Domain-oriented Entity Set Extension Based on Bipartite Graph

Du Liming[a, *], Yahaya Abdulhamid[b], Li Gui, Wang Fengying, Dong Jie

Faculty of Information & Control Engineering, Shenyang Jianzhu University, Shenyang, China 110168

[a]duliboy@163.com, [b]Abdulhamidyahaya1@gmail.com

*Corresponding author

Keywords: Web Data; Data extraction; Data Mining.

Abstract: This paper mainly studies the problem of entity set extension in domain data. On the basis of bipartite graph methods, given an improved algorithm which consider the general case that the edges of the graph have different weights. The designed algorithm can dynamically adjusts the the size of the entity extension sets in the given number of initial iterations, and when the number of iterations exceeds the threshold, the size of the entity extension sets remains unchanged, which can reflect the actual situation more accurately. Finally, the effectiveness of the algorithm is verified by experiments.

1. Introduction

With the development of the network, the amount of available information has been steadily growing every year. Email, social networks, and discussion forums are all contribute to rapid growth of data. It can be said that the World Wide Web has provided a convenient mechanism for document issuance and gradually became the largest public data source in the world [1-3]. For most applications, the user is actually only concerned with certain specific data objects, such as the products of a real estate website. And for each data object, the user only pays attention to certain specific data items, such as housing prices, housing size, floors, and room types [4]. It can be said that domain-oriented services are increasingly important in today's information age.

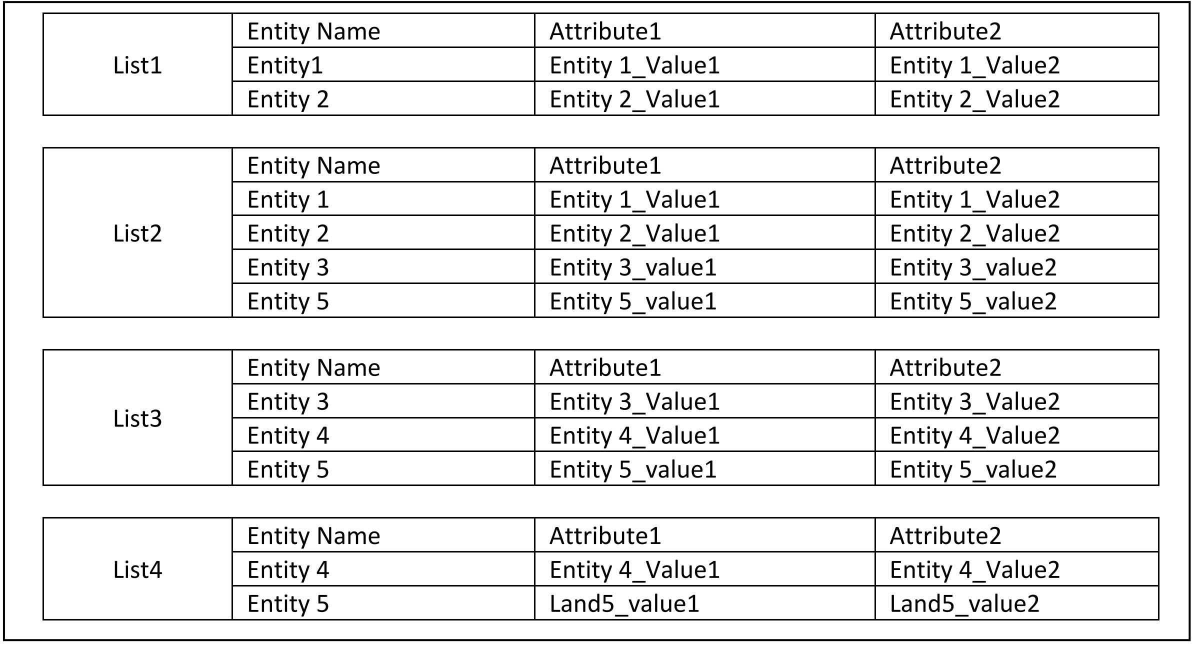
For the domain web data, entities refer to things that exist independently in reality, and entities usually have their own characteristics. That means different entities have specific attributes and can be distinguished from other entities. In general, people specify a name for each entity, therefore, the entity is also called named entity. Similar entities generally have similar attributes and different entities always have different attributes. In the domain applications, the entity can be defined according to the user's target requirements. Entity set expansion refers to expanding the set of seed entities into a more complete collection of entities which belong to the same set of concepts. Obviously, when users use the Web to query data, they essentially perform entity information queries, so solving the entity set expansion problem will greatly reduce the burden on the user for screening and comparison, also improving the user experience.

As mentioned above, the study of entity set expansion has important research significance and application value. The expansion of entity collections has become a current research hotspot. Many researchers and enterprises have paid attention to this topic and achieved some results [5-7]. The most prominent one is Google Sets, which uses some special algorithms for set expansion, but Goggle Set's algorithms are not open source for business confidentiality. Another well-known research result is the Seal system. In ref [6], static threshold algorithm and dynamic threshold algorithm are proposed. Based on these two algorithms, we given an improved algorithm. Compare with the exits method, we introduce a new entity similarity judgment criterion, and proposed a new method to balance the dynamic and static threshold.

The literature introduces the concept of bipartite graph to study the expansion of the entity set. Based on this, this paper introduces a new entity similarity judgment criterion to improve the original algorithm and achieve better results. The effectiveness of the algorithm is verified by experiments.

2. Domain Entity Extension Model

The data in the Web list page of this chapter is used as a research object. The data in the list page is usually obtained from the back-end database and displayed on the page according to a fixed template. It usually contains one or more groups entity object. Such as the example showing in Fig1.



List1	Entity Name	Attribute1	Attribute2
	Entity1	Entity 1_Value1	Entity 1_Value2
	Entity 2	Entity 2_Value1	Entity 2_Value2

List2	Entity Name	Attribute1	Attribute2
	Entity 1	Entity 1_Value1	Entity 1_Value2
	Entity 2	Entity 2_Value1	Entity 2_Value2
	Entity 3	Entity 3_value1	Entity 3_value2
	Entity 5	Entity 5_value1	Entity 5_value2

List3	Entity Name	Attribute1	Attribute2
	Entity 3	Entity 3_Value1	Entity 3_Value2
	Entity 4	Entity 4_Value1	Entity 4_Value2
	Entity 5	Entity 5_value1	Entity 5_value2

List4	Entity Name	Attribute1	Attribute2
	Entity 4	Entity 4_Value1	Entity 4_Value2
	Entity 5	Land5_value1	Land5_value2

Fig 1 Data on list1 to list4

Combining the characteristics of the list data, the plot list data and plot entities in the above figure are modeled as bipartite graphs, as shown in Fig 1. In the extraction process, each list is modeled as a node on the right side of the graph, and each entity that appears in these web page lists is modeled as a node on the left side. In Fig.2, these underlined points "Entity 1" and "Entity 2" is the seed instance, the remaining "Entity 3", "Entity 4" and "Entity 5" are candidate entity. If an entity is included in a web page list, there will be an edge connection between the entity node and the list node. e.g. list2 connect with " Entity 1"、" Entity 2", and "Entity 3", which indicating that all three entities are members of "List2". The weight of the edge can be assigned to the value by additional information (such as the quality score of the webpage). In the experiment, the weight of each edge in the bipartite graph model is set to 1 for the sake of simplicity.

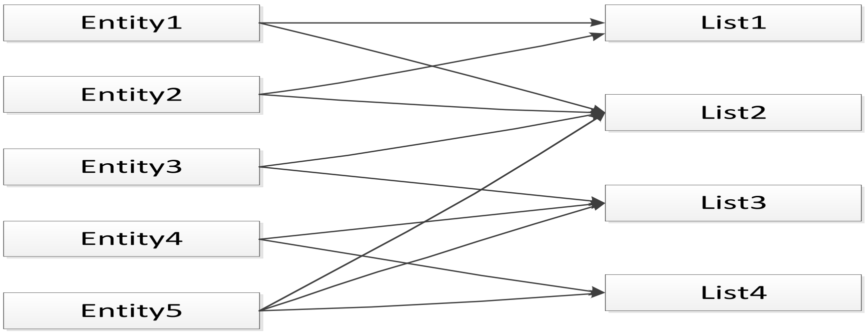


Fig 2 Data model of bipartite graph of Web list

3. Similarity Calculation between Entities

According to a given seed entity, the problem of finding similar entities can be seen as a feature of the node on the right side in the bipartite graph data model, and the problem of finding an entity node similar to the seed node is found. In order to calculate the similarity between entity nodes, this paper uses Jaccard similarity calculation method and cosine similarity calculation method.

Let $n1, n2$ for the two physical nodes on the left side of the bipartite graph model, S_{n1} and S_{n2} be the two sets of webpage list nodes that connect the nodes $n1$ and $n2$ in the bipartite graph model. Then the Jaccard similarities between $n1$ and $n2$ expressed as formula (1).

$$Sim_{jaccard}(n1,n2)=\frac{|S_{n1}\cap S_{n2}|}{|S_{n1}\cup S_{n2}|} \tag{1}$$

Let $n1, n2$ be the two physical nodes on the left side of the bipartite graph model, and use the weight vectors V_{n1} and V_{n2} to represent the weights of the edges connecting the webpage list nodes and the entity nodes n1and n2in the bipartite graph model. Then the cosine similarity of $n1$ and $n2$ can be expressed as (2).

$$Sim_{\text{cosine}}(n1,n2)=\frac{V_{n1}\cdot V_{n2}}{\|V_{n1}\|\|V_{n2}\|} \tag{2}$$

Equations (1) and (2) define the similarity of entities from different aspects. To make the entity similarity criteria more realistic, this paper combines (1) and (2) and introduces new entity similarities. Judgment criteria, as shown in formula (3).

Definition 1. Let $n1, n2$ be the two physical nodes on the left side of the bipartite graph model, $Sim_{jaccard}(n1,n2)$ is the Jaccard similarities between $n1$ and $n2$, $Sim_{\text{cosine}}(n1,n2)$ is the cosine similarity of $n1$ and $n2$ .Then the similarity of $n1$ and $n2$ can be expressed as formula (3).

$$Sim(n1,n2)=\alpha Sim_{jaccard}(n1,n2)+(1-\alpha)Sim_{cosine}(n1,n2), \quad \alpha\in[0,1] \tag{3}$$

Remark 1. In this paper, we use formula (3) to calculate the similaritis between two nodes, it belongs to the combination of (1) and (2). When it is similar, its similarity degenerates to Jaccard similarity. At that time, its similarity is converted to cosine similarity.

4. Extended Entities Sets Quality Assessment

The existing evaluation method first calculates the similarity between each entity and a given seed entity, and then ranks them by the level of the similarity score. This paper considers the extended entities sets as a whole and proposes a simple method to evaluate the quality of extended sets.

Let E is a collection of all entities, X is a set of extended entities where $X\subseteq E$, S is the seed sets where $S\subseteq E$, Sim is defined in (3) which can evaluate the similarity of any two entities, then the similarity of X and S is defined as (4).

$$S_r(\mathrm{X},\mathrm{S})=\frac{1}{|X||S|}*\sum_{x\in X}\sum_{s\in S}Sim(x,s) \tag{4}$$

Remark 2. Formula (4) give us the basis for judging the similarity between the extended set and the seed set. The greater similarity score between the extended entity set and the seed entity, then the better quality of the extended entity set.

Remark 3. Formula(4) cannot fully display the quality of the extended entities sets, for the purpose of set extension is to find a similar and consistent concept set with a given seed entity. In some cases, although extended entities are similar to seed entities, they do not belong to the same concept set.

Let E is set of all entities, X is an extended set with $X\in E$, Sim is defined in (3), which is to evaluate the matrix of similarities between two entities. The consistency of X is expressed as (5).

$$S_c(X)=\frac{1}{|X|^2}*\sum_{i}^{|X|}\sum_{j}^{|X|}Sim(x_i,x_j), \quad x_i,x_j\in X \tag{5}$$

Since the quality of the set of extended entities needs to be expressed simultaneously with similarity and consistency, so in this paper the weighted sum of the (4) and (5) is used in this paper, and give the definition 2.

Definition 2.Let E is set of all entities, X is an extended set with $X \in E$, S is the seed sets where $S \subseteq E$, S_r is defined in (4), and S_c is defined in (5), then the extend quality evaluation function is as (6).

$$Q(X,S) = \beta * S_r(X,S) + (1-\beta) * S_c(X,S),\ \beta \in [0,1] \tag{6}$$

An expanded seed set X consists of higher quality entities in a candidate set. We can use $Q(X,S)$ to evaluate the score, if the score is high, then the X is good.

5. Entity Set Expansion Algorithm

According to the Definition1 and Definition 2, the problem of entity set expansion can be described as a request for solution process: suppose all the candidate entities E and seed entities S are known, we first evaluate the function of two entity similarities, find the initial expand the set of entities X_0, then through iterative method to find $X_1, X_2 \ldots X_n$, until the best is reached. Following will give the algorithm description.

```
Input entity seed set $S$ and the bipartite graph
 For each entityingraph.entitiys do
     Q1_Score:=Q(entity_i, S)
End for
Sort entity_i by Q1_Score[i] desc
$K_0$:=Pick_Threshold(Q1_Score[i])
$X_0$:=the top $K_0$ ranked term by Q1_Score[i]
iter:=1
While true do
      For each entity_iin graph.entitys do
          Q2_Score[i]:=Q(entity_i,$X_{iter-1}$)
          g(entity_i):= $\gamma Q_1\_Score[i] + (1-\gamma) Q_2\_Score[i]$
     End For
    Sort entity_i by g(entity_i) desc
   If iter<=INITIAL_MAX_ITER Then
        $K_{iter}$:=Pick_Threshold(g(entity_i))
        $X_{iter}$:=the top ranked $K_{iter}$entiys by g(entity_i)
        $K$:= $K_{iter}$
        iter++
  Else
        $X'_{iter}$ :=the top ranked K entiys by g(entity_i)
        If $X'_{iter} \neq X_{iter-1}$ Tthen
             Let $r \in X'_{iter}$ be the top ranked entity not in $X_{iter-1}$
              Let $q \in X_{iter}$ be the last ranked entity in $X_{iter-1}$
             $X_{iter} := (X_{iter-1} \cup \{r\} - \{q\})$
        Else
             $X_{iter} := X_{iter-1}$
             break
        EndIf
   iter++
    EndIf
End while
```

Remark 4. Compared with the algorithm mentioned in the literature [6], this algorithm introduces a new calculation formula for entity similarity assessment. It fully considers the weight factors of the edges in the bipartite graph and is therefore more in line with the actual situation.

Remark 5. The threshold K calculated by the first iteration does not accurately reflect the actual size of the entity expansion set. As the iteration continues, the new threshold calculated on the new fraction distribution may be very different from the initial threshold value, so the algorithm dynamically adjusts the threshold in the given number of INITIAL_MAX_ITER iterations. When the number of iterations exceeds the threshold, the threshold remains unchanged. This reflects both the readiness of the threshold and the guaranteed algorithm convergence.

6. Experiment

In the experiment, to verify the effective of our algorithm, we have crawled the land information from some real estate websites such as China's land market(www.landchina.com),and uses XML files to store the land entity information. The experiment result is as Fig.3.

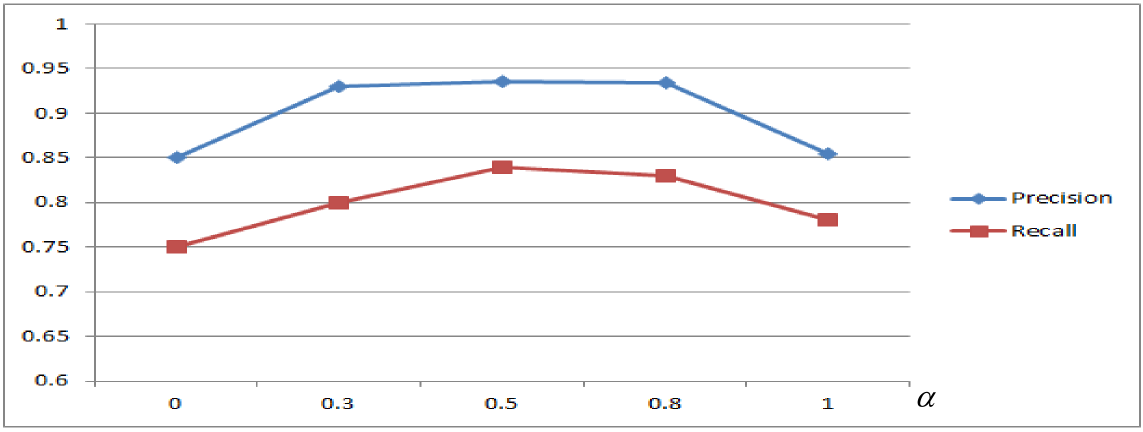


Fig 3 The precision and recall of the experment

From Fig 3, we can see that when α = 0.5, the performance of the algorithm is the best. The precision rate is more than 0.9, verify the effectiveness of the proposed algorithm.

7. Conclusion

This paper first introduced the related concepts of entities, the significance of the expansion of entity collections, the research status and existing problems, and then based on two partite graph methods, given an improved algorithm which consider the general case that the edges of the graph have different weights, and the size of the entity extension sets is dynamically adjusted in the process of the program iteration. Therefore, the improved algorithm can reflect the actual situation more accurately. Finally, the effectiveness of the algorithm is verified by experiments.

Acknowledgement

This work was supported by the The Ministry of Housing and Urban Rural Development Foundation of China under Grants 2017-K8-038

References

[1] Hu D,Meng X. Automatic Data Extraction from Data-Rich Web Pages. The 10th Data System for Advanced Applications (DASFAA), Beijing, 2005

[2] B. Adelberg. NoDoSE–a tool for semi-automatically extracting structured and semistructured data from text documents. In SIGMOD, 1998.

[3] A.Sahuguet and F. Azavant. Web ecology: Recycling HTML pages as XML documents using W4F.

[4] D. M. Blei, A.Y. Ng and M. I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning

Research, vol. 5, no. 3, 2003.

[5] Zheng Jiahen, Li Xin. The Research of Chinese Names Recognition Method Based on Corpus. Journal of Chinese Information Processing, 2013. 14(1): 7-12.

[6] Li Gui, Chen Shaogang, etc, Web based instance extension and attribute value expansion method, Computer science, 2014, 41 (11A)

[7] Zhou Lei, Research of Complex Named Entity Extraction based on Hybrid Method, 2009, Shanghai Jiao Tong University.