\} in Figure 1. Such result discovers an interesting relationship between E2 and E4.

Fig. 1. Events that entities co-occurs in.

2 Problem Formulation

The printing area is 122 mm × 193 mm. The text should be justified to occupy the full line width, so that the right margin is not ragged, with words hyphenated as appropriate. Please fill pages so that the length of the text is no less than 180 mm, if possible.

We formally define the notions related with mining relationships of entities as follows:

**Definition 1.** Event. An event is something that happens at some specific time, and often some specific place, which is usual a phrase or sentence in the web pages including an activity. In our paper, the event of an entity is defined as a model E with five attributes as mentioned in our former work [2], described as follows:

\[ E <\text{subject}, \text{activity}, \{\text{object}\}, \text{time}, \{\text{location}\}> \]

Among them, subject, activity, activity and time elements are required.

In this paper, the entities we concerned are enterprises. So we filter out all the events in which the subject and the object attributes can’t be null and both are enterprises. And then select n entities the occurrence frequency of which is beyond \( \sigma \).

Let \( E^* = \{E1, E2, \ldots, En\} \) be the set of entities and \( T^* = \{t1, t2, \ldots, tm\} \) be the set of all timestamps. A subset of \( E^* \) is called an entityset E. A subset of \( T^* \) is called a
timeset $T$. The size, $|E|$ and $|T|$, is the number of entities and timestamps in $E$ and $T$ respectively.

**Definition 2.** Mining Relationships of Entities. Mining the relationship of entities is to mine all the pair $(E, T)$. A pair $(E, T)$ meets that all entities in $E$ are in the same event at any timestamp in $T$. Specifically, given two minimum thresholds $mine$ and $mint$, where $E = \{E_1, E_2, \ldots, E_q\} \subseteq E^*$ and $T \subseteq T^*$, it needs to satisfy two requirements:

1. $|E| \geq mine$: There should be at least $mine$ entities.
2. $|T| \geq mint$: Entities in $E$ are in the same event for at least $mint$ timestamps.

In addition, in a final pair $(E, T)$, both $E$ and $T$ can’t be enlarged.

**Fig. 2.** Generation of Pairs $(E, T)$ $|E|=2$ and $|T| \geq mint$.

### 3  Mining Relationships of Entities Algorithm

Given the entities we concerned are enterprises, the number of entities $|E|$ is relatively small and the number of timestamps is relatively big.

According to the issue of Mining relationship of entities, a novel method is proposed. Details are presented in Section 3.1 and Section 3.2.

#### 3.1  Mining All the Pairs $(E, T)$ $P$ in Which $|E|$ Equals Two and $|T| \geq mint$.

In this section, for every two entities in $E^*$, we should get all the pairs $(E, T)$ in which $|E|$ equals two and $|T| \geq mint$. Take Figure 1 for example, both $mine$ and $mint$ are set 2. The pairs mined as Figure 2 shows.

From Figure 2, entity $E_1$ and entity $E_2$ have meaningful relationship. The type of relationship can be acquired from activity attributes of events $\{e_5, e_7, e_6\}$.
3.2 Mining All the Pairs (E, T) P’ in Which Both E and T Can’t Be Enlarged.

In this section, a hierarchical clustering method is employed to merge all the pairs (E, T) mined from section 3.1. Details as algorithm 1 shows.

Algorithm 1. Mining all the pairs (E, T) P’ in which both E and T can’t be enlarged

Input: All the pairs queue P (according to time line), support threshold min_e for |E| and min_t for |T|

Output: The final pairs (E, T) P’

begin
P’ ← Φ;
p = Pair(P); /* select left pair from P*/
while (true)
    Booleanflag=true;/*a flag whether merge happened*/
    while (flag)
        s = Pair(P); /* select another left pair from P*/
        n1, n2 = Compare(p, s); /* count the number of the same entities in E and the number of the same timestamps (events) in T */
        if n1 ≥ min_e and n2 ≥ min_t
            p = p.Merge(s);
        else
            flag = false;
        endif
    end
    Push(p, P); /* push p to P from the right*/
    if (!flag)
        break;
    endif
end
P’ = P; /* put these patterns that don’t grow to S’*/
Return P’;
end.

According to Algorithm 1, final pairs (E, T) P’ can be gained from Figure 2. Pair{E_1, E_2, E_4} \{t_1, t_2, t_4\} in \{e_5, e_7, e_6\} and pair{E_2, E_4} \{t_1, t_2, t_3, t_4\} in \{e_5, e_7, e_{11}, e_6\} show the close relationship between E_1, E_2 and E_4. The semantic meaning of the relationship can be gained from the events \{e_5, e_7, e_6\} and \{e_5, e_7, e_{11}, e_6\}.

4 Conclusion

In this paper, we address an important and difficult problem: mining relationships of entities on web. We propose a novel and efficient framework to solve the aforementioned problem. Different from all the existing methods of mining
relationships of entities, our method makes full use of the events that have been extracted already. In addition, our method not only could mine all the relationships between entities and could mine the deep semantic meaning through analyzing the events that the relationships occur in. In the future, multidimensional search and deep analysis are to be performed on these pairs (E, T) and events.

Acknowledgments. The work is supported by the National Natural Science Foundation of China under Grant No.61303005, Science and Technology Development Plan Project of Shandong Province No.2014GGX101019 and No.2015GGX101007, and the Fundamental Research Funds of Shandong University No. 2014JC025, No.2015JC031 and No.2015JC031

References