Association for Information Systems
AIS Electronic Library (AISeL)

ICEB 2009 Proceedings | International Conference on Electronic Business (ICEB)

Winter 12-4-2009

A Review of Approaches to Ensure the Quality of Data Collected on the Internet

Chun-Hung Cheng

Follow this and additional works at: https://aisel.aisnet.org/iceb2009

This material is brought to you by the International Conference on Electronic Business (ICEB) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ICEB 2009 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

A REVIEW OF APPROACHES TO ENSURE THE QUALITY OF DATA COLLECTED ON THE INTERNET

Chun-Hung Cheng
Dept of Systems Engineering & Engineering Management
The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, China
chcheng@se.cuhk.edu.hk

Abstract

Different approaches have been used to detect errors in data collected on the Internet. Some of these existing approaches require prior knowledge of data. Others have to test a large number of parameter values. To address these limitations, two approaches have been recently proposed. In this paper, we review these two approaches.

Introduction

Surveys have been commonly used to understand public opinions and views. Data can be collected by interview and postal-mail. However, interviews are very costly for large samples and postal-mail surveys are slow. With the ubiquitous of personal computers, and the availability of broadband network, Internet surveys have become possible.

With the Internet, surveys can be conducted to reach out a large number of potential survey subjects at very low cost. Nevertheless, the subjects enter their responses without any assistance. Although data-type and data-range checking may be implemented together with an electronic survey form, they are not always effective given a diversity of survey questions and responses. Hence, data collected through this means may contain erroneous values. These erroneous data must be identified to ensure the survey quality. In this work, we shall focus on the detection of unsystematic errors. These errors are those that are not caused by survey design faults.

To assess the quality of data, application-dependent approaches use of prior knowledge of data and, hence, require different quality-checking procedures for different survey applications. Application-independent approaches do not need any such knowledge and provide one general quality-checking procedure for all applications. The flexibility of application-independent approaches makes them more appealing than the application-dependent approaches. However, many existing approaches require the testing of large number of parameters. In this paper, we review two recent approaches to address this limitation.

Review of Existing Approaches

Although the classical database literature considers errors in a database a serious problem (e.g., Felligi and Holt [4] and Naus et al. [13]), few studies propose ways to deal with the problem. Application-dependent approaches such as those by Freund and Hartley [5], Naus et al. [13], and Felligi and Holt [4] are all statistical-based. In detecting errors in a database, these approaches require knowledge of the data. Using these approaches, software developers may have to develop different programs for different database applications.

All application-independent approaches use clustering analysis techniques. Lee et al. [10] first applied a clustering approach. They defined a distance function to measure the difference between two records. Based on a distance matrix, they found the shortest path between a pair of records. Since the determination of the shortest path is an NP-complete problem (Storer and Eastman [17]), the shortest spanning path algorithm (Slagle et al. [14]) is used to find an approximate solution. A link between two records that is longer than the pre-specified threshold value will be broken. Records whose distances are less than the threshold value are similar and are placed in the same group. A record with no similar partners is an outlier.

Storer and Eastman [17] proposed three related clustering approaches. They used the same distance function as defined by Lee et al. [10]. The first approach is called the leader algorithm (Hartigan [7]). The leader algorithm clusters M records into K groups, where M and K are positive integer values and $M \geq K$. It assumes that the distance function between two records and the threshold value for group membership are available. The first record is a leader for the first group. A record is assigned to an existing group if its distance from the group leader is less than the threshold value. It becomes a new leader for a new group if its distance from every existing leader is more than the threshold value.

The second approach is a modification of the leader algorithm that we refer to as an average record leader algorithm. This modified algorithm uses the average record instead of the first record as an initial leader. Therefore, the algorithm can generate a solution independent of record order. On each pass, a record that is furthest from its group leader becomes a leader for a new group. If the algorithm were to produce K groups, it requires K passes through the data.

The third approach is another modification of the leader algorithm. Storer and Eastman [17] called it the greatest distance algorithm. The greatest distance algorithm uses a different criterion for selecting new group leaders. First, Storer and Eastman [17] define a non-deviant cluster as one that has more than one percent of all records. A new leader is the record that is furthest from a leader of a non-deviant cluster and is greater than the average record distance from its cluster leader.

Cheng et al. [2] proposed the use of hierarchical clustering. They demonstrated that the use of hierarchical reduces the number of parameter values to test for data quality. In another work, Cheng et al. [3], they made use of a non-hierarchical clustering based on genetic search.

Data Quality

Currently, application-independent approaches use clustering analysis techniques. Clustering analysis gathers data records into groups or clusters based on their field values. Similar data records occupy the same group while dissimilar records do not coexist in the same group. Records, whose field values make them significantly different from all others, may not find themselves related to any other group members at all (see Figure 1). They are called outliers (Storer and Eastman [17]).

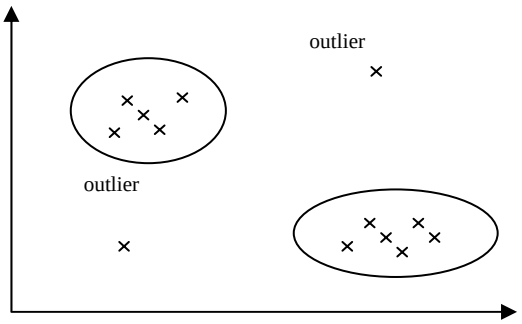

Figure 1: Examples of outliers

A data record, R_i, may be represented by a vector. That is, $R_i = (x_{i1},x_{i2},...,x_{iN})$, where x_{ip} is the value of the *pth* field of R_i, for $p = 1,2,\ldots N$ and $i = 1,2,\ldots,M$. A record can be classified into one of the three types (Lee et al. [10]).

Type I records: All field values in this type of record are numerical. The distance between two records R_i and R_j is defined in equation (1).

$$d_{ij} = \sum_{p=1}^{N} c(x_{ip}, x_{jp}) / N \text{ , where } c(x_{ip}, x_{jp}) = |x_{ip} - x_{jp}| / S_p \text{ , and } S_p = |\max_{1\le i\le M} x_{ip} - \min_{1\le i\le M} x_{ip}| \quad (1)$$

For example, if R_i = (4.5, 3.1, 0.9, -2.1), R_j = (4.1, 2.1, 0.3, -1.1), S_1 = 5.0, S_2 = 4.0, S_3 = 2.0, and S_4 = 2.1, then d_{ij} = 0.2765.

Type II records: All field values in this type of record are non-numerical. The distance between two records R_i and R_j is defined as:

$$d_{ij} = \sum_{p=1}^{N} c(x_{ip}, x_{jp}) / N \quad (2)$$

where

$$c(x_{ip}, x_{jp}) = \begin{cases} 1 \text{ if } x_{ip} \neq x_{jp} \\ 0 \text{ otherwise.} \end{cases}$$

Type III records: Fields in a type III record may assume either numerical or non-numerical values. The distance between two records R_i and R_j is defined as:

$$d_{ij} = \sum_{p=1}^{N} c(x_{ip}, x_{jp}) / N \quad (3)$$

where

for a numerical field *p*,

$c(x_{ip}, x_{jp}) = |x_{ip} - x_{jp}| / S_p$, and

$S_p = |\max_{1\le i\le M} x_{ip} - \min_{1\le i\le M} x_{ip}|$

or for a non-numerical field *p*,

$$c(x_{ip}, x_{jp}) = \begin{cases} 1 \text{ if } x_{ip} \neq x_{jp} \\ 0 \text{ otherwise.} \end{cases}$$

For example, if R_i = (black, black, 3.1, 5.0), R_j = (black, white, 2.1, 5.1), S_3 = 4.0, and S_4 = 5.5, then d_{ij} = 0.3170.

Lee et al. [10], and Storer and Eastman [17] use Euclidean distances or city block distances for type I records, and hamming distances for type II records. There is no upper bound on the value of either distance function. Therefore there are a large number of possible threshold values.

To illustrate the new distance function, consider a simple example with type III records. Table 1 is a personnel database for a hypothetical company.

Table 1: Example

Record	POS[1]	EDU[2]	MON[3]	SAL[4]
1	0	0	15	20,000
2	1	1	10	20,000
3	0	0	11	20,000
4	1	1	35	60,000
5	1	0	17	30,000
6	0	1	17	30,000
7	0	0	16	20,000
8	1	1	33	65,000
9	1	0	16	46,000
10	0	0	50	80,000

Note

1. POS = 1, when an employee has a middle management position; and POS = 0, when an employee has a supervisor position.
2. EDU = 1, when an employee has a college degree; and EDU = 0, when an employee does not have a degree.

3. MON is the number of months an employee has worked for the company.
4. SAL is the current salary of an employee.

Matrix (4) shows the distance value between a pair of records. Note that in the matrix, $d_{ii} = 0$ and $d_{ij} = d_{ji}$. A small distance value between two records implies that they are similar, while a large distance value means that they are different.

An erroneous record, being so different from other records, has large distance values with other records. When records are clustered into groups, erroneous records (i.e., outliers) will not be associated with other records.

Records

	1	2	3	4	5	6	7	8	9	10
1	.00	.52	.02	.73	.29	.29	.01	.73	.34	.36
2	.52	.00	.51	.25	.32	.32	.53	.26	.36	.89
R 3	.02	.51	.00	.74	.31	.31	.03	.75	.36	.89
e 4	.73	.25	.74	.00	.43	.43	.72	.03	.39	.64
c 5	.29	.32	.31	.43	.00	.50	.29	.44	.06	.57
o 6	.29	.32	.31	.43	.50	.00	.29	.44	.56	.57
r 7	.01	.53	.03	.72	.29	.29	.00	.73	.33	.36
d 8	.73	.26	.75	.03	.44	.44	.73	.00	.39	.63
s 9	.34	.36	.36	.39	.06	.56	.33	.39	.00	.53
10	.36	.89	.89	.64	.57	.57	.36	.63	.53	.00

(4)

Records

	1	3	7	4	8	5	9	2	6	10
1	.00	.02	.01	.73	.73	.29	.34	.52	.29	.36
R 3	.02	.00	.03	.74	.75	.31	.36	.51	.31	.89
e 7	.01	.03	.00	.72	.73	.29	.33	.53	.29	.36
c 4	.73	.74	.72	.00	.03	.43	.39	.25	.43	.64
o 8	.73	.75	.73	.03	.00	.44	.39	.26	.44	.63
r 5	.29	.31	.29	.43	.44	.00	.06	.32	.50	.57
d 9	.34	.36	.33	.39	.39	.06	.00	.36	.56	.53
s 2	.52	.51	.53	.25	.26	.32	.36	.00	.32	.89
6	.29	.31	.29	.43	.44	.50	.56	.32	.00	.57
10	.36	.89	.36	.64	.63	.57	.53	.89	.57	.00

(5)

When we rearrange rows and columns in Matrix (4) with the purpose of putting similar records together, we may get one possible solution shown in Matrix (5). It is not difficult to observe that there are three clusters: {1,3,7}, {4,8}, {5,9}. It is also apparent that Records 2, 6, and 10 are not associated with other records in any way. Therefore, they are the outliers.

Hierarchical Clustering

Cheng et al. [2] used hierarchical clustering. A hierarchical clustering technique operates on a distance matrix. It constructs a dendogram that depicts relationships among records. Anderberg [1] discusses seven hierarchical clustering techniques. Among the seven techniques, single linkage, average linkage, and complete linkage clustering are most widely used. In this paper, we choose single linkage clustering. Note that other hierarchical clustering techniques may also apply.

The single linkage-clustering algorithm operates on distance matrix (4) and produces the dendogram in Figure 2.

At this stage, a threshold value is needed to place records into groups. With the help of the dendogram, one can significantly reduce the number of possible threshold values to be examined. For example, Figure 2 shows the three possible threshold values and they are indicated as T1, T2, T3. The highest and second highest value of distance function are 0.89 and 0.57, respectively. A threshold value such as T1, where $0.89 \le T1 \le 0.57$, forms two groups with no outliers (i.e. {1,3,6,7,10}, and {2,4,5,8,9}). Similarly, a threshold value T2 where $0.57 \le T2 \le 0.44$ finds three groups (i.e. {1,3,6,7}, {10}, {2,4,5,8,9}) with record 10 as an outlier.

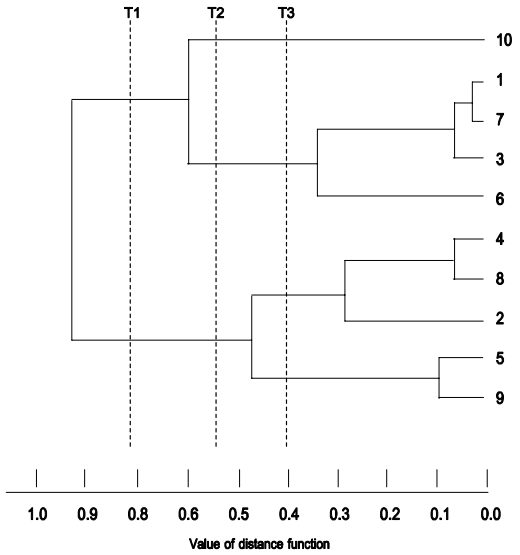

Figure 2: The dendogram

Stanfel [15] developed the classification criteria to classify data records into groups. These criteria seek to minimize the average distance within groups and maximize the average distance between groups. Minimizing the average distance within groups will put similar data records into the same groups. At the same time, maximizing the average distance between groups will put dissimilar data records into different groups.

Let's define:

$$Y_{ij} = \begin{cases} 1 \text{ if records } i \text{ and } j \\ \quad \text{are in the same group} \\ 0 \text{ otherwise} \end{cases} \quad (6)$$

The expression for the average distance *within* groups is given as:

$$\frac{\sum_{i=1}^{M-1}\sum_{j=i}^{M} d_{ij}\left(1-Y_{ij}\right)}{\sum_{i=1}^{M-1}\sum_{j=i}^{M}\left(1-Y_{ij}\right)} \quad (7)$$

While the expression for the average distance *between* groups is given as:

$$\frac{\sum_{i=1}^{M-1}\sum_{j=i}^{M} d_{ij}Y_{ij}}{\sum_{i=1}^{M-1}\sum_{j=i}^{M} Y_{ij}} \quad (8)$$

In order to achieve the objective of maximizing the homogeneity of records within groups as well as the heterogeneity of records between groups, the difference between the average distance within groups and the average distance between groups is minimized as shown in criterion (9):

$$\frac{\sum_{i=1}^{M-1}\sum_{j=i}^{M} d_{ij}Y_{ij}}{\sum_{i=1}^{M-1}\sum_{j=i}^{M} Y_{ij}} - \frac{\sum_{i=1}^{M-1}\sum_{j=i}^{M} d_{ij}\left(1-Y_{ij}\right)}{\sum_{i=1}^{M-1}\sum_{j=i}^{M}\left(1-Y_{ij}\right)} \quad (9)$$

All clustering results obtained by using various threshold values are given in Table 2. Since each result has its own associated outliers, a selection criterion is needed to determine the best clustering result so that the most appropriate outliers can be identified. Among the eight clustering results, solution 5 is the best as the value of its selection criterion is the lowest. Based on solution 5, we conclude that records 2, 6, and 10 are outliers.

Table 2. The possible clustering results

No.	Clustering results	Outlier/s	Val. of criterion
1	{1,3,6,7,10}, {2,4,5,8,9}	no outlier	-0.2581
2	{1,3,6,7}, {2,4,5,8,9}, {10}	10	-0.2775
3	{1,3,6,7}, {2,4,8}, {5,9}, {10}	10	-0.3439
4	{1,3,7}, {2,4,8}, {5,9}, {6}, {10}	6,10	-0.3897
5	{1,3,7}, {4,8}, {5,9}, {2}, {6}, {10}	2,6,10	-0.4431
6	{1,3,7}, {4,8}, {2}, {5}, {6} {9}, {10}	2,5,6,9,10	-0.4402
7	{1,3,7}, {2}, {4}, {5}, {6}, {8}, {9}, {10}	2,4,5,6,8,9,10	-0.4322
8	{1,7}, {2}, {3}, {4}, {5}, {6}, {8}, {9}, {10}	2,3,4,5,6,8,9,10	-0.4244

Non-Hierarchical Clustering

This approach consists of two phases: obtaining a sequence of sample data records, and classifying records into groups. The first phase uses a genetic algorithm and the second phase adopts a classification criterion for grouping.

A chromosome represents an individual. For example, x_1 = (1011001) and x_2 = (0111011) are two distinct individuals. Offspring (new individuals) are generated by crossover. A crossover point will be selected randomly. The parent chromosomes will be split at the chosen point and the segments of those chromosomes will be exchanged. Using this basic crossover operator, two fit individuals may combine their good traits and make fitter offspring.

Nevertheless, the simple representation scheme described above is not suitable for TSP. Instead, three vector representations for TSP were proposed (Michalewicz [12]): adjacency, ordinal, and path. Each representation has its own genetic operators. Among the three representations, the path representation is the most natural representation of a tour. For example, a tour 3 – 4 – 1 – 6 – 5 – 2 – 7 is simply represented by (3 4 1 6 5 2 7). This proposed approach uses this representation.

Initialization involves generating of possible solutions to the problem. The initial population may be generated randomly or with the use of a heuristic. In this approach, the initial population is generated randomly.

Fitness function is used to evaluate the value of the individuals within the population. According to the fitness value scored, the individual is selected as a parent to produce offspring in the next generation or is selected to disappear in the next generation.

In TSP, the total distance is calculated as the distance travelled from the starting city to the last city plus the distance from the last city to the starting city. In this data auditing problem, returning to the starting city (i.e., record) does not have any practical meanings. Therefore, the problem is simplified to the associated Hamiltonian Path Problem (HPP). As the first and last records need not be connected, we may calculate the total distance of a path instead of a tour in our fitness functions.

Let ρ be the permutations of records along the row of the initial matrix. For a sequence of cities (i.e., records): (1 3 7 4 8 5 9 2 6 10), ρ(2) = 3 and ρ(7) = 9. The proposed approach converts the initial sequence of records (specified by the initial matrix) to a new sequence that minimizes the following fitness function:

$$\sum_{i=1}^{n-1} \partial_{\rho(i)\rho(i+1)} \qquad (10)$$

where n = number of records (i.e., rows or columns).

Parent selection is a process that allocates reproductive opportunities to individuals. There are several selection schemes: roulette wheel selection, scaling techniques, ranking, etc. (Goldberg [6]).

As the process continues, the variation in fitness range will be reduced. This often leads to the problem of premature convergence in which a few super-fit individuals receive high reproductive trials and rapidly dominate the population. If such individuals correspond to local optima, the search will be trapped like hill climbing.

Fitness ranking is used to solve the problem of premature convergence (Whitley [18]). Individuals are sorted according to their fitness values, the number of reproductive trails are then allocated according to their rank.

Several TSP crossover operators are defined: partially-mapped (PMX), order (OX), cycle (CX), and edge recombination (ER) crossover. Whitley et al. [18] found that ER is the most efficient crossover operator for TSP. Starkweather et al. [16] proposed an enhancement to ER and find it more efficient than the original operator.

Cheng et al. [3] used the EER operator. Since the EER operator incorporates random selection to a break tie, this mechanism creates an effect similar to mutation. In our approach, we do not use any mutation operator.

Mutation is applied to each child individually after crossover according to the mutation rate. It provides a small amount of random search and helps ensure that no point in the search space has a zero probability of being examined. Several mutation operations have been suggested by Michalewicz [12]. We do not plan to use mutation operation. This is because the crossover operator used incorporates a random selection in completing a legal permutation and the effect is similar to a mutation.

In each generation, only two individuals are replaced. In other words, parents and offspring may co-exist in the population. The genetic process is repeated until a termination criterion is met. In this case, we use a pre-specified maximum number of generations as a termination criterion. The same classification criteria by Stanfel [15] may be used to classify data records into groups.

Conclusion

In this work, we discussed the use of clustering algorithms for assessing the quality of data collected on the Internet. Limitations of some existing approaches were identified. Two recent approaches to address these limitations have been reviewed.

Reference

[1]. Anderberg, M.R., 1993. *Cluster analysis for ap plications*, Academic Press, New York.

[2]. Cheng, C.H., Goh, C.H., and Lee-Post, A., 2006. Data auditing by hierarchical clustering, *Inter natinal Journal of Applied Management & Tec hnology*, Vol.4, No.1, pp. 153-163.

[3]. Cheng, C.H., Goh, C.H., and Lee-Post, A., 20 07. A two-staged approach for assessing the q uality of Internet survey data. *The 3rd Internat ional Conference on Web Information Systems and Technologies*, Barcelona, Spain.

[4]. Felligi, I.P. and Holt, D., 1976. A systematic ap proach to automatic editing and imputation, *Jou rnal of American Statistics Assocication*, Vol. 7 1, pp. 17-35.

[5]. Freund, R.J. and Hartley, H.O., 1967. A proced ure for automatic data editing, *J Journal of Ame rican Statistics Assocication*, Vol. 62, pp. 341-3 52.

[6]. Goldberg, D.E., 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Massachusetts: Addison Wesley.

[7]. Hartigan, J.A., 1975. *Clustering Algorithms*, Mc Graw-Hill, New York.

[8]. Haupt, R.L. and Haupt, S.E., 1998. *Practical Genetic Algorithms*. New York: John Wiley & Sons.

[9]. Holland, J.H., 1975. *Adaptation in Natural an d Artificial Systems*. Michigan: Michigan Pres s.

[10]. Lee, R.C., Slagle, J.R., and Mong, C.T., 1978. T owards automatic auditing of records, *IEEE Tra nsactions on Software Engineering.*, Vol. SE-4, pp. 441-448.

[11]. Lenstra, J.K. and Kan Rinnooy, A.H.G., 1975. Some Simple Applications of the Traveling S alesman Problem, *Operations Research Quart erly*, Vol. 26**,** pp. 717-733.

[12]. Michalewicz, Z., 1999. *Genetic Algorithms + Data Structures = Evolution Programs*. Third, Revised and Extended Edition, Hong Kong: S pringer.

[13]. Naus, J.I., Johnson, T.G., and Montalvo, R., 197 2. A probabilistic model for identifying errors a

nd data editing, *Journal of American Statistics A ssocication*, Vol. 67, pp. 943-950.

[14]. Slagle, J.R., Chang, C.L., and Heller, S.R., 1975. A clustering and data-reorganizing algorithm, *I EEE Transactions on Systems, Man, and Cyber netics*, Vol. SMC-5, pp. 125-128.

[15]. Stanfel, L.E., 1983. Applications of clustering t o information system design, *Information Proce ssing & Management.*, Vol. 19, pp. 37-50.

[16]. Starkweather, T., McDaniel, S., Mathias, K., Whitley, D., and Whitley, C., 1991. A Compar ison of Genetic Sequencing Operators, *Procee dings of the fourth International Conference o n. Genetic Algorithms and their Applications*, pp.69-76

[17]. Storer, W.F. and Eastman, C.M., 1990. Some E xperiments in the use of clustering for data valid ation, *Information Systems.*, Vol. 15, pp. 537-54 2.

[18]. Whitley, D., 1989. The Genitor Algorithm and Selection Pressure: Why Rank-based Allocati on of Reproductive Trials Is Best, *Proeedings* of. the Third International Conference on Gen etic Algorithms, pp.116-121