Testing Various Similarity Metrics and their Permutations with Clustering Approach in Context Free Data Cleaning

Sohil D. Pandya sohilpandya@gmail.com
Lecturer, MCA Department
Sardar Vallabhbhai Patel Institute of Technology
Vasad, 388306, India

Dr. Paresh V. Virparia pvvirparia@yahoo.com
Reader, G H Patel PG Dept. of Comp. Sci. & Technology,
Sardar Patel University,
Vallabh Vidyanagr, 388120, India

Abstract

Organizations can sustain growth in this knowledge era by proficient data analysis, which heavily relies on quality of data. This paper emphasizes on usage of sequence similarity metric with clustering approach in context free data cleaning to improve the quality of data by reducing noise. Authors propose an algorithm to test suitability of value to correct other values of attribute based on distance between them. The sequence similarity metrics like Needlemen-Wunch, Jaro-Winkler, Chapman Ordered Name Similarity and Smith-Waterman are used to find distance of two values. Experimental results show that how the approach can effectively clean the data without reference data.

Keywords: Context free data cleaning, Clustering, Sequence similarity metrics.

1. INTRODUCTION

The management of any organizations is intense to sustain growth in markets and to achieve this; it heavily relies on data analysis. Data analysis provides an impending to proficient way for by and large picture and brings to float up detailed and hidden Information. The accuracy and relevance of data analysis heavily relies on the quality of data. The data quality measures like completeness, valid, consistent, timeliness, accurate, relevance etc. allow quantifying data in order to achieve high performance results of various data analyses. Because of human interventions and computations at various levels, noise is added in data before it got stored [3]. Noise is “irrelevant or meaningless data” [1], which leads to deterioration of outcome of data analysis. The data cleaning is a process of maintaining data quality in Information Systems (IS) [2]. The data cleaning processes mainly focused on detection and removal of noise. Using similarity metrics in data cleaning process to identify and replace incorrect sequence with correct sequence based on distance between them is an interesting way of cleaning the data [6]. Here the distance for various similarity metrics may be based numbers of characters, number of replacements needed to convert one sequence to another, number of re-arrangements required, most similar characters, any combinations of all above, etc. The distance between two sequence ranges from 0.0 to 1.0. For example, the distance between *Malaysia* and *Mallayssia* for various similarity metrics is shown Table 1.

The purpose of this paper is to apply similarity metrics and their permutations in *context free data cleaning* using *clustering* approach. The context free data cleaning means to examine and

transform values of attributes without taking account further values of attribute [4].

Similarity Metrics	Distance
Needlemen-Wunch	0.8000
Smith-Waterman	0.8750
Chapman Ordered Name Compound Similarity	0.9375
Jaro-Winkler	0.9533

TABLE 1: Values of Distance for Various Similarity Metrics.

The core idea is based on the frequency of values and based on matching between them it is decided whether they should be transformed or not? Clustering is the assignment of a set of observations into subset so that observations in the same clusters are similar in some sense, which has various applications in machine learning, data mining, pattern recognition, bioinformatics, etc. [7]. The later sections describe an algorithm, its experimental results and concluding remarks.

2. USAGE OF SIMILARITY METRICS IN CONTEXT FREE DATA CLEANING

The proposed algorithm has two major components, viz. clustering and similarity, and one important parameter *acceptableDist*, which is minimum acceptable distance required during matching and transformation. To measure distance we used following Similarity Metrics and their permutations:

1. Needlemen-Wuch
2. Jaro-Winkler
3. Smith-Waterman
4. Champan Ordered Name Compound Similarity

The Needleman-Wunch algorithm, as in (1) performs a global alignment on two sequences and commonly used in Bioinformatics to align protein sequences [8].

$$F_{0j} = d * j$$
$$F_{i0} = d * i \tag{1}$$
$$F_{ij} = \max(F_{i-1,j-1} + S(S_{1i}, S_{2j}), F_{i,j-1} + d, F_{i-1,j} + d$$

Where $S(S_{1i},S_{2j})$ is the similarity of characters i and j; d is gap penalty.

The Jaro-Winkler distance, as in (2), is the major of similarity between two strings [8]. It is a variant of Jaro distance [8].

$$Jaro-Winkler(S_1, S_2) = Jaro(S_1, S_2) + (L * p(1 - Jaro(S_1, S_2))$$
$$Jaro(S_1, S_2) = \frac{1}{3}\left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m}\right) \tag{2}$$

Where m is number of matching characters and t is number of transpositions required; L is length of common prefix and p is scaling factor (standard value 0.1).

The Smith-Waterman algorithm, as in (3) is well-known algorithm for performing local sequence alignment, i.e. for determining similar regions between two protein sequences. It compares segments of all possible lengths and optimizes the similarity measures using substitution matrix and gap scoring scheme [8].

$$H(i,0) = 0, 0 \le i \le m$$
$$H(0,j) = 0, 0 \le j \le n \tag{3}$$
$$H(i,j) = \max\begin{Bmatrix} 0 \\ H(i-1,j-1) + w(S_{1i}, S_{2j}), Mismatch \\ H(i-1,j) + w(S_{1i}, -), Deletion \\ H(i,j-1) + w(-, S_{2j}), Insertion \end{Bmatrix}$$

Where S1, S2 are strings and m, n are their lengths; H (i, j) is the maximum similarity between strings of S1 of length i and S2 of length j; w(c,d) represents gap scoring scheme.

Chapman Ordered Name Compound Similarity tests similarity upon the most similar terms of token-based name where later name are valued higher than earlier names [8].

2.1 Algorithm

***Step-1*:** Start.
***Step-2*:** Values for a selected attributed transformed into uppercase, after removal of non-alphanumeric characters.
***Step-3*:** Derive frequencies in descending order, for all the distinct sequences. Refer the group of distinct values as clusters and the sequences as cluster identifiers.
Step-4: Select any of the sequence similarity metrics for comparing two values of an attribute and decide acceptableDist.
***Step-5*:** Compare the cluster identifier with other cluster identifiers, beginning with first to last cluster, to decide distance between them.
***Step-6*:** If the distance is less than *acceptableDist* then it forms transformation and/or validation rules for particular *acceptableDist* that can be utilized in further cleaning process (e.g., second pass of the same algorithm, context dependant cleaning) and the values of comparables can be transformed in to comparator, else comparables remains as separate clusters.
***Step-7*:** Stop.

[Note: The extended version of the above algorithm is used for usage of various permutations of two Similarity Metrics, where we had two parameters – one for each Similarity Metrics, i.e. *accetableDist1* and *acceptableDist2* [5]. In the extended version we perform *Step-6* for both Similarity Metrics. The results for both approach is shown in Section 3]

2.2 Assumptions & Limitations

In the above experiments we made certain assumptions like (a) Most of data entry is done correctly, only 2 to 20 percent data injected is not correct, and (b) Entered incorrect values are typographic errors. The algorithm has limitations like (a) It may incorrectly altered values those may be correct in real world, (b) Correct values which are typographically similar may be transformed, (c) The result varies when same *acceptableDist* and different Similarity Metrics (or its combinations) upon a same dataset, which leads to confusion upon selection of Similarity Metrics.

3. EXPERIMENTAL RESULTS

The algorithm is tested using a sample data derived from Internet. The data consisting of attributes named First Name, Middle Name, Last Name, Address, City, Pin code, District, State, Country, Phone number, and Email. District attribute is selected the testing purpose. There were about 13,074 records out of which 551 (4.22 %) values for the selected attribute were identified as incorrect and required corrections. During the execution of algorithm, 359 clusters were identified for the selected attribute. After identification of clusters and their identifiers, algorithm is tested for various similarity metrics value. For selected similarity metrics various results, like how many records updated (total, correctly & incorrectly), were found and are discussed in this section. Following results, percentage of correctly altered (CA %), percentage of incorrectly altered (IA %) and percentages of unaltered values (UA %) were derived as in (4).

$$CA(\%) = \frac{CA}{TotalAlteration} * 100$$
$$IA(\%) = \frac{IA}{TotalAlteration} * 100 \quad (4)$$
$$UA(\%) = \frac{UA}{NumberofIncorrectValues} * 100$$

Results found on testing of algorithm are:

1. It can be observed from Figure 1 that application of Similarity Metrics and their permutations that the percentage of values altered is growing with increase of *acceptableDist* as the tolerance of matching criteria. (See Table 2 for Legend description). For instance, using Chapman Ordered Name Compound Similarity with distance values 1, 0.9, and 0.8 (of each) there were 0.35%, 4.60%, 9.81% values altered respectively out of all values.

2. It can also be observed from Figure 2 and 3 that as the increment of *acceptableDist*, the percentage of incorrectly altered values also gets increased. For instance, using Chapman Ordered Name Compound Similarity with distance values 1, 0.9, and 0.8 (of each) there were 7.14%, 38.57%, and 57.16% values altered incorrectly out of total values altered.

3. The efficiency of algorithm is increased, if we use permutation of Similarity Metrics instead of using a single Similarity Metric.

Sr. No.	Notation	Similarity Metric – I	Similarity Metric-II
1	NW	Needlemen-Wunch	-
2	JW	Jaro-Winkler	-
3	CONS	Chapman Ordered Name Compound Similarity	-
4	SW	Smith-Waterman	-
5	NWJW	Needlemen-Wunch	Jaro-Winkler
6	NWCONS	Needlemen-Wunch	Chapman Ordered Name Compound Similarity
7	NWSW	Needlemen-Wunch	Smith-Waterman
8	JWCONS	Jaro-Winkler	Chapman Ordered Name Compound Similarity
9	JWSW	Jaro-Winkler	Smith-Waterman
10	CONSSW	Chapman Ordered Name Similarity	Smith-Waterman

TABLE 2: Legend Description.

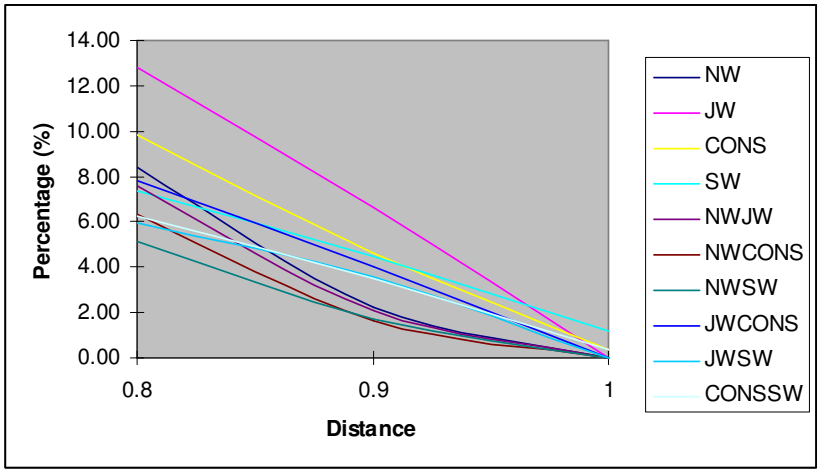

FIGURE 1: Percentage Alteration.

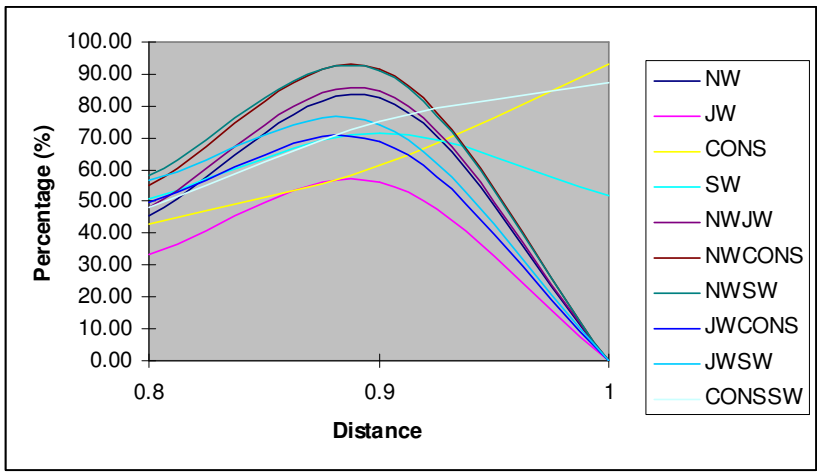

FIGURE 2: Percentage of Correctly Altered Values.

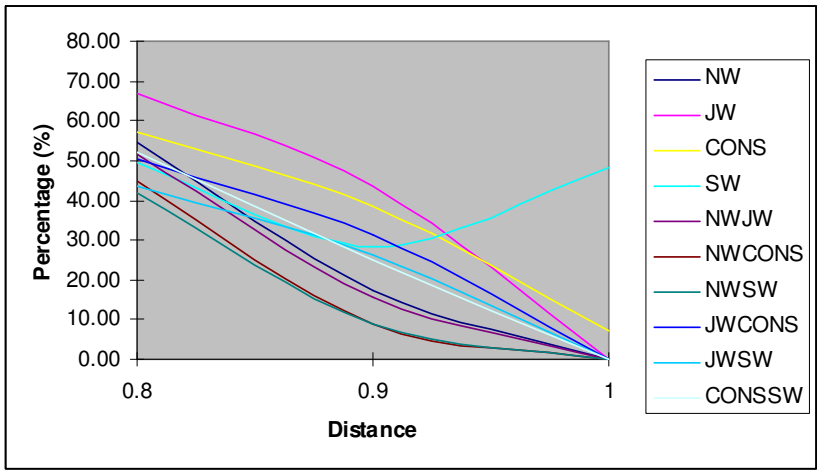

FIGURE 3: Percentage of Incorrectly Altered Values.

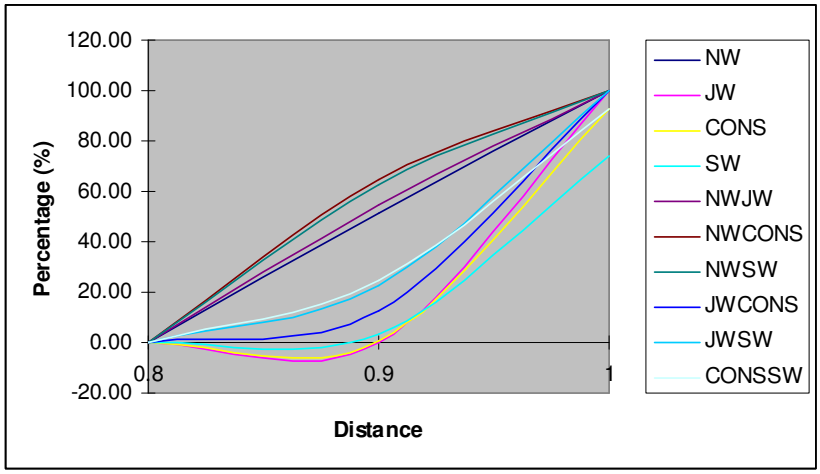

FIGURE 4: Percentage of Unaltered Values.

4. CONSLUSION

The results of the experiment verify the correctness of the algorithm and which motivate us to use it for data cleaning. The major advantage of it, where the reference/correct dataset is not given and still the data cleaning is achieved. However the various percentages shown in results depend on Similarity Metric(s), parameter(s), and dataset, i.e. for different dataset may require different aspects of said dependants. It may be possible that other Similarity Metrics or their permutations may give more precise data cleaning, that yet to be explored and future experiments.

5. REFERENCES

1. Hui Xiong, Gaurav Pandey, Michael Steinbach, Vipin Kumar. *"Enhancing Data Analysis with Noise Removal"*. IEEE Transaction on Knowledge & Data Engineering, 18(3):304-319, 2006.
2. Lukasz Ciszak. *"Applications of Clustering and Association Methods in Data Cleaning"*. In Proceedings of the International Multiconference on Computer Science and Information Technology. 2008.
3. Sohil D Pandya, Dr. Paresh V Virparia. *"Data Cleaning in Knowledge Discovery in Databases: Various Approaches"*. In Proceedings of the National Seminar on Current Trends in ICT, INDIA, 2009.
4. Sohil D Pandya, Dr. Paresh V Virparia. *"Clustering Approach in Context Free Data Cleaning"*. National Journal on System & Information Technology, 2(1):83-90, 2009.
5. Sohil D Pandya, Dr. Paresh V Virparia. *"Application of Various Permutations of Similarity Metrics with Clustering Approach in Context Free Data Cleaning"*. In Proceedings of the National Symposium on Indian IT @ CROXRoads, INDIA, 2009.
6. W Cohen, P Ravikumar, S Fienberg. *"A Comparison of String Distance Metrics for Name-Matching Tasks"*. In the Proceedings of the IJCAI, 2003.
7. http://en.wikipedia.org/
8. http://www.dcs.shef.ac.uk/~sam/simmetric.html