PROCEEDINGS OF SCIENCE

The SNM Algorithm Based on Multiple Edit Distances and Variable Window

Qiaoqiao Yang[1]
Data science and software engineering institute of Qingdao university, Qingdao, 266000, China
Email: 18354264193@163.com
Zhenbo Guo
The college of computer science and technology of Qingdao university Qingdao， 266000,China
Email: gzb@qdu.edu.cn
Kaixi Wang
The college of computer science and technology of Qingdao university Qingdao， 266000,China
Email: kxwang@qdu.edu.cn

In order to solve the drawback of SNM algorithm, which leads to inefficiency in detecting precision on approximately duplicate records when multiple sources need to be integrated, the improved SNM algorithm based on the variety of edit distance and the variable window is proposed in this paper. According to the record pattern of edit distance and the size of edit distance, the approximately duplicate records of data are deleted to reduce the number of comparisons. Then we put forward the mechanism of variable window to solve the problem. Empirical results show that this algorithm can effectively solve the problem that the window is too big or too small to leak. The experiment shows that the improved SNM algorithm can solve many problems in the integration of multi-source , and it has obvious advantages in ensuring precision and boosting efficiency.

CENet2017
22-23 July 2017
Shanghai, China

1Foundation Items: NSFC-General technical foundation research Joint fund (U1536113)

PoS(CENet2017)063

1.Introduction

The data integration can solve the problem effectively on "Information Island". The problem of heterogeneous data source integration has become a hot research topic, and the research is carried out on the data storage, analysis and mining, etc. The enterprises adopt different database management systems, which will create many problems in the process of integration [1]. How to identify similar duplicate records and eliminate redundant "dirty data" has become one of the key problems of heterogeneous database integration [2]. In the process of dealing with huge amounts of data integration, the primary task is to identify the approximately duplicate records. The sorted-neighborhood method (SNM) is a common algorithm in detecting approximately duplicate records, but the precision and efficiency of time are not high. Therefore, an improved SNM algorithm is proposed to solve this problem.

Eliminating duplication in large databases has drawn great attention. The recognition algorithm is composed of SNM algorithm [3], MPN algorithm [4] and KNN algorithm [5]. The SNM algorithm is a standard way to detect the similar duplicate records. The rest methods are mostly developed on the basis of this thought [6]. The SNM algorithm has good efficiency when the duplicate records are identified in the sliding window. The number of data records is N, The number of times between records is (N - 1)! . The number of times between records is W * N by SNM algorithm (W is a fixed window size and w<<N). However, the duplicate records are only a handful of data sets, even in the process of heterogeneous database, the vast majority of the records in the database are not similar to duplicate records. SNM algorithm has the following disadvantages: 1, The algorithm is dependent on keywords sorting; 2. The size of the sliding window is fixed and difficult to control; 3. It's necessary to improve the time with the number increasing. 4, The attribute similarity is 0 in the process of data integration when this attribute is loss. That may lead to wrong judgment [7].

2.Edit Distance

The database on keywords retrieval is often combined with the edit distance calculation in many applications. Edit distance is commonly used to compare two records of similarity measurement algorithm [8]. The literature makes the entire record as a string, then through calculating the edit distance of two strings to determine whether two records are similar [9]. The literature is calculated separately, then this algorithm combines the fields of edit distance to judge whether the record is similar [10]. The problems in the edit distance algorithm are: on the one hand, they only consider the number of editing regardless of the attribute loss. The different name attributes and largest public strings (LCS)have big effect on the similarity in record; on the other hand, the edit distance works well for English language, but cannot suit well for Chinese,because Chinese is letter language and edit operation is not flexible in dealing with the exchange between characters [11]. This article puts forward the similarity algorithm based on a variety of edit distance, such as pattern edit distance and attribute value edit distance.

2.1Pattern Edit Distance

The data comes from different databases in heterogeneous database. There are some problems such as inconsistent semantic representation, inconsistent spelling [12]which may lead to record mismatching when two records have different numbers of attribute, namely. Illustrated record M with five attributes {type, name, type, intelligence, note}, respectively corresponding to the attribute value {BS22, rongsheng, more open, intelligent, silver air-cooled}. Record N has

four attributes {type, name, type, note}, respectively corresponding to the attribute value {22BS, rongsheng, three open, air-cooled silver}, but N record clearly misses the contribution on intelligence. We need to refer the similarity of two records in missing attributes to determine whether the miss attributes would participatein calculation.

If the record of the missing attribute similarity is close to 1 or 0, it lacks the attributes . The attribute is invalid, the record has no contribution to the record. The attribute is not involved in sentence; If the record of the missing attribute similarity isn’t closed to 1 or 0, it lacks the attributes .The attribute is valid, the record has contribution to the record. The weight of each attribute completely depends on experiments combined with the mathematical statistics, the pattern edit distance is 1.

2.2Attribute Value Edit Distance

Edit distance, a.k.a. Levenshtein distance [13], is a common measurement of textual similarity. Formally, there are string A and B, the edit distance denoted edit (A, B) is the minimum number of edit operations (insertions, deletions, and substitutions) of single characters that are needed to transform A to B. Edit distance reflects the absolute difference of two string size. The smaller the edit distance, the greater the similarity. For instance,

edit（“edit operations”, “editt operationy”）= 2.

In particular, we can remove the fourth character “t” in the first string, and substitute the last character “s” with “y” to transform it to the second string. Similarly, edit （“ Jerod Descendant”, “Jerod Descendnd”）=2. It is known that the complexity of computing the edit distance between string A and B, $\text{edit}（\text{A,B}）\in\left[\left|M-N\right|,\max(M,N)\right]$, where N and M are the lengths of A and B. The similarity of two strings is Sim (A， B)=1- edit (A, B)/max(M, N).

It's a famous calculation formula about edit distance, which is due to Levenshtein again. By using the matrix can easily get edit (A, B):

- The matrix D[M][N] is established which have M row and N column. D[0][0] =0.
- The first-line should be initialized from 0 to N, the first column should be initialized from 0 to M. The letter of i test each character of s from 1 to N, and the letter of j test each character of s from 1 to M.
- If A[i] = B[j], edit (A, B) =0, else edit (A, B) =1.
- If the first string of the characters is equal to the second first string character, f (i, j) =1; else, f (i, j) =0.

When we use the traditional edit distance algorithm in Chinese, it can lead to the wrong conclusions. For example, the string has the same length but the order is different A = "Air cool optical control", B = "optical control Air cool". It can take 4 by using the traditional string edit distance algorithm, the similarity is zero. However, A and B are express the same meaning. In this case, the need to optimize the traditional edit distance formula. When we deal with the string with different lengths, it no longer belongs to an equal length exchange, then the traditional SNM algorithm is chosen.

If i >=1 and j >=1 and A[i] = B[j-1] and A[i-1] = B[j], D（i，j）=

$$\min\{D(i-1,j)+1, D(i,j-1)+1, D(i-1,j-1)+1，\ \text{edit}(i-2,j-2)+1\} \quad (2.1)$$

Through the expansion of the traditional edit distance algorithm, the operation has a wide range of applications in Chinese by increasing exchange operation. The operation can also

solve the problem of different name attribute. The Literature was given to solve the different properties of the solution. There is a way to distinguish the same attribute through comparing the attribute valuesaccording to the above example. The edit distance algorithm is 1 by using the improved edit, the similarity is 0.75. This value can reflect the similarityof the string. In many cases, the model and number are represented by English. The attribute value is given priority by Chinese word and we should choose the improved edit distance algorithm.

3.Dynamic Variable Window

The SNM algorithm based on sliding window has a problem. The size of the experimental window is estimate and fixed [14]. On the one hand, the duplicate records are more than the size of the window in a window containing. There are some similar duplicate records cannot be identified in a window containing; On the other hand, the window is too small while there is not similar duplicate records in window. The designed thought is: The data is in order according to the keywords. We set a fixed size window to solve cold start problems. The first data is participated in the operation with the second data; The first data is participated in the operation with the last data in this window. According to the numbers of records and the similarity in a window, the sliding window needs dynamic increase or decrease.

The initial value of the window is set to W. The window has a number data. The similarity was calculated on Sim (R_1, R_2), The similarity was calculated on Sim (R_2, R_3), ... , The similarity was calculated on Sim (R_m, R_{m+1}), The similarity was calculated on Sim (R_1, R_n).

Diff（R_1，R_2）=1- Sim（R_1，R_2），Diff（R_1，R_n）=1- Sim（R_1，R_n），

Diff（R_m）=1/m（Diff（R_1，R_2）w_1+ Diff（R_2，R_3）w_2+…+Diff（R_m，R_{m+1}）w_m)m<<n

The values of w refer to the concept of weighted moving average. The value far from the objective function has less influence, so they should be given less weight while the nearest value can best predict the case, they should be given the biggest weight, $w_1 < w_2 <...< w_m$, w_1 w_2 + +...+ $w_m = 1$.

If Diff(R_1, R_n)>(n-1)*(Diff(R_m)), then W'=(1- Diff $_{threshold\ value}$/Diff(R_1，R_n)) *W, W_{new}=W-W'. The minimum value is no less than the keyword number of records that have the same sort of minimum.

If Diff(R_1, R_n)< (n-1)* (Diff(R_m)), then W'=(1- Diff (R_1, R_n）/Diff $_{threshold\ value}$）*W，W_{new}=W+W'. The maximum value is not more than keyword number of records that have the same sort of maximum.

4. Improved SNM Algorithm

The SNM algorithm based on a variety of edit distance and variable window is improved. The basic steps are as follows:

Input: record set C, the effective factor V = {0, 1}, record the similarity threshold value U, cold start and fixed window size L, attribute weights of W.

Output: all could pose a similar duplicate records set.

Step 1: the keywords are sorted by preprocessing to avoid that expression difference caused on duplicate records in different position. The SNM algorithm is sensitive to keyword, so the order of the data is set according to the sort key before the treatment. The date of form - DD YYYY - MM, MM/DD/YYYY, YY - MM - DD and other forms in the basis of the time series data set. The keyword is unified format for ISO 8601 W3CDTF specification, and we use

YYYY - MM - DD format.

Step 2：the size of the window is adjusted by the relationship between Diff (R_1， R_n) and n* Diff（R_m). if Diff(R_1， R_n)>(n-1)*(Diff（R_m)), then W'=（1- Diff $_{threshold\ value}$/Diff(R_1， R_n)）*W， W_{new}=W-W'. if Diff(R_1， R_n)< (n-1)* (Diff(R_m)), then W'=（1- Diff（R_1， R_n）/Diff $_{threshold\ value}$）*W, W_{new}=W+W'.

Step 3：The calculation of similarity method should be chosen by the record pattern edit distance. If the pattern edit distance is 1, the attribute is missing. The attribute will be set effective factors and dynamic weights. The weights of attribute is determined by the expert's experience combined with the mathematical statistics and subjective way. If the lack of records attribute similarity is close to 1 or 0, the attributes are given the effective factor of 0. It is not involved in sentence, which does not affect the record similarity. If the missing field similarity is not close to 0 or 1. The similarity of each field remains the same. According to the different string edit distance algorithm, the similarity calculation record is chosen in English or Chinese.

Step 4： the similar duplicate records are output until all records are compared.

The specific flow chart is shown in Figure 1.

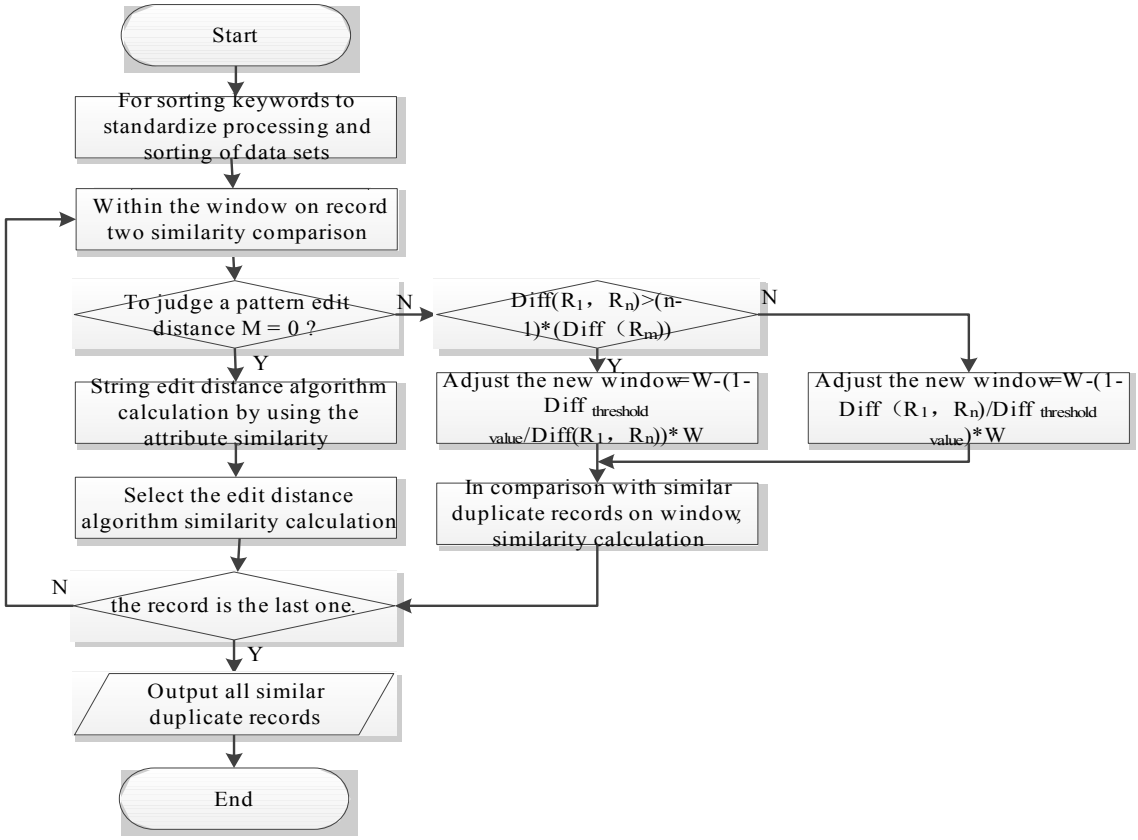

Figure 1: Algorithm Flow Chart

5. Simulation Experiments and Analysis

5.1. Simulation Experiments

The improved SNM algorithm needs to be verified in this paper. The experiment's data set comes from home appliance industry management system of refrigerator data. For the results evaluation, this article adopts the method of artificial judging. The first refrigerator database data contain nine attributes, the second table contain eight properties. Two tables have 4000

data. The second refrigerator database table misses the attribute of goods gross weight. In addition to the attribute of goods gross weight, the two record similarity is close to 1 or 0, then the attributes of goods gross weight are given the effective factor is 0.

Through a large number of experimental verification, the experiment comes to the conclusion that the similarity threshold value is set from 0.72 to 0.78, but it adopts the whole data samples from hospital [14]. On the choice of parameters, the experimental process similarity threshold value is set to 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78; the size of the sliding window initial value is set to 30, 35, 40, 45, 50; through repeating comparison and verification, we can find that different values of similarity calculation results will produce important influence. The experimental process similarity threshold value subject to change with the sample data, which has very strong practical significance. In order to make the result of the experiment to have reasonable operation value, we have carried out many tests and set the operation cost ultimately: The experimental process similarity threshold value is set to 0.75, the size of the sliding window initial value is set to 40. The sample data is shownin table 1.

Refrigerator Model	Intelligent Type	Freezer Models	……	Door sSructure	Energy Efficiency
430WEZ50	not support	Cold storage refrigeration	……	Split a two-door type	Level 1
312WDPM	not support	Straight cold	……	Many type	Level 3
456WDGK	not support	Cold storage refrigeration	……	More than a fridge	Level 2
……	……	……	……	……	……
WDGV1308	support	Air cooling	……	Cross across the hall	Level 2

Table 1: Refrigerator Information Sample Data

5.2. Results Analysis

The results detected similar article duplicate records and compare the number with the correct recognition of similar duplicate records using artificial statistics. Respectively using the improved algorithm and SNM duplicate records similar detection algorithm, according to the record of detected respectively two algorithm precision rates and recall rates. The recall ratio and the precision ratio are defined as:

the recall ratio=TP/(TP+FN) *100%;the precision ratio=TP/(TP+FP) *100%

(TP—number of true positives; FN—number of false negatives; FP—number of false positives)

The results as shown in Table 2 and Table 3.

Algorithm Record	SNM algorithm	The improved algorithm
1000	91.3%	91.5%
2000	91.0%	91.1%
3000	90.5%	90.8%
4000	90.3%	90.6%

Table 2: Precision Ratio

Algorithm Record	SNM algorithm	The improved algorithm
1000	83.8%	83.8%
2000	83.5%	83.5%
3000	82.9%	82.8%
4000	82.3%	82.3%

Table 3: Recall Ratio

The experiment proves that the improved algorithm precision is better than the traditional

SNM algorithm without reducing the rate of recall.

6. Conclusion

The method bases on a variety of edit distance and variable window, and a combination of both can solve the problems of recognizing the big data's duplicate records effectively. But, there are still unsolved problems, such as improving the recall ratio and dealing with the non-standard sample. The records of difference values are set according to the experience because it has exerted certain influence on the size of the window. We can continue to study as well as set a dynamic threshold value of specific issues in our later work.

References

[1] Pahwa P, Arora R, Thakur G. *An efficient algorithm for data cleaning*[J]. International Journal of Knowledge-Based Organizations(IJKBO), 2011(4): 56—71.

[2] ZM Guo, AY Zhou. *Research on data quality and data cleaning: A survey*[J]. Journal of Software, 2002, 13(11): 2076-2082.

[3] Draisbach U, Naumann F, Szott S, et al. *Adaptive windows for duplicate detection*[C]//Data Engineering (ICDE), 2012 IEEE 28th International Conference on. IEEE, 2012: 1073-1083.

[4] J Zhang, Z Fang, Y Xiong, et al. *Optimization algorithm for cleaning data based on SNM* [J]. Journal of Central South University (Science and Technology), 2010, 6: 034.

[5] H Zhang, Berg A C, Maire M, et al. *SVM-KNN: Discriminative nearest neighbor classification for visual category recognition*[C]//Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. IEEE, 2006, 2: 2126-2136.

[6] Beskales G, Ilyas I F, Golab L, et al. *On the relative trust between inconsistent data and inaccurate constraints*[C]// International Conference on Data Engineering. IEEE, 2013:541-552.

[7] ZN Zhang, L He, YZ Tan, et al. *A heuristic approximately duplicate records detection algorithm based on attributes analysis*[J]. International Journal of Digital Content Technology & its Applications,2012,6(4):259-267.

[8] Q SHAO, K YE *Based on improved edit distance and similarity of Chinese character string matching* [J]. Electronic science and technology, 2016, 29(9): 7-11.

[9] L Jin, C Li, Mehrotra S. *Efficient record linkage in large data sets*[C]//Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings. Eighth International Conference on. IEEE, 2003: 137-146.

[10] YF QIU, ZP TIAN, YZ JI. *An efficient detection method of similar duplicate records* [J]. Journal of computer, 2001, 24(1): 69—77

[11] WH ZHU, J YIN, YH . *Big data environment fast duplicate detection method of high-dimensional data* [J]. Research and development of the computer, 2016, 53(3): 559-570.

[12] RZ LI, M FANG. *Based on the integration of heterogeneous metadata and its multiple table dynamic query algorithm* [J]. Computer engineering, 2007, 33(17): 111-113.

[13] Levenshtein V I. *Binary codes capable of correcting deletions, insertions and reversals*[J]. Soviet Physics Doklady, 1966, 10(1):707-710.

[14] W CHEN. *Research and application of data cleansing key technology and software platform* [D]. Nanjing university of aeronautics and astronautics, 2005.