A Novel Holistic-based Entity Unifying Method for Heterogeneous Data

Jiecai Zheng\textsuperscript{1,*}, Zhaoxia Lu\textsuperscript{2}, Xinxiao Zhao\textsuperscript{3} and Xueqing Li\textsuperscript{1,*}\textsuperscript{b}

\textsuperscript{1}School of Sport Communication & Information Technology, Shandong Sport University, Jinan, China
\textsuperscript{2} School of Sport Communication & Information Technology, Shandong Sport University, Jinan, China
\textsuperscript{3}School of Computer Science and Technology, Shandong University, Jinan, China
\textsuperscript{a}Email: zhengjiec@gmail.sdu.edu.cn
\textsuperscript{b}Email: xqli@sdu.edu.cn
\textsuperscript{*}Corresponding Author

Abstract. With the fast growing of the informatic technologies and extensive accumulation of informative sources. The information is gradually presented with more characteristics such as messiness and heterogeneity. Traditional entity unifying methods tend to have good measurement results for structured data with uniform structure and relevant attributes. For unstructured multi-sourced data, the entity similarity cannot be accurately measured. To overcome the defects of the traditional methods, this paper proposes an entity unifying method for heterogeneous data by calculating the entity similarity based on "relation". Extensive experiments on different types of datasets by have been conducted to demonstrate the superiority of our method.

1. Introduction

With the fast growing of the informatic technologies and extensive accumulation of informative sources like digit city, digit sky and digit earth et al, the related content of information is no longer limited to monotonous Web pages and simple structured databases. The information is gradually presented with more characteristics such as messiness and heterogeneity\textsuperscript{[1][2]}. At the same time, heterogeneous data fusion, multi-source data integration and other related technologies in the field of information technology have also been widely concerned and continuously applied.

The traditional entity unifying methods are normally build upon property features\textsuperscript{[3][4]}. The core of these methods is to determine whether different representations point to the same entity by calculating the similarity of each property feature and the preset similarity threshold. Traditional methods tend to have good measurement results for structured data with uniform structure and relevant attributes. For unstructured multi-sourced data (the data structure is not uniform, the data attributes are not complete, and the data source does not describe the same object in the same way), often the similarity cannot be accurately measured. In addition, the corresponding Web data obtained by crawling technology has more diversified data forms, and its data expressions often do not follow a unified data representation. Therefore, the similarity of data entities cannot be guaranteed by the property similarity.

To overcome the defects of the traditional methods, this paper proposes an entity unifying method for heterogeneous data by calculating the entity similarity based on "relation". The main contributions of this method are as follows:
In this paper, the method of entity unifying is proposed, which optimizes the mutual influence of similarity matching on entities, and ultimately realizes the entity unifying of all representations.

In the process of the similarity calculation, this method makes comprehensive use of "property features" and "contexts" to measure the similarity. Besides, the method makes use of the improved "quasi-group" data structure to representing the similarity measure of "relationship" between different appearances, which enhances the accuracy of entity similarity.

Extensive experiments on different types of datasets by have been conducted to demonstrate the superiority of our method.

2. Method of Measuring Entity Similarity

According to the introduction of traditional entity unifying methods, the basic idea of the traditional unifying algorithm is as follows \([5][6]\): compute the pair wise similarity measure based on the image attributes, compare several property pairs in the attribute class in turn and complete the overall similarity measure of entity representation by transitive closure. When the similarity of two representations exceeds the preset threshold, two representations can be considered to point to the same entity. Otherwise, the two representations point to different entities. When the corresponding property features of the image to be compared are relatively complete, it is easy to make accurate judgments as to whether the entity representation can be unified through the similarity measure of the image attribute classes. However, under the circumstance of multi-source heterogeneous data, it may not be possible to draw a correct conclusion by relying on the appearance attributes alone to make an assessment of the similarity of appearances. Therefore, we consider supplementing the judgments of appearance similarity from the "context" and the relation between them, expecting to get relatively accurate determines of similarities. However, there may still be some problem: suppose that there is a possibility of entity unifying between representation A and representation B, C, D, the similarity judgment between representations A and B may affect the similarity measure between A and the other representations (C, D).

Based on the above considerations, using the definition of "group" in Section 4.2, each representation is considered as a class, and the interrelationship between the representations is represented in the form of graphs, given the preset similarity threshold. Subsequently, pairs of all representations are made grouped and similarity measures are computed from the pair. The result of the similarity measure is stored in the priority queue. The algorithm proposed in this chapter will extract the most similar representations from the priority queue and merge them each time to form a new representation class. The newly merged foresaid class needs to update the similarity of other related foreigner classes and also update the priority queue. Then select the pair of most similar pairs from the updated priority queue, and repeat the above merge and update operations until there are no similarity table pairs greater than the threshold in the priority queue. Through constant loop matching, the algorithm finally achieves the overall entity unifying. Each output of the class can considered as one entity, and all the representations inside each class are uniformly directed to the entity.

The algorithm is implemented as follows:
Algorithm 4.1 Algorithm on Holistic based Entity Unifying

**Input:** entity set $S$ to be compared, preset threshold $\gamma$

**Output:** entity set after unifying $C'$

1. $C = \text{InitClustering}(S)$; // Initialization hierarchical cluster class set
2. $\text{priQ} = \text{InitQueue}();$ // Initialize prior queue
3. for each cluster pair $<c_i, c_{i+1}>$
4. $\text{sim}(c_i, c_{i+1}) = \text{Calcsim}(c_i, c_{i+1});$
5. $\text{priQ}.\text{Enqueue}(\text{sim}(c_i, c_{i+1}));$
6. end for
7. while $\text{priQ}.\text{length}() \neq 0$
8. item = $\text{priQ}.\text{Dequeue}();$ // get the current most similar match pair $<c_x, c_y>$ in queue
9. simValue = Calcsim(item.first, item.second);
10. if simValue $< \gamma$ then break;
11. end if
12. $c_{xy} = \text{Merge}(c_x, c_y);$ // merge two origin classes
13. $\text{UpdateCluster}(C, c_{xy});$
14. end while
15. return $C;$

This algorithm initializes the hierarchical clustering operation by specifying the similarity threshold $\gamma$, and performs similarity calculation based on the class sets. The final output unified entity set. The 1st line of the algorithm initializes the collection of entity representations, and uses a cohesive hierarchical clustering to complete the initialization of the class sets.

The line 3-6 performs a similarity measure on the initialization class set and stores the similarity calculation result in the priority queue priQ.

Starting from 7th line of the algorithm, the matching set of the class set with the largest similarity value is obtained from the priority queue priQ in turn. Looping start from the following steps:

Step 1: Algorithm line 8, the first item of the priQ (similar value descending order) pop out, and determine whether the similarity exceeds the threshold $\gamma$.

Step 2: Algorithm line 10, if the similar measurements of the current item is lower than the threshold, quit from the while loop.

Step 3: Algorithm line 11, if the similarity value of the current item is greater than the threshold value, the merge operation of the class set $(c_x, c_y)$ is performed, and the similarity measure of the class set in the queue priQ is updated in the 13th line of the algorithm.

Step 4: After the updating queue priQ, loop again starts from step 1 until the similarity value of the current item is less than the threshold.

In the last step, the algorithm outputs the unified entity class set C, and the algorithm ends.

From the algorithm described above, we can easily find that, in the holistic-based entity unifying algorithm, the document representations need to be paired and measured the similarity. The computation needs to be recalculated in the update operation. The time complexity of the algorithm is $O(n^2)$ (n is the number of the entity).

In addition, in order to further reduce the number of class matching pairs, we can also use Blocking technique to sort and segment preliminary similar representations and compare similarity in smaller data blocks to reduce the number of time complexity. However, it should be noted that different block feature selection may miss some candidate matching pairs. The current solution is to use multiple block partitions to prevent the loss of candidate matching pairs.

3. Experiment Results and Analysis
The document entity unifying method proposed in this chapter can be used to the scene with relatively abundant appearance association such as: product information dataset, document information dataset and knowledge-based source dataset. In this experiment, we select three typical datasets for testing:
Deep net based product dataset (DBPI), Document-based DBLP dataset and Internet based knowledge dataset (DBKI) \[7\]. These three datasets all have been pre-processed. The ground truth is also labelled to evaluate the results.

We performed three different entity unifying experiments on three typical datasets. For each dataset, we separately use the three similarity measures described above to perform differential entity unifying experiments: using only property similarity, combing property similarity measurement with contextual similarity and combing all three of them.

We compare our results with the ground truth and classified the results. The precision is computed accordingly. Experiment results are shown as follows:

![Precision On DBPI](image.png) ![Precision On DBKI](image.png) ![Precision On DBLP](image.png)

**Figure 1.** Precision on Different Dataset

From the figures above, the overall performance of holistic-based improves significantly by comprehensive using of various measures of similarity. The reason is that the accuracy of matching decisions can be guaranteed when more features of similarities are considered, but it is often difficult to make matching decisions when there is "insufficient evidence". Experiment shows that comprehensive using of various measures of similarity can improve the overall performance of the entity unifying compared to the single similarity measurement.

4. Conclusion

This paper proposes a holistic-based entity unifying approach which is based on the cohesive hierarchical clustering. This approach generates iterative entity unifying by utilizing the mutual influence between different matching decisions, and finally achieves the entity unification of all the representations. Experiments on the typical datasets show that the algorithm proposed in this paper can generate an effective entity unifying and achieve good precision score.

5. Acknowledgments

The work is supported by Research and Construction of Comprehensive Analysis and Processing System for Big Data of Multi-source Heterogeneous Constitution Health(2017G006045).

6. References

[1] Sowmya R, Suneetha K R. Data mining with big data[C]. In Proceedings of 11th International Conference on Intelligent Systems and Control (ISCO), 2017, pp.246-250.

[2] Halevy A, Rajaraman A, Ordille J. Data integration: the teenage years[C]. In Proceedings of the 32nd international conference on Very large data bases (VLDB) Endowment, 2006, pp. 9-16.

[3] I. Fellegi, A. Sunter. A theory for record linkage [J]. Journal of the American Statistical Association, 1969,64(328):1183-1210.

[4] M. Hernandez, S. Stolfo. The merge/purge problem for large databases [C]. In Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, 1995, pp.127-138.
[5] H. Garcia-Molina. Pair-wise Entity Resolution: Overview and Challenges [C]. In Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, 2006, pp.1.

[6] O. Benjelloun, H. Garcia-Molina, D. Menestrina, S. E. Whang, Q. Su, J. Widom. Swoosh: A Generic Approach to Entity Resolution [J]. The Very-Large Database Journal (VLDBJ), 2008,18(1):255-276.

[7] http://dbis.uni-trier.de/Mitarbeiter/reuther_files/private/reuther.shtml