Acta Technica *62* (2017), No. 4A, 565–574

Identity recognition of complex entity across heterogeneous sources

Qiang Meng[2], Jianfeng Wang[2,3]

Abstract. As part of the Web big data, information about the same entity may distribute widely in multiple heterogeneous sources, which challenges traditional entity recognition and clustering methods. Aiming to address the characteristics of inconsistency and irrelevance of data across heterogeneous sources, in this paper, we pro-pose a joint iterative method for entity recognition based on object similarity measurement and characteristic relevance analysis. In this work, we first construct a model of non-linear similarity measurement and pro-pose a method of optimizing multidimensional weight parameters for measuring the similarity between objects; then we establish an iterative model to optimize object relevance, expand training set and analyze characteristic relativity. We also propose a method to estimate the weights and parameters concerning un-known characteristic data (they do not appear in training data) for ultimately achieving joint identity recognition on data across heterogeneous sources. We experiment on both homogeneous and heterogeneous datasets and compare with three state-of-the-art methods. The results validate better accuracy and adaptability of our method.

Key words. Entity identity, heterogeneous cross-source, iterative algorithm, complex characteristic, nonlinearity relationship.

1. Introduction

Entity recognition is a critical foundation for data fusion and data cleaning, which directly affects data quality and data analysis. Entity recognition is a process to group the data describing the same real-world entity from one source or multiple sources, as well as a particular clustering issue [1] different from normal clustering [2]. According to the type of objects to be recognized, it is mainly divided into named

[1]Acknowledgement-This research is based upon work supported in part by the National Natural Science Foundation of China (No.61272109, 61502350), in part by the Natural Science Foundation of Hubei Province of China (2014CFB289). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any of the above organizations or any person connected with them.

[2]Workshop 1 - Physical Educaion College of Zhengzhou University, Zhengzhou 450044, China

[3]Corresponding author: Jianfeng Wang

entity recognition [3] (natural language processing) and object entity recognition [4] (structures data). Named entity recognition is to match the identity between two texts via semantic analysis and relationship among lexical structures. Object entity recognition is to estimate the identity between two objects via comparing similarity among the features of entities. The essential of both of the aforementioned methods is to match by means of calculating the similarity between two objects. Apparently the effective similarity calculation are very important for entity recognition. Therefore, entity recognition has an important significance to improve big data applications as well as other research fields epitomized as follows:

(1)Information retrieval. Plenty of repetitive data produced by retrieval results from multiple sources would affect the efficiency of information acquisition. Effectively recognizing and integrating data describing the same entity from multiple sources can improve the effectiveness of homogeneous data fusion and search results diversity.

(2)Data fusion. As the base of data fusion, entity identity recognition matches and merges the entity objects from different sources, thus completing the feature properties as well as improving the fusion effectiveness and data quality.

2. The Identity Judgment on Heterogeneous Entities

2.1. The General Form of Similarity Function for Heterogeneous Entities

We denote two heterogeneous objects as:

$$\begin{aligned} D_1 &= \{< K_{11}, V_{11} >, < K_{12}, V_{12} >, \cdots, < K_{1p}, V_{1p} >\} \\ D_2 &= \{< K_{21}, V_{21} >, < K_{22}, V_{22} >, \cdots, < K_{2q}, V_{2q} >\} \end{aligned} \tag{1}$$

According to the previous work, we equivalently process the features based on similarity and range. The features with equivalence relation from two object data are denoted as one, and the other features are mutually independent. Thus the above two object data can be merged as:

$$\begin{aligned} D_1 = \{&< K_1, V_{11} >, < K_2, V_{12} >, \cdots, < K_n, V_{1p} >, \\ &< K_{11}, V_{11} > \cdots, < K_{1(p-n)}, V_{1(p-n)} >\} \\ D_2 = \{&< K_1, V_{21} >, < K_2, V_{22} >, \cdots, < K_n, V_{2n} >, \\ &< K_{21}, V_{21} > \cdots, < K_{2(p-n)}, V_{2(p-n)} >\} \end{aligned} \tag{2}$$

where n represents the number of features with the same attributes in D_1 and D_2, the attribute K is re-ordered. Considering that different features matter differently in the similarity, we suppose the other features are irrelevant.

Therefore the similarity of D_1 and D_2 can be represented as:

$$Sim(D_1, D_2) = \frac{\sum_{i=1}^{n} \omega_i Sim_V(V_{1i}, V_{2i})}{\sum_{i=1}^{n} \omega_i} \tag{3}$$

where ω_i denotes the weight of feature i, the range of ω_i is [0,1]. For the heterogeneity of data, the data features are different. $\sum_{i=1}^{n} \omega_i$is not always equal to 1. $Sim_V(V_{1i}, V_{2i})$ denotes the similarity of two values on feature i. For the difference of attributes and distributions under heterogeneous environment, the calculations of features similarity are different.

$$Sim_V(V_{1i}, V_{2i}) = \begin{cases} Sim_V(V_{1i}, V_{2i}), (a) \\ \frac{(\frac{Sim_V+u-1}{u})^t+(\frac{1-u}{u})^t}{(\frac{1-u}{u})^t+1}, (b) \end{cases} \tag{4}$$

Optimal Algorithm of Characteristic Weights: We can discover the different importance of the features by studying the training data. We assign different weights to characteristics according to how the features affect the identity. As to the training set TCR, it is generated s equivalent sets of identity as R1,R2,...,Rs. Suppose the data from training set are consistent and effective, our goal is to output the results that arbitrary two objects from the same equivalent set have high similarity, as well as arbitrary two objects from different equivalent set have low similarity. In order to satisfy the weight characteristic parameters of training results, we construct the loss function OPT as follows:

$$OPT = \sum_{X \in R_i, Y \in R_j, i=j} (1 - Sim(X, Y)) + \sum_{X \in R_i, Y \in R_j, i \neq j} Sim(X, Y) \tag{5}$$

When the OPT is smallest, i.e. the similarity of the objects approaches to 1 in the same equivalent set and the similarity of the objects approaches to 0 in different equivalent set, the characteristic weight in this situation is the optimum value. However, the training set may be an uneven sample from huge samples. If we intensify the effect of training data when qualifying the weights, over-fit phenomena occurs. On the one hand, the data of training set perform excellent while the others deviate from our expectancy value. On the other hand, for the optimal function is linear, the data pairs of the similarity are easily generated within the field of 0.5. Then the reliability of recognition is lowered.

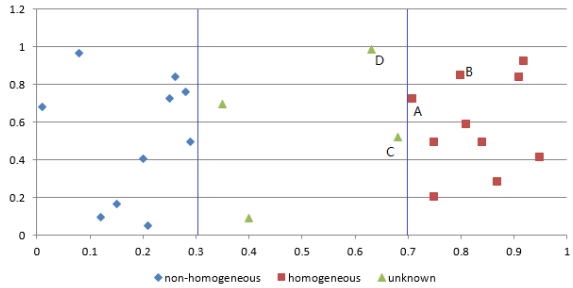

Fig. 1. Distribution of similarity and homoousia of data pairs

According to the specialty of the penalty function, we propose$f(x) = e^{bx} - 1$. When $x > 0$, $f(x)$ increases quickly as x increases. Thus the error values of deviated data are increased, and the precision of optimal function is improved. When$x < 0$,

$f(x) < 0$, $f(x)$ decreases slowly as x decreases. Then the speed is slowed down when close to the frontier, the over-fit is relieved. The object loss function is optimized as:

$$OPT = \sum_{X \in R_i, Y \in R_j, i=j} e^{\alpha(\varepsilon_2 - Sim(X,Y))} - 1 + \sum_{X \in R_i, Y \in R_j, i \neq j} e^{\beta(Sim(X,Y) - \varepsilon_1)} - 1 \tag{6}$$

Therefore, the objective parameters to be solved are $\alpha, \beta, \omega_1, \omega_2, \cdots, \omega_n, u_1, u_2, \cdots, u_n$, $t_1, t_2, \cdots, t_n$etc. We utilize Stochastic Gradient Descent(SGD) to optimize the characteristic weights and calculate the partial derivatives of all parameters. For instance the Gradient descent direction of ω_k is:

$$\begin{aligned} &\Delta\omega_k = \frac{\partial OPT}{\partial \omega_k} = \\ &\sum_{X \in R_i, Y \in R_j, i=j} -\alpha e^{\alpha(\varepsilon_2 - S(X,Y))} \bullet \frac{\partial S(X,Y)}{\partial \omega_k} \\ &+ \sum_{X \in R_i, Y \in R_j, i \neq j} \beta e^{\beta(S(X,Y) - \varepsilon_1)} \bullet \frac{\partial S(X,Y)}{\partial \omega_k} \\ &= \sum_{X \in R_i, Y \in R_j, i=j} -\alpha e^{\alpha(\varepsilon_2 - S(X,Y))} \bullet \frac{S_{Vk} \sum_{i=1}^{n} \omega_i - \sum_{i=1}^{n} \omega_i S_{Vi}}{[\sum_{i=1}^{n} \omega_i]^2} \\ &+ \sum_{X \in R_i, Y \in R_j, i \neq j} \beta e^{\beta(S(X,Y) - \varepsilon_1)} \bullet \frac{S_{Vk} \sum_{i=1}^{n} \omega_i - \sum_{i=1}^{n} \omega_i S_{Vi}}{[\sum_{i=1}^{n} \omega_i]^2} \end{aligned} \tag{7}$$

where $S(X,Y)$ is the objects similarity function expressed by Eq.(13) calculated by Eq.(14) represents the similarity of values on the feature i between object X and Y. Besides, $\sum_{i=1}^{n} \omega_i$is the characteristic intersection for its corresponding data pair(x,y), and ω_i in different (x,y) may be not necessarily identical. For each group of sample pair (x,y) composed of two data, it is iterated along with the direction of gradient descent in the objective function as follows:

$$\omega_k(t+1) \leftarrow \omega_k(t) - \eta(\frac{\partial OPT}{\partial \omega_k}) \tag{8}$$

where $\eta \in \mathrm{Re}^{+}$represents the studying speed, i.e. descent velocity. In our work we utilize Gaussian function to control step length of gradient descent, thus the renewal process is represented as:

$$\omega_k(t+1) = \omega_k(t) - \frac{1}{\sqrt{2\pi}} \times e^{-\Delta\omega_k(t)^2} \times \Delta\omega_k(t) \tag{9}$$

In the initial situation, we set all the weights as 0.1 by default for increasing, set all the u_i as 0.5 by default for decreasing, set all the t_i as 1 by default for decreasing. The training set studies according to Stochastic Gradient-Descent to solve the weights of characteristic attributes and relative parameters. Thus the two objects which satisfy Eq. (19) can be inferred as the identical entity.

$$Sim(D_X, D_Y) = \frac{\sum_{i=1}^{n} \omega_i Sim_V(V_{Xi}, V_{Yi})}{\sum_{i=1}^{n} \omega_i} \geq 1 - \varepsilon \tag{10}$$

3. Experiments

Experimental Settings:Our experiments use real-world datasets to measure the algorithm performance. In order to analyze the entity characteristics and performance variation in different environment, two datasets are collected including:(1)homogeneous datasets from mobile application: constructed by 4 different sources, 20109 records. The data stored by relational database possess the same characteristic. The attributes cover name, content, category, rating, size, version number, download, developer, last update; (2) heterogeneous commodity datasets: constructed by 3 sources, 698921 records. The data stored by Key-Value database possess different characteristics. Besides the common attributes such as name, content, price, seller, category, there are some different attributes such as brand, color, material, features, support phone, weight. On the average, each data possesses 12 attributes. In the complete set, there are 3918 different characteristic attributes in all.

We train the parameters by sample set, and validate the effect by calibration of testing set. In our experiments, as to datasets of mobile application, we randomly select 1000 items as sample set and 2000 items as testing set. As to commodity datasets, we randomly select 1500 items as sample set and 3000 as testing set. The Ground truths of both sample set and testing set are all collected from amazon crowdsourcing analysis and arranged as set $\{<D_{11},D_{12},\ldots>,<D_{21},D_{22},\ldots>,\ldots\}$. Each subset represents the data objects of the identical entity. We adopt complete set for computation. Besides comparing the results from testing set, we randomly select partial sample data to manually annotate for supplementing the ground truths of testing set.

All experiments are performed on a PC cluster with Intel(R) Core(TM) CPU i7-4790 3.60 GHz and 8 GB memory, running on a Windows 2008 operating system. Our algorithms are implemented using the C++ language and each execution is performed as a single process (i.e., no parallel processes), where very minor simplifications are done in the implemented versions.

Evaluation Criteria:As to n items of the testing set, we construct $n \times n$ data pairs $< Di, Dj >$, let 0-1 matrix $TM_{n\times n}$ represent the testing results, where each element represents the identity of data pair calculated by former algorithm; let 0-1 matrix $RM_{n\times n}$ represent the truth, where each element represents the identity of data pair based on ground truth.

Validation and Analysis:Entity recognition is a special type of clustering issue. In order to validate our work, we conduct experiments on homogeneous mobile application datasets and heterogeneous commodity datasets separately, and then compare on the same datasets with: (1) traditional clustering benchmark(Cop-K-means): A semi-supervised clustering algorithm based on pairwise constraints; (2) GAC (Genetic Algorithm-based Clustering Technique); (3) RELDC: entity recognition based on similarity; (4) DepGraph[10]: entity recognition algorithm based on relationship of object dependence.

(1) Performance analysis of homogeneous datasets

In order to measure how training study influences the algorithm performances, remaining the scale of the training set, we sample sizes of 200, 500, 800 and 1000

sets from homogeneous datasets for training. It is indicated in Figure 3, when the training sets are sufficient, RELDC and GAC perform better than DepGraph and Cop-K-means. But as the sets reducing, the performances of RELDC and GAC decrease rapidly. This is because DepGraph analyzes the relationship between characteristics via a few sets and Cop-K-means extends semi-supervised learning via pairwise constraints information, which lowers the demand for training sets. At the same time, RELDC and GAC are in high demand for training sets. The experiments demonstrate when training sets are insufficient, IBJI-MHE performs approximately as DepGraph and when training sets are sufficient, it performs approximately as RELDC.

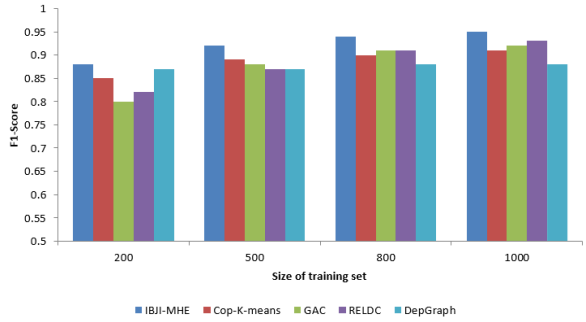

Fig. 2. comparison in different training size (homogeneous)

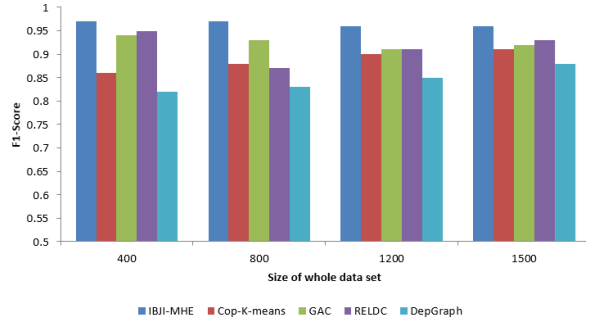

Fig. 3. comparison in different universal size(homogeneous)

In order to compare the algorithm performances in different dataset sizes, we uniform scale the testing sets and training sets, and sample testing sets with the sizes of 400, 800, 1200, 1500 for validation, then geometric diminish the corresponding training sets. It is indicated in Figure 4 that as the dataset increasing, Cop-K-means and DepGraph improve their performances. It is because they utilize the characteristic relation between data and their abilities of characteristic learning are enhanced when the datasets are sufficient. However the performances of RELDC and GAC are degraded for increasing datasets exacerbate discreteness. The experimental results demonstrate IBJI-MHE maintains superior stability and performs excellent regardless of dataset size.

(2) Performance analysis of heterogeneous datasets

In order to measure how training study influences the algorithm performances in heterogeneous datasets, remaining the scale of the training set, we sample sizes of 200, 500, 800 and 1000 sets from heterogeneous datasets for training. It is indicated in Figure 5 RELDC and GAC are strongly dependent on training sets. It is mainly because depending on association graph and characteristic attributes they are unable to effectively process unknown characteristics of heterogeneous data in collections. Cop-K-means performs better than DepGraph under heterogeneous environment. It is because Cop-K-means combines training and testing sets to adjust convergent function while DepGraph is unable to jointly model unknown characteristics. The experimental results demonstrate IBJI-MHE is independent on training sets under heterogeneous environment, and performs generally better than other algorithms.

In order to compare the algorithm performances in different dataset sizes under heterogeneous environment, we uniform scale the testing sets and training sets, and sample testing sets with the sizes of 600, 6000, 2000, 3000 for validation, then geometric diminish the corresponding training sets. It is indicated in Figure 6 that as the dataset increasing, Cop-K-means and DepGraph improve their performances. Cop-K-means improves obviously when dataset is small and then varies little. It is because Cop-K-means enhanced training effect to some extent after learning training and testing sets. While DepGraph improves slowly for it is unable to integrally establish dependency relation graph by computing adjacent objects under heterogeneous environment and the performance improves limitedly. At same time the performances of RELDC and GAC are decreased, especially GAC decreases obviously. It is because as the heterogeneous datasets increasing, the dispersion enhanced sharply and the performance is decreased immensely. The experimental results demonstrate IBJI-MHE maintains superior stability and performs excellent regardless of dataset size.

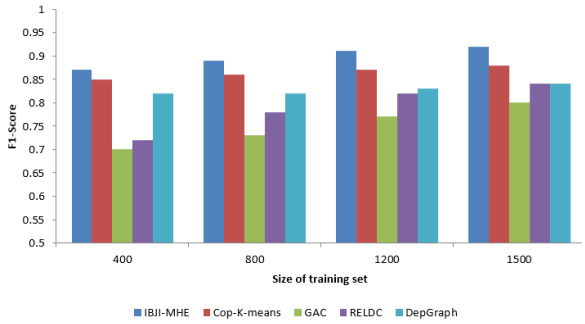

Fig. 4. comparison in different training size(heterogeneous)

4. Conclusion

We have proposed a joint iterative method of entity recognition for heterogeneous complex data based on measurement of objects similarity and analysis of character-

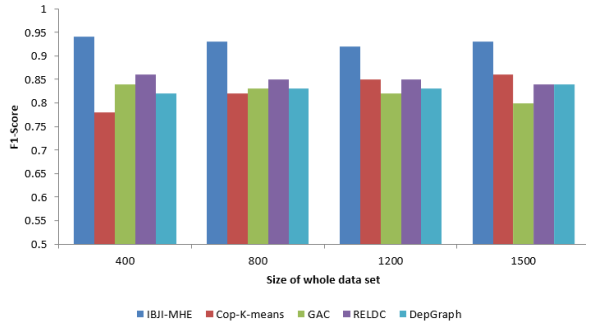

Fig. 5. comparison in different whole size(heterogeneous)

istic relevance. The experimental results have demonstrated our techniques outperform other methods on both accuracy and adaptability. Our contributions mainly are:

(1) We propose a measurement of similarity between non-linear description data.

(2)We present objective loss function to optimize global characteristic parameters.

(3)We construct a combined iterative associated solution.

These methods have effectively solved entity recognition of complex data under heterogeneous environment. However, both theory and experiments of our works are conducted based on data with no obvious conflicts. We only consider situations of data missing, multiple-valued attributes and polymorphism description but ignore data errors, data conflicts and strong inconsistency. In the future, we would like to investigate (??) entity recognition on inconsistency data based related workand (??) performance optimism in big data to improve adaptability and efficiency.

References

[1] Bouveyron. C,Brunet-Saumard. C: *Model-based clustering of high-dimensional data: A review.*Computa-tional Statistics & Data Analysis *71* (2013),52–78.

[2] Montalvo. S, Martínez. R, Casillas. A,Fresno. V: *Multilingual news clustering: Feature translation vs. identification of cognate named entities.* Pattern Recogni-tion Letters *28* (2007), No. 16, 2305–2311.

[3] Vavliakis. K. N, Symeonidis. A. L,Mitkas. P. A: *Event identification in web social media through named entity recognition and topic modeling.*Data & Knowledge Engineering *88* (2013),1–24.

[4] Elmagarmid. A. K, Ipeirotis. P. G,Verykios. V. S: *Duplicate record detection: A survey.*IEEE Transactions on Knowledge and Data Engineering *19* (2017), No. 1, 1–16.

[5] Thor. A, Rahm. E: *MOMA-A Mapping-based Object Matching System.* In Proc. 3th Biennial Conference on Innovative Data Systems Research (2007),247–258.

[6] Benjelloun. O,Garcia-Molina. H, Menestrina. D,: *Swoosh: a generic approach to entity resolution. The VLDB Journal.*The International Journal on Very Large Data Bases *18* (2009) , No. 1,255–276.

[7] Singla. P, Domingos. P: *Entity resolution with markov logic.*In Proc. 6th IEEE International Conference on Data Mining(2006),572–582.

[8] Bhattacharya. I, Getoor. L: *Iterative record linkage for cleaning and integration.* In Proc. 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. ACM (2014),11–18.

Received November 16, 2016