Mixed Attributes Two-Stage-Clustering Entity Resolution

LEI Gang¹,²
¹School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330013, China
²Software College, Jiangxi Normal University, Nanchang 330022, China

Abstract: Record matching and clustering are two essential steps in the process of entity resolution, and the single text similarity clustering based on tf-idf (term frequency-inverse document frequency) feature often leads to poor precision in spots entity resolution. The paper outlines a mixed attributes two-stage-clustering entity resolution framework (abbreviated in MATC-ER) and designs an approach to measure the similarity by mixing spot name and spot introduction, which makes good use of the record information at different stages. Then the paper proves its efficiency based on the comparative experiments on the real data of travel spots.

Key words: Mixed attributes, two-stage clustering, entity resolution, spots entity.

1. Introduction

With the spring of travel business and the spread of travel information, varieties of websites about e-business, review on travel, strategies of travel and web portals about travel have mushroomed. Web data about travel are increasing rapidly. Therefore researches about information integration are hot and entity resolution is a key step toward information integration [1]. There are different descriptions about the same travel spots on different websites. So it is necessary to comprehensively analyze the spot name and spot introduction. In the process of spot entity resolution, term frequency is sometimes difficult to distinguish spots information caused by following factors—the same entity with different descriptions and the same description with different entities. The traditional way of text similarity based on tf-idf feature [2] easily leads to false negative (the same spot with the different entity) and false positive (the different spot with the same entity).

 Different points of view often have different forms of the same thing, that is, comprehensive understanding of things has more granularity through the multiple angles of observation and analysis in order to achieve. With regard to the deficiency of the traditional approach, we put forwards a mixed attributes two-stage-clustering entity resolution framework (abbreviated in MATC-ER). The MATC-ER method makes greatest use of the relevant information in the two stages by extracting the characteristic information from the different travel data and their different attributes.

2. Related Work

Record matching and clustering are two essential steps in the process of entity resolution, which have been introduced systematically in Ref. [3]. The paper focuses to extract features from text and utilize them effectively in the process of two-stage-clustering. The following is related work.

Venkatesan et al. [4] present a system for linking a given text document with relevant structure data by viewing the structured data as a predefined set of entities in the text document. Anitha et al. [5] present a method matching unstructured product offer to structured product specifications by making optimal parse of (unstructured) offer with regard to
specification. Dalvi et al. [6] use a generative language model to match reviews to objects. Hanna et al. [7] present a method of tailoring entity resolution for matching unstructured product offers by extracting product code. Michelson and Knoblock [8] create attribute reference set to extract attributes from text document to matching documents. Wu et al. [9] match unstructured product offers by mining the associations of hidden attributes as the signatures of objects in the unstructured data. All of above include two kinds of record matching including structured to unstructured, and unstructured to unstructured. While the paper focuses to match structured to structured using unstructured features.

There exist two kinds of research [10-16] about two-stage clustering algorithm for matching record: (1) First stage clustering is only partial clustering based on strong features, and the exact results of the first stage are used to extract extended features, which are used to match the rest records of the first stage in the second stage clustering [10-14]; (2) First stage clustering is preliminary, and the results of the first stage are used to extract extended features, which are used to exactly match all records in the basis of the first stage in the second stage clustering [15, 16]. The paper belongs to the latter.

3. MATC-ER Method

There are three concepts in MATC-ER: spot entity, spot record and spot entity cluster.

Definition 1: spot entity: means spot in reality and tells one from the other.

Definition 2: spot record: means the descriptions about the spot in the different websites, including several attributes, such as spot name and spot introduction.

Definition 3: spot entity cluster: means the classification of spot record. The ideal cluster would be the following: (1) The same spot entity record belongs to a same cluster; (2) The same spot entity cluster exclusively covers spot records of a spot entity.

There are two stages in MATC-ER method:
(1) First-clustering stage: the TF-IDF feature of nouns in spot introduction of spot record will be extracted and cosine similarity will be computerized, then K-Means clustering will be developed;
(2) Second-clustering stage: bucket clustering will be developed by computerizing jaccard similarity of spot name, and person nouns and place nouns of spot introduction.

3.1 First-Clustering Stage

The first-clustering aims at spot entity cluster in order to realize preliminary spot entity resolution. The paper adopts an improved K-Means algorithm which combines Initial Clustering Center Selection Based on Maximum Distance [17] and Evaluation Function Based on Silhouette Coefficient [18]. The algorithm is efficient to set up initial clustering center and cluster number. In this way, the clustering effectiveness will be better than random method and trial method. The algorithm is described like the following:

Input: N sample documents
Output: clustering result having the best cluster number

Algorithm:
(1) For \( i = 2 \) to \( N \)
(1.1) confirm \( i \) initial clustering centers by Initial clustering center selection algorithm;
(1.2) distribute all documents to the nearest clusters;
(1.3) update the mean of all clusters, and let them new clustering centers;
(1.4) repeat (1.2) to (1.3), until convergency;
(1.5) compute the silhouette coefficient \( s \) of \( i \) clusters;
(2) find the cluster number \( i \) responding to the biggest \( s \).

3.2 Second-Clustering Stage

3.2.1 Spot Mixed Attributes Similarity

(1) Spot name attribute similarity
Though different webs name spots differently, the theme-topic words in similar spot entity are often similar to each other. For example, the theme-topic words Zi zhu yuan Park and Qing liang gu appear respectively in the spot names Beijing Zi zhu yuan Park and Qing liang gu Travel Spot. So the similar spot names are the significant clues to tell whether they are in the same spot entity or not. The computerization of the exploring jaccard similarity between A spot name and B spot name is as follows:

$$s_n(A, B) = \frac{|T_{An} \cap T_{Bn}|}{|T_{An} \cup T_{Bn}|}$$  \hspace{1cm} (1)

Where, $T_{An}$ are the terms of A spot name, and $T_{Bn}$ are the terms of B spot name.

(2) Spot introduction attribute similarity

There is an intuitive idea coming from the manual observation to a large quantity of spot introduction texts, that is, place nouns and person nouns in the spot introduction text make a great distinction among spots, which means that place nouns and person nouns in the introduction to the similar spot entity will align with each other and vice versa. In terms of this feature, the jaccard similarity measurement between A spot introduction and B spot introduction is defined as follows:

$$s_i(A, B) = \frac{|T_{nh_{Ai}} \cap T_{nh_{Bi}}| + |T_{ns_{Ai}} \cap T_{ns_{Bi}}|}{|T_{nh_{Ai}} \cup T_{nh_{Bi}}| + |T_{ns_{Ai}} \cup T_{ns_{Bi}}|}$$  \hspace{1cm} (2)

Where, $T_{nh_{Ai}}$ are the terms of person noun of A spot introduction; $T_{ns_{Ai}}$ are the terms of place noun of A spot introduction.

(3) Mixed attributes similarity

The spot record similarity will be computed by weighting and summing spot name attribute similarity and spot introduction attribute similarity. The equation is as follows:

$$s_s(A, B) = \omega_1 * s_n(A, B) + \omega_2 * s_i(A, B)$$  \hspace{1cm} (3)

Where, $\omega_1$ is the weight coefficient of spot name attribute similarity, and $\omega_2$ is the weight coefficient of spot introduction attribute similarity. They are trained by the average perception algorithm [19].

3.2.2 Bucket Clustering Algorithm

Bucket algorithm [20, 21] means putting the spot records of each spot entity into the corresponding buckets. Hence the spot entity resolution can be realized precisely. Bucket algorithm is described like the following:

Input: The first-clustering records
Output: The records in the buckets
Algorithm:
(1) create an empty stack;
(2) put first record to stack, and let it to be the typical record of a new bucket;
(3) choose next record, do following operations:
(3.1) Repeat
(3.1.1) successively match with the record of stack;
(3.1.2) If similarity is greater than threshold Then match with all records of bucket responding to the typical record; If all similarities are greater than threshold Then put the record on this bucket, break;
(3.1.3) Until match with all records of stack;
(3.2) If the record has not been put into any bucket then put it on stack, and let it to be the typical record of a new bucket;
(4) repeat (3), until all record to have been searched.

4. Experiments

4.1 Database and Background

The database derives from the spots introduced in the xinxin travel website (the most powerful travel e-business website in China, http://www.cncn.com) and China travel website (the most powerful website, http://www.zgly.cn). The total number of spot records amounts to 32,059. The data is collected by using the software MetaSeeker1, developed by GooSeeker. The spot records are extracted in text and stored in XML form. The information about the spot names and the

Mixed Attributes Two-stage-clustering Entity Resolution

<table>
<thead>
<tr>
<th>Method</th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
<th>Precision</th>
<th>Recall</th>
<th>F-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATC-ER</td>
<td>0.46</td>
<td>0.54</td>
<td>0.995</td>
<td>0.895</td>
<td>0.943</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.16</td>
<td>0.84</td>
<td>0.949</td>
<td>0.904</td>
<td>0.926</td>
</tr>
</tbody>
</table>

Fig. 1 Comparison of precision.

Table 1 Comparison of experiment results.

spot introductions is extracted into the database. The LTP\(^2\) platform, developed by HIT (Harbin Institute of Technology, China), is employed to make Chinese word segmentation for text.

The equipments on which experiments were run are Intel Core i7-4700MQ 2.40GHz(CPU), 16GB (memory capacity), 1TB(hard drive capacity), Microsoft Windows 8.1 professional(OS).

4.2 Assessment and Baseline

(1) Assessment

The precision tops the list of standards of assessing the experiment of the entity resolution. Then F-value which combined precision and recall, will be taken into consideration. The standards are listed as follows:

Precision = correctly matched record pairs/matched record pairs;

Recall = correctly matched record pairs /actually matched record pairs;

F-value = (precision*recall*2)/(precision + recall).

(2) Baseline experiment adopts single stage clustering (bucket algorithm), and hybrid similarity including jaccard similarity of the spot name attribute and cosine similarity of the spot introduction attribute covering all the nouns.

4.3 Parameters

There are several parameters in the paper such as K-Means clustering parameter, attribute similarity weight, and similarity threshold of bucket.

Among them, the similarity threshold of bucket is defined as 0.4 according to trail method.

4.3.1 K-Means Clustering Parameter

(1) Numbers of cluster are 352. Majorities of similar spot records will be clustered in the same cluster and only minorities of similar records will be mistakenly clustered in the different clusters. In the process of the second-clustering, there are few interferences.

(2) Termination threshold is 0.99. Though the mean number changes subtly, the clustering results are not influenced distinctly, which can be considered as steady clustering.

4.3.2 Attribute Similarity Weight

The two groups of mixed attributes are used to learn the average perception, the weight are proved efficient

Mixed Attributes Two-stage-clustering Entity Resolution

(1) For the spot name, the place nouns and person nouns of the spot introduction, $\omega_1$ is 0.46 and $\omega_2$ is 0.54;

(2) For the spot name and all nouns of the spot introduction, $\omega_1$ is 0.16 and $\omega_2$ is 0.84.

4.4 Results

Fig. 1 shows that the precision of MATC-ER method is higher than baseline method, because the former has considered many fine granularity features in spot similarity and made full use of them by two stage clustering. Fig. 2 shows that the recall of MATC-ER method is slightly lower than baseline method, because the first clustering in the former has caused the omission of matched records. Fig. 3 shows that the F-value of MATC-ER method is higher than baseline method, which proves the MATC-ER method is more effective.

5. Conclusions

The paper outlines the MATC-ER framework and designs an approach to measure the similarity by mixing spot name and spot introduction, which makes good use of the entity information at different stages. Then the paper proves its efficiency based on the comparative experiments on the real data of travel spots. In the future research, the spot information from Webs and encyclopedia can be used to extend the features of the first-clustering for the sake of the more
efficient second-clustering.

Acknowledgement

This work was financially supported by the Jiangxi Province Postgraduate Innovation Foundation in 2012(YC2012-B021).

References


