A New Error Correction Algorithm for Chinese Placename

Ou Ye, Jing Zhang, Junhuai Li
School of Computer Science and Engineering, Xi’an University of Technology,
Xi’an 710048, P.R.China
E-mail:785669070@qq.com

Abstract—**In order to improve the accuracy of error correction for Chinese placename, the cleansing framework based on the address-name index table and bi-direction placename error correction method are proposed. Firstly, the error correction algorithms in data cleansing area are reviewed. Secondly, the cleansing framework based on the address-name index table is proposed for the features of Chinese placename and the low accuracy of existing algorithms. In this framework, the structure of address-name index table and relevant concepts are defined, and a new bi-direction placename error correction method is used to improve the accuracy of algorithm by using the address-name index table. In this method, the concepts of partial similarity and whole similarity are introduced, and Chinese placename are matched and corrected in forward firstly, then they are searched from address-name index table and corrected in reverse. Finally, extensive simulation experiments are conducted to prove the feasibility and rationality of method, the results of which show that the bi-direction placename error correction method is better than the others in the performance metrics of implementation precision ratio and recall ratio for error correction of Chinese placename.**

Index Terms—**address-name index table, error correction, data cleansing; bi-direction placename error correction method, Chinese placename**

I. Introduction

Chinese placename is the name that can flag the geographic elements of China, such as “××Province ××City×××County××Village”. The error correction of Chinese placename is to correct the non-word-error and real-word-error Chinese placename. Among that, non-word-error is also named word- misspelled. It means that the word which is split by word boundary in the text data, and it not exists in dictionary. Non-word-error mainly includes insert-error, delete-error, substitute-error, and others. Real-word-error means that the word which is generated by inattention and other human factors, and the word is still existed in the dictionary, but there is grammar mistake for this word in the sentence, such as translocation word, ect[1]. Because the error Chinese placename belongs to the dirty data (the inconsistency or inaccurate data, old data and wrong data[2]), and the technologies of data cleansing are mainly used to solve the problem of dirty data [3-5] currently, we can use the technologies of data cleansing to correct the error Chinese placename. At present, there are some results of error correction mainly to be reflected in following respects: 1) Optimization algorithms of error correction [6-10]; 2) The application of algorithms for error correction [11-17], ect. These results are mainly used to detect and correct error data in the English text by using the statistical algorithms [18-19], clustering algorithms [20], rules-based algorithms [21] and others. Among that, it also contains the error correct of placename.

However, English word can directly reflect semantic meaning itself, and detection and revision for the English words should be primarily considered. But there are great differences between Chinese and English language. For example, the Chinese placename “shannxisheng” is corresponding to the English placename “Shannxi Province”. It shows two differences: 1) Chinese placename have not any separators, but English placename have; 2) Chinese placename have the homophonous words, but English placename not have. Recently, there are less researches on error correction of Chinese placename, and there are some shortcomings of existing algorithms as follows: 1) currently, the algorithms most pay attention to the error detection, and ignore the error correction, it will cause some error data can not be automatically detected and corrected, or correcting suggestions can not be provided; 2) there are less researches of error correction for placename, person name and other data.

Our motivation in studying the problem is to propose a cleansing framework based on the address-name index table and a bi-direction placename error correction method. The cleansing framework based on the address-name index table can appropriate for the error correction of Chinese placename. In this cleansing framework, the bi-direction placename error correction method is introduced to improve the accuracy of existing error correction algorithms. Among that, the partial similarity calculation and whole similarity calculation are

Manuscript received July 5, 2012; revised Match 22, 2013; accepted April 16, 2013.
Corresponding author: Ou Ye; email: 785669070@qq.com;
This research was supported in part by the National Natural Science Foundation of China under Grant (No. 61172018), ect.

© 2013 ACADEMY PUBLISHER
doi:10.4304/jsw.8.10.2397-2405

introduced into the bi-direction error correction method to calculate the similarity between Chinese placename.

Ⅱ. Related Work

The detection and error correction of error data mainly includes two aspects: 1) error correction of misspelling-word; 2) error correction of context-dependent. Due to this paper mainly research the error correction of Chinese placename, and misspelling-word often appears, this paper mainly reviews the existing algorithms of misspelling-word. At present, there are mainly two error correction methods of misspelling-word: 1) the statistics method; 2) the detection method based on dictionary. In this section, there are two common algorithms of error correction will be reviewed.

(1)Minimum Edit Distance Algorithm

Minimum edit distance algorithm [6, 9] is a common error correction algorithm. This algorithm is based on letters to matching two words, and to calculate the minimum number of edit operations (i.e., insertions, deletions, and substitutions) of single characters needed to transform one string to another, then to calculate the similarity between two words that are in the dictionary, finally to determine candidate words of error correction and to provide the correcting suggestions. The larger the minimum number of edit operations, the smaller the similarity between two words, and the more likely the word is the wrong data.

The advantage of this algorithm is that insertion data, deletion data, and substitution data can be detected and corrected by this way, and it has high precision and recall ratio. However, the semantic relation between two words is not considered, and this algorithm only is to replace the word in the word list or dictionary, it is difficult to detect and correct translocation word. Then this algorithm is based on English letters or Chinese characters to match, but Chinese characters don't all have semantic meaning, so the accuracy of matching will down. Finally, the time complexity of minimum edit distance algorithm is O(m*n), so this algorithm has high computational complexity.

(2)N-gram algorithm

N-gram algorithm [13, 22] is a common algorithm for error correction, and it is based on statistics method. At present, N-gram (N=2) model and N-gram (N=3) model are often used to detect and correct the error data.

The advantage of this algorithm is that the way is convenient and easily to use, and it has high recall ratio. However, because of the limitation for corpus, the effect of error correction for placename, person name and other proper-noun data is poor, the precision ratio is low. And this algorithm needs to through the statistics for large-scale text to obtain the transition probability matrix, so the running time of this algorithm is high.

III. A Cleansing Framework based on the Address-name Index Table

A.Address-name Index Table and Relevant Concept

Definition 1 Geographical Area: the name which is confirmed by combining the geographic space of area with demarcation rules of national administrative areas. Geographical area is represented as Ar.

If A area is bigger than B, then Ar(A)＞Ar(B).

Definition 2 Placename Domain: it is a full name that contains several geographical areas Ar(A) 、Ar(B)、…and Ar(N), placename domain is represented as L. The ith placename domain L(i) = Ar(A)Ar(B)…Ar(N), and Ar(A) ⊇ Ar(B) ⊇…⊇Ar(N).

Definition 3 N-Class Domain Value: it is the sequence that is corresponding to the geographi- cal area Ar(i), and Ar(i) is included the placename domain L(i). N-Class Domain Value is represented as NAr, and 0≤N(Ar)≤N.

Definition 4 Address-name Index Table: it is a table that contains all correct geographical areas of same class and other relevant information. Address-name index table is represented as GT.

Definition 5 Relevant Information of Placename: it contains all information between geographical area and placename domain. Relevant information of placename is represented as ADr, and ADr(i)⊇Ar(i).

Corollary 1 In the placename domain, the lower the sequence of index for geographical area, the smaller the N of N-Class domain value, and the bigger the geographic space of geographical area.

Corollary 2 In the placename domain, the lower the sequence for geographical area, the smaller the N of N-Class domain value, and the smaller the probability of geographical area repeated appears.

Corollary 3 For any placename domain L(i) = Ar(A)Ar(B)…Ar(N), according to the sequence of geographical area, the lower the sequence, the bigger the geographical area, and NAr (A) ＜NAr (B) ＜…＜NAr (N), Ar(A) ⊇ Ar(B) ⊇…⊇Ar(N).

Corollary 4 For any Address-name index table GT(j), there is at least one superior address-name index table GT(i) or at least one junior address-name index table GT(k) to corresponding with GT(j). Among that, Ar(i)＞Ar(j) ＞Ar(k).

Corollary 5 The number of the category for geographical area is same as the number of address-name index table.

B. The Data Structure of Address-name Index Table

From the definition 4, we can see that all the geographical area of same category and other relevant information are included in the address-name index table. Among that, the relevant information mainly contains abbrevia- tion and pinyin, ect. Address-name index table is closely related to the category of geographical area and demarcation rules of national administrative areas. The data structure of address-name index table is shown in figure 1:

© 2013 ACADEMY PUBLISHER

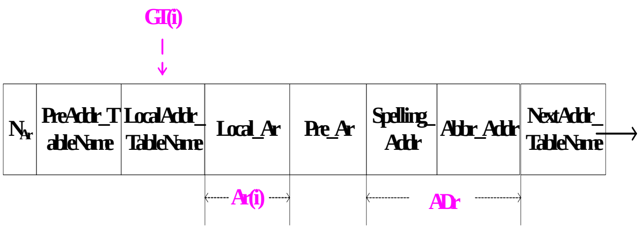


Figure 1.The data structure of address-name index table

The meaning of each field in the figure 1 is shown below:

NAr: The sequence that is corresponding to the geographical area Ar(i), and Ar(i) is included the placename domain L(i).

PreAddr_TableName: The superior name of address-name index table, such as GT(i-1). If GT(i-1) is not exist , it is null.

LocalAddr_TableName: the name of address -name index table, such as GT(i), it includes the geographical area Ar(i).

Local_Ar: the geographical area Ar(i) that is included in the standard(correct) placename.

Pre_Ar: the superior geographical area Ar(i-1).

Spelling_Addr: pinyin information of the geographical area Ar(i).

Abbr_Addr: the abbreviation of the geographical area Ar(i).

NextAddr_TableName: the junior name of address-name index table, such as GT(i+1). If GT(i+1) is not exist ,it is null.

Among that, Abbr_Addr and Spelling_Addr are included in the relevant information of placename ADr(i), and some information can be inserted or deleted in the ADr(i).

Any placename all is corresponding to a placename domain L(i), and there are relationship between geographical area Ar(i) and address- name index table GT(i). The relation -ship is shown in figure 2.

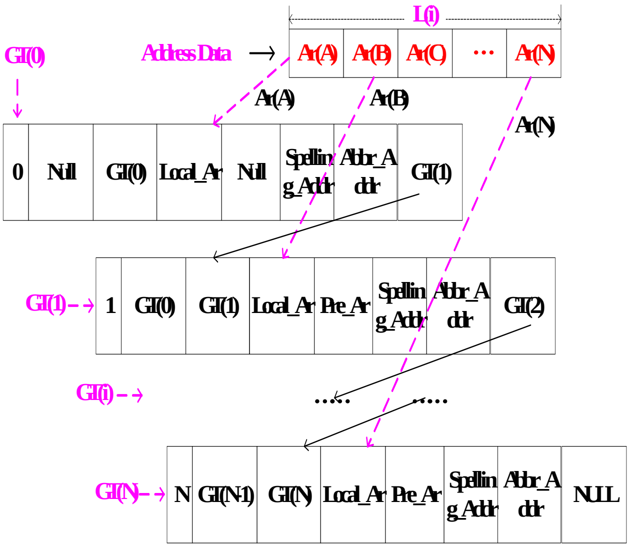


Figure 2.The relationship between placename data and address-name index table

From the figure 2, it can be seen that a placename is corresponding to the placename domain L(i), and L(i) = Ar(A)Ar(B)…Ar(N). These geographical areas are gradually sorted by areas of the geographic space and demarcation rules of national administrative areas, such as “Country”, “Province”, “City”, ect. Any one Ar(i) is corresponding to a address-name index table, and it is included in this table. Also there is a relationship between address-name index table, the junior Ar(i+1) can be searched in the junior address-name index table by the field “NextAddr_TableName”, and the superior geographical area Ar(i-1) also can be searched by the field “PreAddr_TableName”. So the superior or junior geographical area can be searched by Ar(i), and address-name index table GT(i) is corresponding to one superior or junior address-name index table, too, it is useful to search the placename accurately.

C. A Cleansing Framework based on the Address-name Index Table

Two problems need to be solved during the data cleansing for Chinese placename. One is non-word-error; another is real-word-error. For example, a placename “Xi’anCityshenmu CountymengVillage” has the real-word-error. In fact, “shenmuCounty” is not in the “Xi’anCity”, these errors are difficult to distinguish. For the error correction, there are homophones, nearly pronunciation words and other factors may cause the error. Due to the different habits of the abbreviation, the abbreviation has the variety and instability. Therefore, it is difficult to deal with incorrect abbreviation data. In a word, many factors are needed to be considered with detection and error correction for Chinese placename. At present, the existing algorithms and strategies are difficult to detect and correct the incorrect Chinese placename comprehensively. Therefore, in order to detect and correct the wrong Chinese placename more accurately, this paper proposes a cleansing framework based on the address-name index table. The cleansing framework is illustrated in figure.3.

It can be seen from the figure 3 that the framework of error correction and cleansing for Chinese placename mainly contains five steps.

(1) Input data: The test data (they are stored in the Excel document initially) finally are stored in the database by the format detection step and inserting pinyin step. Among that, field names are detected so that to avoid the confused data in the format detection step.

(2)Data normalization: Chinese placename is normalized in this step: 1) the stop words must be deleted, such as “and”, “of” and others; 2) to split the digits and characters of placename, such as placename “XStreetNo56” can be normalized to the “X street/56”; 3) to normalized the placename. For example, “Shanghai” should be normalized to “Shanghai City”.

(3)Data detection of schema layer: The structural conflict of field and restriction of property are detected in this step. Such as conflict of data structure, conflict of keywords and others.

(4)Data detection of instance layer: Chinese placename are detected and corrected in this step. Firstly, test data match with the standard data. These standard data are all the names of geographical areas, and they are stored in the address-name index table. Then homophones are selected by pinyin field between two Chinese placename. If pinyin are same, two Chinese placename may be matched successfully. Among that, if two Chinese placename are different, homophone needs to be corrected by standard data. And if pinyin field is not same, the bi-direction placename error correction method is used to calculate the similarity value, and according to the similarity to judge whether two placename have the similarity. If two placename satisfy the similarity condition, Chinese placename in test data need to be corrected by using standard placename. Otherwise, the suggestion of error correction will be provided to users.

(5)Output data: Error correction Logs that contain the revised Chinese placename and suggestions are provided to the users.

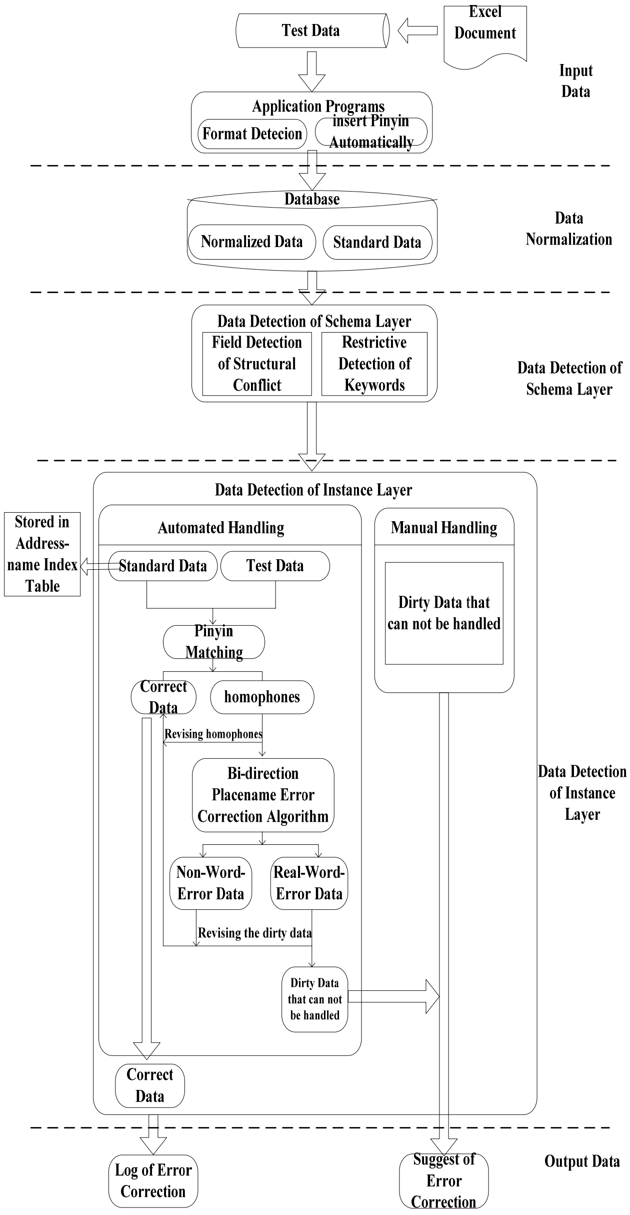

Test Data
Excel Document
Input Data
Application Programs
Format Detecion
insert Pinyin Automatically
Database
Normalized Data
Standard Data
Data Normalization
Data Detection of Schema Layer
Field Detection of Structural Conflict
Restrictive Detection of Keywords
Data Detection of Schema Layer
Data Detection of Instance Layer
Stored in Address-name Index Table
Automated Handling
Manual Handling
Standard Data
Test Data
Pinyin Matching
Correct Data
homophones
Revising homophones
Bi-direction Placename Error Correction Algorithm
Non-Word-Error Data
Real-Word-Error Data
Revising the dirty data
Dirty Data that can not be handled
Dirty Data that can not be handled
Data Detection of Instance Layer
Correct Data
Log of Error Correction
Suggest of Error Correction
Output Data

Figure 3.The cleansing framework based on the address-name Index Table

IV. The Bi-direction Placename Error Correction Method

The main idea of the bi-direction error correction method on Chinese placename is that: firstly, by the partitions of geographical units, geographical units are matched and corrected by the partial similarity calculation in forward, and then all the geographical units are indexed and corrected by the partial similarity calculation for reverse; Finally, two results of forward and reverse are combined to construct the correct Chinese placename by using the whole similarity calculation. The example of main idea is shown in figure 4.

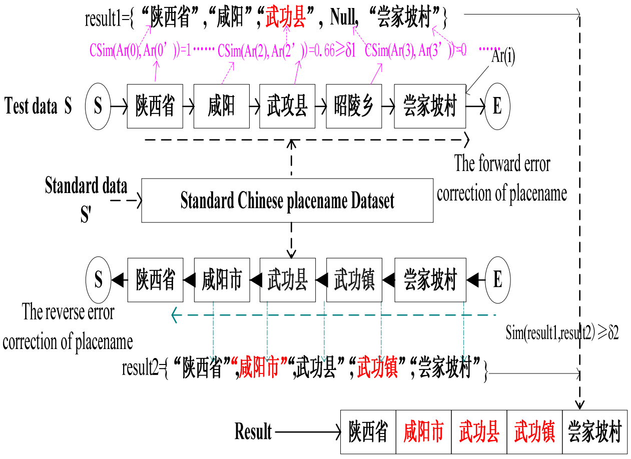

result1={ “陕西省”,“咸阳”,“武功县”, Null, “尝家坡村”}
CSim(Ar(0), Ar(0’))=1 ······ CSim(Ar(2), Ar(2’))=0. 66≥δ1 CSim(Ar(3), Ar(3’))=0 ······
Ar(i)
Test data S
S
陕西省
咸阳
武攻县
昭陵乡
尝家坡村
E
The forward error correction of placename
Standard data S'
Standard Chinese placename Dataset
S
陕西省
咸阳市
武功县
武功镇
尝家坡村
E
The reverse error correction of placename
Sim(result1,result2)≥δ2
result2={ “陕西省”,“咸阳市”,“武功县”,“武功镇”,“尝家坡村”}
Result
陕西省
咸阳市
武功县
武功镇
尝家坡村

Figure 4.The main idea of bi-direction error correction method

From the figure 4, it can be seen that Chinese placename are detected in forward and reverse, and according to the standard placename and index ways in the address-name index table to calculate the partial similarity and whole similarity. After that, similarity will be calculated. If the similarity value is more than the given threshold, Chinese placename can be corrected to the standard placename, otherwise the suggestion of error correction will be provided to users. The specific methods to achieve as following steps:

Step 1 Data input: to read the normalized Chinese placename from the database;

Step 2 Segmentation of Chinese placename: according to the identification symbols of national administrative areas (such as “City”, “County” and others), Chinese placename domain L(i) can be obtained by Chinese-word- segmentation algorithm. Finally, these data can be converted to the L(i) = Ar(A)Ar(B)…Ar(N), and Ar(A) $\supseteq$ Ar(B) $\supseteq$…$\supseteq$Ar(N);

Step 3 The forward error correction of Chinese placename: first of all, the geographical area Ar(i) in placename domain L(i) is to be read. Then according to the N-class domain value NAr(i), geographical area Ar(j) that is similar as the Ar(i) in address-name index table GT(i) is to be searched. Among that, address-name index table GT(i) is corresponding to the NAr(i), and Ar(j) is included in the standard placename and the address-name index table GT(i). After that, the similarity algorithm CSim(Ar(i), Ar(j)) is used to calculate the similarity value of geographical areas. Among that, CSim(Ar(i), Ar(j)) is the partial similarity. If CSim(Ar(i), Ar(j))≥δ1(δ1 is a

© 2013 ACADEMY PUBLISHER

given threshold), it shows that Ar(i) is similar to Ar(j), and the Ar(i) can be reivsed to Ar(j). Then Ar(i) is to be revised to Ar(j), and Ar(j) is stored in the arraylist LRAddr[]. If the result of searching is null, then the value “null” is stored in the arraylist LRAddr[]. Finally, after all the geographical areas in placename domain L(i) are searched, the arraylist LRAddr[] can be obtained. In the arraylist LRAddr[], all elements are correspond -ing to the geographical areas, and the value is Ar or null;

Step 4 The reverse error correction of placename: In the arraylist LRAddr[], from the last element to search the geographical areas in address-name index table reversely, and the element must be not null. Then the superior geographical areas can be obtained constantly until NAr = 0, and they are not null. Finally, the search results are stored in the arraylist RLAddr[]. All the non-null elements of arraylist RLAddr[] are standard and correct geographical areas;

Step 5 similarity calculation and bi-direction error correction: First of all, the elements of LRAddr[] are used to match with the elements of RLAddr[] reversely, and the similarity algorithm SSim(LRAddr[], RLAddr[])is used to calculate the similarity value of Chinese placename. Among that, SSim(LRAddr[], RLAddr[]) is the whole similarity. If SSim (LRAddr[], RLAddr[])≥δ2(δ2 is the given threshold), it shows that the Chinese placename in test data is similar as the standard placename (correct data), then it can be substituted for standard placename. Otherwise, the non-null elements of LRAddr[] and RLAddr[] are used for the suggestion of error correction to provide the users;

Step 6 The next Chinese placename is read, and then begin to jump to Step 1;

Step 7 Until the last Chinese placename is read, the test is finished.

The figure 5 describes the process of bi-direction placename error correction method. As shown as figure.5, Due to Chinese characters of placename can not reflect the semantic meaning itself directly, and there are not any separators between Chinese words, the Chinese placename need to be segmented to the several geographical areas by Chinese-word-segmenta- tion algorithm. So the geographical areas Ar and placename domain L(i) can be obtained. On this basis, for any one geographical area Ar(j), the forward error correction method can be used to match with the similar geographical area Ar(j’). Among that, Ar(j’) is the standard(correct) geographical areas in the address-name index table. If Ar(j’) exists and not null, it shows that Ar(j) is corresponding to the correct geographi -cal area Ar(j’), then Ar(j) is substituted for Ar(j’) and updated it, the junior geographical area need to be searched in address-name index table constantly until the last geographical area of L(i). By this way, the results can be obtained and stored in the arraylist LRAddr[].

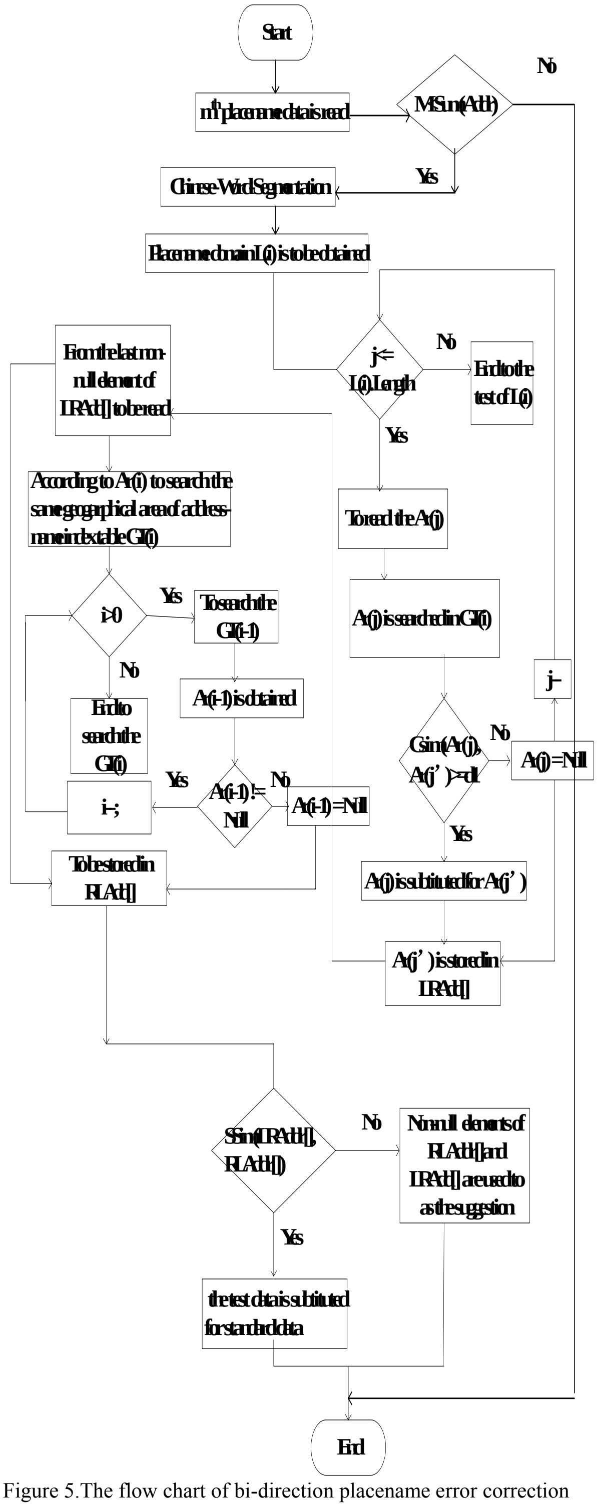

Start
mth placename data is read
MSum(Addr)
No
Yes
Chinese-Word-Segmentation
Placename domain L(i) is to be obtained
j<= L(i).Length
No
End to the test of L(i)
Yes
From the last non-null element of LRAddr[] to be read
According to Ar(i) to search the same geographical area of address-name index table GT(i)
To read the Ar(j)
Ar(j) is searched in GT(i)
i>0
Yes
To search the GT(i-1)
No
End to search the GT(i)
Ar(i-1) is obtained
j--
Csim(Ar(j), Ar(j’))>=d1
No
Ar(j)=Null
Yes
Ar(i-1)!=Null
No
i--;
Ar(i-1)=Null
Yes
To be stored in RLAddr[]
Ar(j) is subtituted for Ar(j’)
Ar(j’) is stored in LRAddr[]
SSim(LRAddr[], RLAddr[])
No
Non-null elements of RLAddr[] and LRAddr[] are used to as the suggestion
Yes
the test data is subtituted for standard data
End

Figure 5.The flow chart of bi-direction placename error correction method

After that, according to the non-null elements of LRAddr[], the geographical areas will be searched reversely in address-name index table, and the results are stored in the RLAddr[]. Finally, similarity value of the elements in LRAddr[] and RLAddr[] are calculated, and according to the relation of similarity value and threshold to judge the similarity. If the similarity value is more than threshold, Chinese placename of test data is revised to standard placename, and the log of error correction is provided to the users. Otherwise, the results of LRAddr[]

© 2013 ACADEMY PUBLISHER

and RLAddr[] are as the suggestion to provided to the users.

(1)（Forward placename error correction method）ForwardAddrErrorCorrection(L(i).Ar(j), GT(j)):

```
Input: L(i);
Output: LRAddr[];
Auxiliary Variable: L(i), LRAddr[],GT(0)～GT(n) , δ1, δ, SimStr;
Initialization: L(i)= Ar(0)Ar(1)…Ar(n), LRAddr[] = Null, SameSize=0, δ1 = 0.6, SimStr = “”;
Begin:
1.   for(int j=0; j < L(i).length; j++)      // L(i).length=n
2. SimStr = SearchSimStr(Ar(j).ToString(), GT(j));
     if(SimStr != “”)
3.       δ = CSim( SimStr, Ar(j).ToString());
           if(δ >=δ1)
4.           if(δ == 1)
               LRAddr.Add(SimStr);
             else
5.             Update(Ar(j).ToString()←SimStr);
               LRAddr.Add(SimStr);
           else
6.             LRAddr.Add(Null);
  End
```

(2)（Reverse placename error correction method）ReverseAddrErrorCorrection(LRAddr[],Ar(j),GT(j)):

```
Input: Ar(j), LRAddr[],GT(0)～GT(n);
Output: RLAddr[];
Auxiliary Variable:,RLAddr[],StandardStr;
Initialization: RLAddr[] = Null, StandardStr = “”;
Begin:
 for(int k=LRAddr.count -1; k >=0; k--)
   if(LRAddr[k].ToString() != Null)
StandardStr = SearchSimStr(LRAddr[k].ToString(), GT(k));
       if(SimStr != Null)
             RLAddr.Add(StandardStr);
       else
              RLAddr.Add(Null);
   else
       RLAddr.Add(Null);
End
```

Definition 6 the partial similarity: it is the ratio of the same Chinese characters between two geographical units. The calculation formula of the partial similarity is shown in formula 1. Among that, the Ti similarity [23] algorithm is used to the right side of formula to calculate the partial similarity:

$$CSim(S,S')=\min(\frac{U\{s_i\,|\,s_i\in S\cap s_i\in S'\}}{|S|},\frac{U\{s_i\,|\,s_i\in S\cap s_i\in S'\}}{|S'|}) \tag{1}$$

Among that, the parameter S is a geographical unit; S′ is the standard geographical unit, and S and S′ are geographical units of same areas; Si is the same Chinese character of S and S′; the partial similarity value can be calculated by matching with S and S′. If the value of CSim(S, S′) is more than the given threshold value δ1, it shows that S and S′ are similar, and then S is revised to S′.

Definition 7 the whole similarity: it is the ratio of the same geographical units between two Chinese placename. The calculation formula of the whole similarity $Sim(S,S')$ is shown in below. Among that, $asim(S,S')$ is the semantic factor; $ccsim(S,S')$ is the structure factor.

$$asim(S,S') = Min(\frac{\sum_{i=0}^{n} S_i}{|S|},\frac{\sum_{j=0}^{m} S'_j}{|S'|})\times(1-\rho)+\rho\times\frac{\sum_{i=0}^{n} S_i}{|S|} \tag{2}$$

$$ccsim(S,S')=\delta\times\eta\times Min(\sum_{i=0}^{n} Weight_i\,|S_i,\sum_{j=0}^{m} Weight_j\,|S'_j) \tag{3}$$

$$Sim(S,S')=\alpha\times asim(S,S')+\beta\times ccsim(S,S')$$
$$(\alpha+\beta=1\,,\,\beta\le\alpha\,) \tag{4}$$

Among that, the variable ρ is the abbreviation factor, if S is the abbreviation, ρ= 1; otherwise, ρ=0; the parameter η is the impact factor of swapping positions among same geographical units; the parameter δ is the position factor, $\sum_{i=0}^{n} Weight_i\,|S_i$ is the sum of weight of Si; the parameter α is the factor of semantic similarity, the parameter β is the factor of structure similarity.

V. Experiment Evaluation

In this section, the bi-direction error correction method for Chinese placename (BDAM) is compared with other 3 common error correction algorithms in order to verify accuracy of method. These three common error correction algorithms are that: 1) Word lists for spelling correction algorithm (WLSC); 2) Minimum edit distance algorithm (MED) [6, 9]; 3) N-gram algorithm (N=2) [22]. They are compared from time complexity and precision ratio. In order to ensure the fairness and accuracy of experiments, all the algorithms are performed in the same software and hardware environment(C# language (OS: Windows Server 2003; CPU: Core (TM) 2 Duo CPU T6570 2.1GHZ; Memory: 2G; Platform: Microsoft Visual Studio 2005)). On this basis, 10156 Chinese placename as the test data in data table “Planning management table” and the data table “Projects management table” are selected in the rural drinking water information system (RDWIS) to test. By the experiments, the results of experiments can be obtained, and as shown in figure 6, figure 7, figure 8 and figure 9. Among that, figure 6 and figure 8 use the given thresholds δ1=0.6 and δ2=0.6, figure 7 and figure 9 use δ1=0.6 and δ2=0.8.

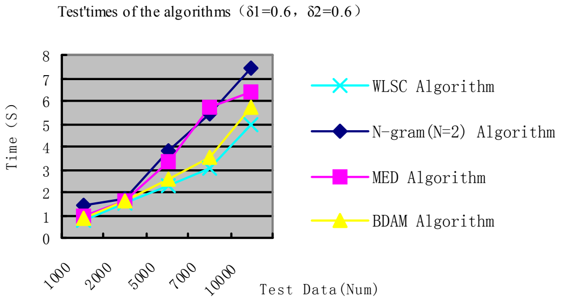


Figure 6.The running time of algorithms (δ1=0.6, δ2=0.6)

© 2013 ACADEMY PUBLISHER

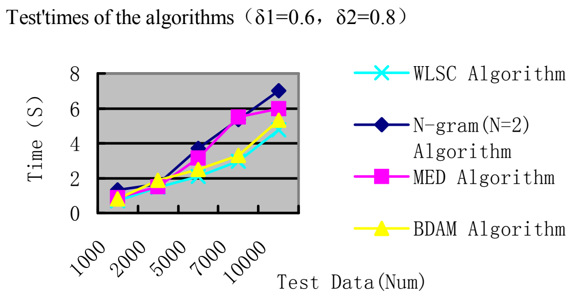


Figure 7.The running time of algorithms ((δ1=0.6, δ2=0.8)

For the different thresholds, the running times of four algorithms are as shown in figure 6 and figure 7. It can be seen from the figure 6 that the WLSC algorithm has the shortest running time. The reason is that WLSC algorithm mainly is suitable to detect and correct the misspelling word, this algorithm is relatively simple, and the detection and error correction are integrated, so the efficiently of algorithm is high, and the time complexity is low; the bi-direction error correction method for Chinese placename (BDAM) are mainly used to the forward and reverse error correction algorithms to revise the wrong Chinese placename. The time complexity is between O(n) and O(2n), so the running time of BDAM method is longer than WLSC algorithm. For the minimum edit distance algorithm (MED), because the idea of this algorithm is similar as the edit distance algorithm, the time complexity is also O(m*n), its time complexity is more than BDAM method, and the running time of MED algorithm is more than BDAM method's. Finally, due to the N-gram is the statistical algorithm, and the large-scale Chinese placename must be counted so that to obtain the transition probability matrix between words, with increasing on a large scale of test data, running time of this algorithm will increase. Therefore, the running time of N-gram algorithm is more than MED algorithm's. It can be seen from the figure 7 that the change of given threshold δ2 will lead to some Chinese placename not to satisfy the similarity, so the error correction for Chinese placename will decrease. In general speaking, the difference of running times between four algorithms are not obvious from the whole situation, but while threshold δ2=0.8, the running times of all four algorithms are less than threshold δ2=0.6's.

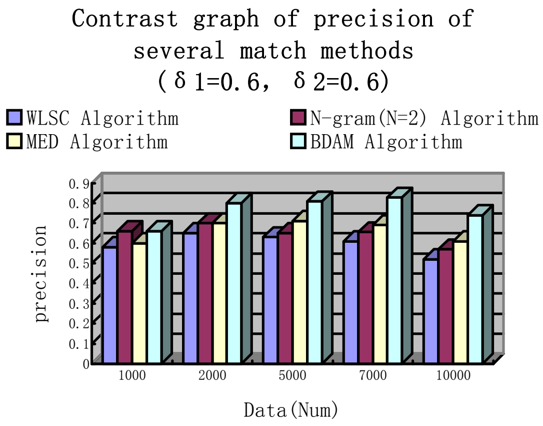


Figure 8. Precision ratio of four algorithms (δ1=0.6,δ2=0.6)

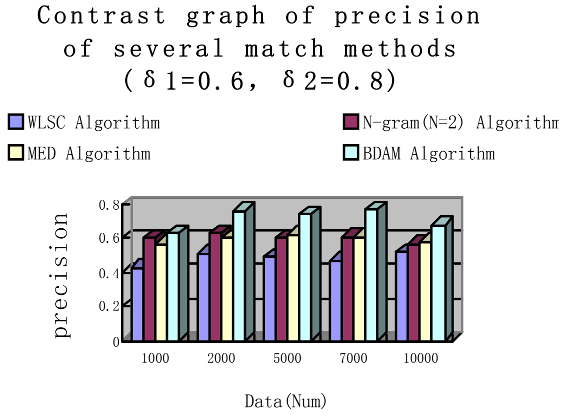


Figure 9. Precision ratio of four algorithms(δ1=0.6，δ2=0.8)

For the different thresholds, the precision ratios of four algorithms are as shown in figure 8 and figure 9. It can be seen from the figure 8 that: because the WLSC algorithm is difficult to ensure the un-ambiguity and comprehensiveness of misspelling dictionary, the precision of this algorithm is low; N-gram is a common statistical algorithm, if the scale of data is small, N-gram algorithm can ensure the precision ratio. However, due to some Chinese placename are not often to appear in the text, the precision of N-gram algorithm will decrease with the increasing of data scale; for the minimum edit distance algorithm (MED), it may be able to revise the error Chinese placename of insertions types, deletions types and substitutions types effectively, and the precision and recall ratio of this algorithm are not affected by the scale of data severely, so the precision ratio can be ensure. However, the MED algorithm is difficult to correct the error data of translocation type, so some these data could not be detected and corrected, and precision ratio of MED algorithm will fall. For the bi-direction error correction method for Chinese placename(BDAM), Chinese placename can be searched and corrected twice, and it combines the standard data in address- name index table with partial and whole similarity calculation to detect and correct the non-word error and real-word error data, so the precision ratio is high. However, some wrong placename can not be corrected by this method, it must use the way of manual intervention to correct the wrong placename. It can be seen from the figure 9 that the change of a given threshold δ2 will lead to some Chinese placename not to satisfy the similarity, so the error correction will decrease.

Through the experimental statistics, the experimental result of 1000 test data is as shown Table.1 The precision and recall ratio of 1000 test data in table 1.

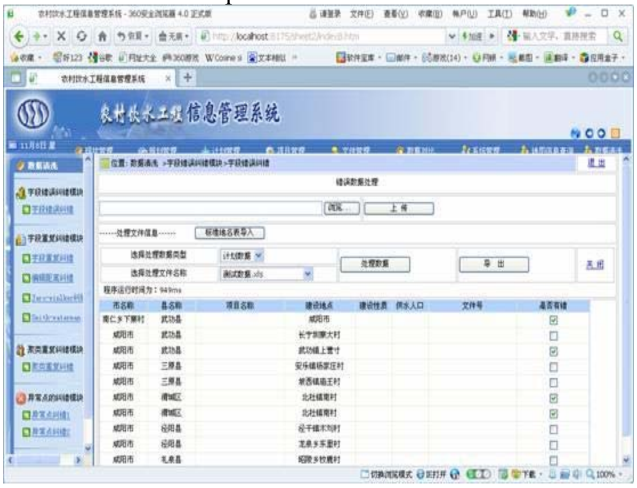
TABLE I

THE PRECISION AND RECALL RATIO OF 1000 TEST DATA

Test items / Algorithm	Correct placename data	Revised data	Non-detection Data	Wrong data	Precision ratio	Recall ratio
WLSC	74	129	223	352	0.57	0.21
N-gram	110	169	201	352	0.65	0.31
MED	130	212	140	352	0.61	0.37
BDAM	172	260	92	352	0.66	0.49

The result of figure 8 can be verified by table1, and it can be seen from table 1 that, the precision and recall ratio of bi-direction error correction method (BDAM) is highest in the four algorithms. The reason is that BDAM method can detect the most error Chinese placename, and the most error placename can be corrected. Then the recall ratios of MED and N-gram algorithms are more than WLSC algorithm.

Finally, this paper shows the experiment result of 1000 test data, and it is shown in figure 10. Firstly, the test data are stored in the database, then by using the algorithm to correct the error data. If the Chinese placename is wrong, it will be flag "√", and the log or suggestion of error correction will to be provided to the users.



位置：数据清洗 >字段错误纠错模块>字段错误纠错

错误数据处理

------处理文件信息------

选择处理数据类型：计划数据

选择处理文件名称：测试数据.xls

程序运行时间为：949ms

市名称	县名称	项目名称	建设地点	建设性质	供水人口	文件号	是否有错
南仁乡下寨村	武功县		咸阳市				☑
咸阳市	武功县		长宁镇康大村				□
咸阳市	武功县		武功镇上营村				☑
咸阳市	三原县		安乐镇杨家庄村				□
咸阳市	三原县		[illegible]				□
咸阳市	渭城区		北杜镇南村				☑
咸阳市	渭城区		北杜镇南村				☑
咸阳市	泾阳县		[illegible]				□
咸阳市	泾阳县		[illegible]				□
咸阳市	礼泉县		[illegible]				□

Figure 10. The result of error correction for Chinese placename

VI. CONCLUSIONS

This paper proposes a cleansing framework based on the address-name index table and bi-direction placename error correction method so that to detect and correct the wrong Chinese placename. And by the theoretical analysis and experiments' result, the feasibility and rationality of the algorithm can be proved, and it is benefit to enhance the accuracy of the bi-direction placename error correction method. The future works are focus on three points: 1) to optimize the data structure of address-name index table and to improve the efficiency of detection; 2) to combine the semantic similarity and structure similarity to improve the accuracy of matching; 3) to improve the comprehensive suggestion of error correction. In the future, we will to solve those problems in order to enhance the accuracy of error correction about Chinese placename.

ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China under Grant (No. 61172018); the Major Program of the National Soft Science Research (sub-project): the research of judicial information network and community correction information system strategy based on the cloud computing and internet of things; the Water Conservancy Science and Technology Plan Project of Shannxi Province: the research and development of the key technology of cloud computing platform in the new water conservancy information network.

REFERENCES

[1] Yangsen Zhang, Shiwen Yu.Summary of Text Automatic Proofreading Technology [J].In: Application Research of Computers, 2006, 23(6):8-12.

[2] Wenfei Fan.Extending dependencies with conditions for data cleaning[J].In:2008 8th IEEE International Conference on: Computer and Information Technology: 2008; 185-190.

[3] P. Neely. :Data Quality tools for Data Warehousing: a small sample survey.In: proceedings of MIT conference on Information Quality. Center for technology in government.University at Albany/ SUNY: Germany (1998).

[4] GUO Zhi-mao, ZHOU Ao-ying.:Research on data quality and data cleaning: A Survey[J].In: Journal of Software, 2002, 13(11): 2076-2082

[5] Cihan Varol, Coskun Bayrak.Risk Assessment of Error-Prone Personal Information in Data Quality Tools [J].Journal of Computers, 2011, 6(2): 155-161.

[6] A.V.Aho, T.G.Peterson.:A minimum distance error-correcting parser for context free languages[J].In: SIAM J.Comput, 1972, 12, 1(4): 305-312.

[7] R.C.Angell, G.E.Freund, P.Willett.:Automatic spelling correction using a trigram similarity measure[J].In: Inf.Process.Manage, 1983, 19, 255-261.

[8] P.Brown, S.Della Pietra, V.Della Pietra, R.Mercer.: Word sense disambiguation using statistical methods[C].In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (Berkeley, Calif.), ACL, 1991, 6, 264-270.

[9] Karen Kukich. Techniques for Automatically Correcting Words in Text[J].ACM Computing Surveys, 1992, 24(4):377-438.

[10] Yang-sen Zhang, Yuan-da Cao, Shi-wen Yu. : A Hybrid Model of Combining Rule- based and Statistics-based Approaches for Automatic Detecting Errors in Chinese Text[J].In: Journal of Chinese Information Processing, 2006, 6, 20(4): 1-7.

[11] James L Peterson. : Computer Programs for Detecting and Correcting Spelling Errors[J].In: Communication of the ACM, 1980,(12):676-687.

[12] Joseph J Pollock. :Automatic Spelling Correction in Scientific and Scholarly Text[J].In: Communication of the ACM, 1984,(4):358-368.

[13] E M Risenman. :A Contextual Post- processing System for Error Correction Using Binary N-gram[J].1974, 22 (5):480-493.

[14] Golding A R, Schabes Yves. : Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction [C].In: SantaCruz: Proceeding of the 34th Annual Meeting of the Association for Computational Linguistics, 1996.71-78.

[15] Liangju Wang, Huihui Zhang, Wanlian Li.Analysis of Causality between Tourism and Economic Growth Based on Computational Econometrics[J].Journal of Computers, 2012, 7(9):2152-2159.

[16] Lei Yang, Hu Lin, Dongfeng Yue, Tianrong Gao.A New Error-Correcting Transmission Method for Dual Ring Fieldbus in CNC System[J].Journal of Computers, 2012, 7(10): 2432-2438.

[17] Dengyue Li, Shengrui Wang, Zhen Mei.: Approximate Address Matching[J].In: 2010 IEEE International Conference on: P2P, Parallel, Grid, Cloud and Internet Computing: 2010; 264-269.

[18] R.J.Bock, W.Krischer.: The Data Analysis Briefbook [M].Springer, 1998.

[19] R.A.Johnson, D.W.Wichern.:Applied Multivariate Statistical Analysis[J].In: 4th ed., Prentice Hall, 1998.

[20] T.Zhang, R Ramakrishnan, M Liv ny.BIRCH: A New Data Clustering Algorithm and Its Applications[J].In: Data Mining and Knowledge Discovery, 1997, 1, 141-182.

[21] A.Marcus, J.I.Maletic.: Utilizing Association Rules for the Identification of Errors in Data[J].In: The University of Memphis, Division of Computer Science, Memphis, Technical Report TR-14-2000, 2000,3.

[22] E D Ward, D M Riseman. In: Contextual Word Recognition Using Binary Digrams[J].In: IEEE on Computers, 1971,20(4): 397-403.

[23] Sam Y, Sung Zhao Li, Peng Sun.: A Fast Filtering Scheme for Large Database Cleaning[A].In: K Wong, F Lam, N Lam.Proceedings of the eleventh international conference on Information and knowledge management.Boston: ACM .2002.76-83.

***Corresponding Author**: Ou Ye,
E-mail: 785669070@qq.com
Address: School of Computer science and engineering, Xi'an University of technology, Xi'an 710048, P.R.China
Authors: Jing Zhang, Ph.D; JunHuai Li. Ph.D
E-mail: zhangjing@xaut.edu.cn
Address: School of Computer science and engineering, Xi'an University of technology, Xi'an 710048, P.R.China