

# **AN ALGEBRAIC APPROACH TO QUALITY METRICS FOR CUSTOMER RECOGNITION SYSTEMS**

(Research-in-Progress)  
IQ Metrics

**John R. Talburt**  
Axiom Corporation  
[jtalbu@axiom.com](mailto:jtalbu@axiom.com)

**Emily Kuo**  
MIT IQ Program  
[emilykuo@mit.edu](mailto:emilykuo@mit.edu)

**Richard Wang**  
MIT IQ Program  
[rwang@mit.edu](mailto:rwang@mit.edu)

**Kimberly Hess**  
Axiom Corporation  
[khess@axiom.com](mailto:khess@axiom.com)

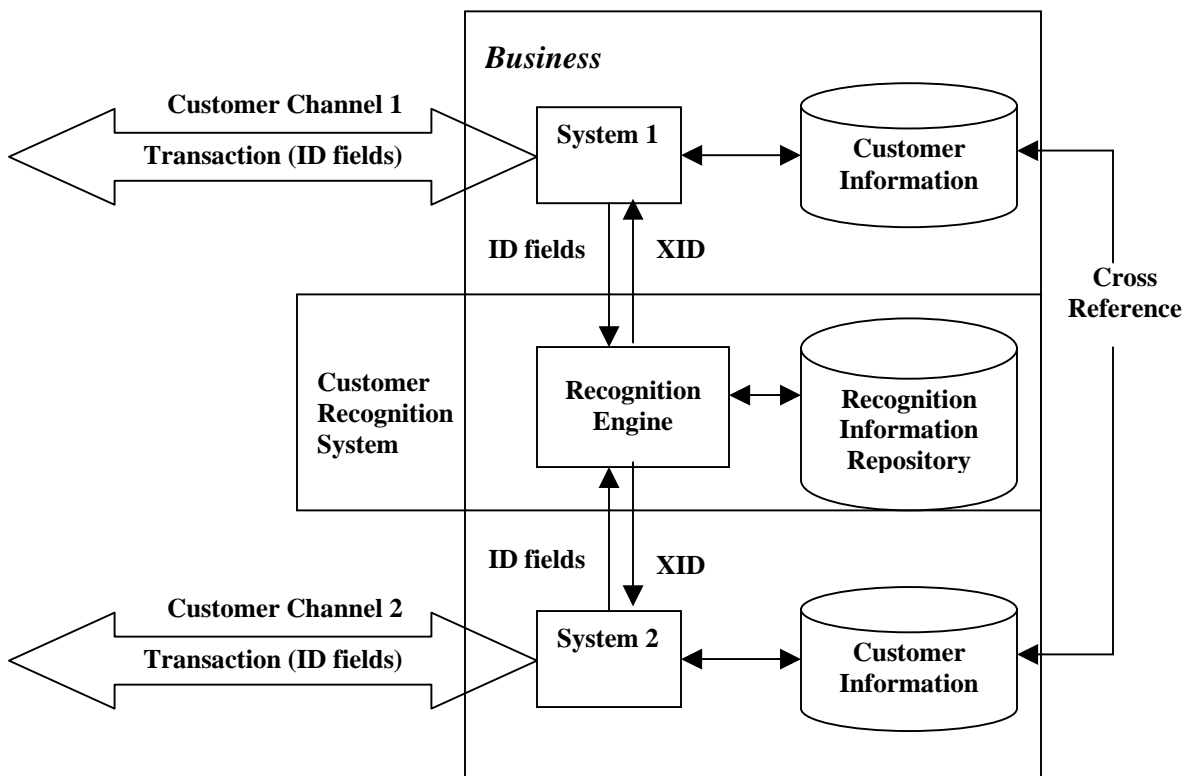
**Abstract:** Success in implementing a Customer Relationship Management (CRM) system requires close attention to data quality issues. However, most of the literature focuses on the quality of the input streams rather than the quality of the customer data integration (CDI) and customer recognition outcomes. This paper describes some preliminary research into the creation and validation of quality metrics for customer data integration and customer recognition systems. The approach is based on an algebraic view of the system as producing a partition of the set of customer transactions it processes. Comparing one system to another, or even changes to the same system, becomes a matter of quantifying the similarity between the partitions they produce. The authors discuss three methods for measuring the similarity between partitions, suggest the use of these measurements in creating metrics for customer recognition accuracy and consistency, and report on early experimental results.

**Key Words:** Data Quality, Information Quality, Metrics, Customer Data Integration, Customer Recognition Systems, Partitions, Partition Similarity

## 1.0 INTRODUCTION

Most modern businesses interact with their customers through several channels that carry transactions to and from internal systems. Channels may represent different lines of business (homeowners versus auto for an insurance company), different sales channels (in-bound telephone sales versus online sales for a retailer), or different geographic locations. It is common for each channel to have its own form of internal customer recognition based on one or more items of identification information. The identification information in the transaction may include some type of internally assigned customer key specific to that particular channel. Even within a single channel, key-based recognition is not perfect. The same customer may be assigned different identifying keys for a number of reasons. The White Paper on Customer-Centric Information Quality Management [7] published through the MITIQ program gives a more complete discussion of the factors that impact the quality of customer recognition.

In a multi-channel business the problem is further compounded by the need to recognize and profile customers across channels and synchronize the keys assigned by different channels. Figure 1.1 shows a typical configuration for a Customer Recognition System that manages recognition across channels.



**Figure 1.1: Block Diagram of a Multi-Channel Recognition System**

In Figure 1.1, the customer transactions coming through the channels include one or more items of identifying information. The two channels are connected to a recognition engine, which has access to a repository of recognition information that has been collected from both channels. The information in the repository is organized in such a way that the transactions belonging to the same customer are assigned a

unique cross-reference identifier, the XID shown in the diagram. The XID represents the customer's single, enterprise identity and is used to bring together the various internal (system) keys that the customer may have been assigned at different times or through different channels.

Despite the fact that customer recognition is a critical factor in successful CRM solutions, there is little guidance in the literature on metrics specific to customer recognition quality. This paper attempts to describe a formal approach to customer recognition quality metrics similar to what has been done by Wang, Lee, and others for database systems [8] and information products in general [3].

## 2.0 AN ALGEBRAIC MODEL FOR CUSTOMER RECOGNITION

Despite the complexity involved in an actual Customer Recognition System implementation, its function can be described relatively simply in terms of "equivalence relation" from basic abstract algebra. In this model there are three critical elements. Let

$$T = \{t_1, t_2, \dots, t_n\}$$

Represent a finite set of "n" customer transactions that have been processed in a particular order through a given Recognition Engine. As shown in Figure 1, the recognition will assign to each transaction an XID.

**Definition 2.1:** For a given Recognition Engine E, and a given order of the transactions T, define the **binary relation  $R_E$**  on the set of transactions T by

$$R_E \subset T \times T, \text{ such that}$$

$$(t_i, t_j) \in R_E \Leftrightarrow \text{The Recognition Engine E assigns } t_i \text{ and } t_j \text{ the same XID}$$

Because E will assign one and only one XID to each transaction it processes, it follows that the binary relation  $R_E$  defined in this way is an Equivalence Relation, i.e.,

1.  $R_E$  is reflexive,  $(t_i, t_i) \in R_E \quad \forall t_i \in T$
2.  $R_E$  is symmetric,  $(t_i, t_j) \in R_E \Rightarrow (t_j, t_i) \in R_E$
3.  $R_E$  is transitive,  $(t_i, t_j) \in R_E, (t_j, t_k) \in R_E \Rightarrow (t_i, t_k) \in R_E$

**Definition 2.2:** If P is a set of subsets of a set T, i.e.,  $A \in P \Rightarrow A \subseteq T$ , then P is said to be a **partition of T** if and only if

$$A \in P \text{ and } B \in P \Rightarrow \text{either } A \cap B = \phi \text{ or } A = B,$$

$$\text{and, } \bigcup_{A \in P} A = T$$

Because the binary relation  $R_E$  defined on particular ordering of T by a Recognition Engine E is an equivalence relation, the set of all equivalence classes of R is a partition  $P_R$  of T, i.e.

If  $P_i = \{t_j \mid (t_j, t_i) \in R\}$ , Then  $P_E = \{P_i \mid 1 \leq i \leq n\}$  is a partition of T

Each equivalence class  $P_i$  represents all of the transactions belonging to the same customer as determined by the Recognition Engine.

**Definition 2.3:** If E is a Customer Recognition Engine, T is a set of transactions,  $\alpha$  is a particular ordering of T, and  $P_E$  is the partition of T generated by the equivalence relation  $R_E$ , then  **$\{E, T, \alpha, P_E\}$  is a Customer Recognition Model.**

Different recognition engines, different transactions sets, or even different orderings of the same transaction set will produce different models. However, the models are considered equivalent if they produce the same partition of the transaction set.

**Definition 2.4:** Two Customer Recognition Models  $\{R, T, \alpha, P_R\}$  and  $\{S, T, \beta, P_S\}$  are equivalent over the same transaction set T if and only if  $P_R = P_S$ .

Note that Definition 2.4 requires that the models be defined over the same set of transactions. However, different engines and different orderings of the transactions comprise different models, which may or may not be equivalent.

As a simple example, suppose that R assigns an incoming customer transaction an XID that is the same as the first previously processed transaction where the Last Names are no more than one character different, and the Street Numbers are the same.

Order ( $\alpha$ )	Transactions (T)	XID
T <sub>1</sub>	(Smithe, 101 Oak St)	A
T <sub>2</sub>	(Smith, 101 Elm St)	A
T <sub>3</sub>	(Smith, 202 Oak St)	B
T <sub>4</sub>	(Smythe, 101 Pine St)	A

**Table 2.1**

Table 2.1 shows that the four transactions processed in the order shown would be classified into two partition classes  $\{T_1, T_2, T_4\}$  and  $\{T_3\}$ . The first transaction would be assigned an XID of “A”. The second transaction would be compared to the first, and because “Smithe” and “Smith” are only one character different, and the street numbers are the same, it would also be assigned “A”. The third transaction has a street number that does not match either the first or second transaction, and would therefore receive a different XID of “B”. Finally, the fourth transaction would be assigned “A” because when compared to the first transaction, “Smythe” is only one character different than “Smithe” and the street numbers are the same.

Order ( $\beta$ )	Transactions (T)	XID
T <sub>4</sub>	(Smythe, 101 Pine St)	A
T <sub>3</sub>	(Smith, 202 Oak St)	B
T <sub>2</sub>	(Smith, 101 Elm St)	C
T <sub>1</sub>	(Smithe, 101 Oak St)	A

**Table 2.2**

On the other hand, Table 2.2 shows the outcome of processing the same set of transactions with the same recognition rules, but reversing the order of processing. In this case, the four transactions are classified into three partition classes  $\{T_1, T_4\}$ ,  $\{T_2\}$ , and  $\{T_3\}$ . In this processing order, the third transaction processed (T<sub>2</sub>) does not match the first transaction (T<sub>4</sub>) because “Smythe” and “Smith” differ by two characters, and does not match the second transaction (T<sub>3</sub>) because the street numbers are different.

**Definition 2.5:** A Recognition Engine R is said to be **Order Invariant** over a set of transactions T if and only if R produces the same partition for every ordering of T.

### 3.0 PARTITION SIMILARITY

Definition 2.4 relates the equality (equivalence) of two Recognition Models to the equality of the partitions they produce. In the same way, the relative similarity of two Recognition Models can be based on the relative similarity of the partitions they produce. However in this case, the definition of similarity between partitions is less clear. A number of similarity “indices” have been developed in statistics in connection with cluster analysis. The primary consideration in the selecting a particular index for an application is the extent to which it provides adequate discrimination (sensitivity) for a particular application. As a starting point in the initial research, the authors have chosen to test three indices, the Rand Index [5], and the Adjusted Rand Index [9], in the initial research, and the TW Index developed by the authors and described in this paper

The T-W Index was designed by the authors to provide an easily calculated baseline measure. The Rand Index and Adjusted Rand Index have been taken from the literature on cluster analysis and recommended for cases where the two partitions have a different number of partition classes [4]. These indices have a more complex calculation than the T-W Index that involves the formula for counting the combinations of n things taken 2 at a time,  $C(n,2)$ . Because transaction sets can be on the order of hundreds of thousand or even millions of records, the combination calculations for the Rand and Adjusted Rand Indices can exceed the limits of single precision for some statistical packages. Moreover, the lack of symmetry in the calculations for these indices requires that either a very large amount of main memory to be available to make all of the calculations in a single pass of the transactions, or that the transactions be sorted and processed twice.

#### *T-W Index*

**Definition 3.1:** If A and B are two partitions of a set T, define  $\Phi(A, B)$ , the Partition Overlap of A and B, as follows:

$$\Phi(A, B) = \sum_{i=1}^{|A|} \left| \{B_j \in B \mid B_j \cap A_i \neq \phi\} \right|$$

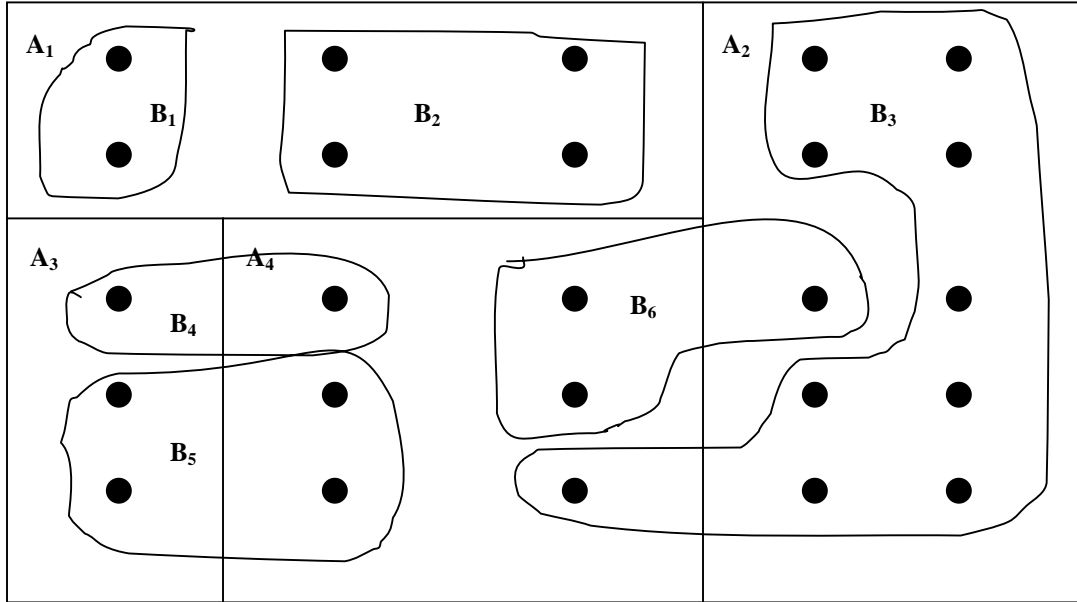
For a given partition class of partition A, it counts how many partition classes of partition B have a non-empty intersection with it. These are summed over all partition classes of A.

**Theorem 3.1:** If A and B are two partitions of a set T, then  $\Phi(A, B) = \Phi(B, A)$ .

Proof: It is easy to see that the definitions of  $\Phi(A, B)$  and  $\Phi(B, A)$  are symmetric.

**Definition 3.2:** If A and B are two partitions of a set T, define  $\Delta(A, B)$ , the T-W Index between A and B, as follows:

$$\Delta(A, B) = \frac{|A| \cdot |B|}{(\Phi(A, B))^2}$$



**Figure 3.1: Array Diagram of Two Partitions A and B**

Figure 3.1 shows a 5 by 5 array of 25 points that represents an underlying set T. The four partition classes of partition A are represented as rectangles labeled A<sub>1</sub> through A<sub>4</sub>, and the six partition classes of partition B are represented by the oval shapes labeled B<sub>1</sub> through B<sub>6</sub>.

The calculation of the Overlap of A and B for this example is:

$$\begin{aligned}
 \Phi(A, B) &= |\{B_j \in B \mid B_j \cap A_1 \neq \phi\}| + |\{B_j \in B \mid B_j \cap A_2 \neq \phi\}| + |\{B_j \in B \mid B_j \cap A_3 \neq \phi\}| \\
 &+ |\{B_j \in B \mid B_j \cap A_4 \neq \phi\}| \\
 &= |\{B_1, B_2\}| + |\{B_3, B_6\}| + |\{B_4, B_5\}| + |\{B_3, B_4, B_5, B_6\}| \\
 &= 2 + 2 + 2 + 4 = 10
 \end{aligned}$$

Therefore,

$$\Delta(A, B) = \frac{|A| \cdot |B|}{(\Phi(A, B))^2} = \frac{4 \cdot 6}{10^2} = 0.24$$

**Corollary 3.2:** If A and B are partitions of the set T, then  $\Delta(A, B) = \Delta(B, A)$  .

**Definition 3.3:** If A and B are partitions of the set T, partition A is said to be a “refinement” of partition B, if and only if

$$A_i \in A \Rightarrow A_i \subseteq B_j \text{ for some } j, 1 \leq j \leq |B|$$

i.e., every partition class of partition A is a subset of some partition class of partition B.

**Theorem 3.3:** If A and B are partitions of the set T, and A is a refinement of B, then

$$\Delta(A, B) = \frac{|B|}{|A|}$$

Proof: If A is a refinement of B, then every partition class of A will intersect only one partition class of B. Therefore

$$\Phi(A, B) = \sum_{i=1}^{|A|} |\{B_j \in B \mid B_j \cap A_i \neq \phi\}| = \sum_{i=1}^{|A|} (1) = |A|$$

Therefore,

$$\Delta(A, B) = \frac{|A| \cdot |B|}{(\Phi(A, B))^2} = \frac{|A| \cdot |B|}{|A|^2} = \frac{|B|}{|A|}$$

From Definition 3.1, it is easy to see that

$$\Phi(A, B) \geq \max(|A|, |B|)$$

Consequently, by Definition 3.2,

$$\Delta(A, B) \leq 1$$

The following Theorem shows that the T-W Index is equal to one only when the partitions are identical.

**Theorem 3.4:** A and B are identical partitions of T, if and only if  $\Delta(A, B) = 1$ .

Proof: Suppose the A and B are identical partitions of T. Then A must be a refinement of B. By Theorem 3.3,

$$\Delta(A, B) = \frac{|B|}{|A|}$$

However, because A and B are identical,  $|A| = |B|$ . Consequently,  $\Delta(A, B) = 1$ .

The converse can be demonstrated by observing that Definition 3.1 requires that  $\Phi(A, B) \geq \max\{|A|, |B|\}$

Any difference between partitions A and B will mean that either  $\Phi(A, B) > |A|$  or  $\Phi(A, B) > |B|$  and consequently,

$$\Delta(A, B) = \frac{|A| \cdot |B|}{(\Phi(A, B))^2} < 1$$

**Corollary 3.5:** If A is any partition of T, and B is the “trivial partition” of T, i.e.,  $B = \{T\}$ , then

$$\Delta(A, B) = \frac{1}{|A|}$$

Proof: Every partition is a refinement of the trivial partition. Therefore by Theorem 3.3,

$$\Delta(A, B) = \frac{|B|}{|A|} = \frac{1}{|A|}$$

**Corollary 3.6:** If A is the “point partition” of T, i.e.,  $A = \{\{t_1\}, \{t_2\}, \dots, \{t_n\}\}$  where each partition class of A contains only one element of T, and B is any partition of T, then

$$\Delta(A, B) = \frac{|B|}{|T|}$$

Proof: The “point partition” is a refinement of every partition. Again by Theorem 3.3,

$$\Delta(A, B) = \frac{|B|}{|A|} = \frac{|B|}{|T|}$$

**Corollary 3.7:** If A is the “point partition” of T, and B is the trivial partition of T, then

$$\Delta(A, B) = \frac{1}{|T|}$$

Proof: Apply Corollaries 3.5 and 3.6 together.

Although the T-W Index will always be greater than zero, Theorem 3.7 shows that it approaches zero for the point partition of an arbitrarily large set T. Therefore, the T-W Index takes on values in the half open interval (0,1].

### ***Rand Index and Adjust Rand Index***

The Rand Index [5] and the Adjusted Rand Index [9] are both commonly used indices to compare clustering results against external criteria [4]. The computation of these indices is best explained using a tabular representation of the overlap between two partitions.

If A and B are two partitions of the set T, the overlap between A and B can be represented in Table 3.1.

A\B	B <sub>1</sub>	B <sub>2</sub>	...	B <sub>n</sub>	Sums
A <sub>1</sub>	C <sub>11</sub>	C <sub>12</sub>	...	C <sub>1n</sub>	S <sub>1*</sub>
A <sub>2</sub>	C <sub>21</sub>	C <sub>22</sub>	...	C <sub>2n</sub>	S <sub>2*</sub>
...	...	...	...	...	...
A <sub>m</sub>	C <sub>m1</sub>	C <sub>m2</sub>	...	C <sub>mn</sub>	S <sub>m*</sub>
Sums	S <sub>*1</sub>	S <sub>*2</sub>	...	S <sub>*n</sub>	S <sub>mn</sub>

**Table 3.1**

In Table 3.1, the row and column entry C<sub>ij</sub> represents the count of elements in the intersection between partition class A<sub>i</sub> of partition A and the partition class B<sub>j</sub> of partition B. Each row sum S<sub>i\*</sub> is equal to the number of elements in the partition class A<sub>i</sub>, and the column sum S<sub>\*j</sub> is equal to the number of elements in the partition class B<sub>j</sub>. The sum S<sub>mn</sub> is equal to the number of elements in the underlying set T.

The calculation of both the Rand Index and Adjusted Rand Index can be expressed in terms of four values, x, y, z, and w, defined as follows:



$$x = \sum_{i,j} \binom{C_{ij}}{2}, \text{ where } \binom{N}{2} = \frac{N \cdot (N-1)}{2}$$

$$y = \sum_i \binom{S_{i^*}}{2} - x$$

$$z = \sum_j \binom{S_{*j}}{2} - x$$

$$w = \binom{S_{mn}}{2} - x - y - z$$

Based on these values

$$\text{Rand Index} = \frac{x + w}{x + y + z + w}$$

$$\text{Adjusted Rand Index} = \frac{x - \left( \frac{(y+x) \cdot (z+x)}{x+y+z+w} \right)}{\frac{(y+z+2x)}{2} - \left( \frac{(y+x) \cdot (z+x)}{x+y+z+w} \right)}$$

The primary difference is that the Adjusted Rand takes on a wider range of values thus increasing its sensitivity.

Transforming the example of Figure 3.1 into tabular form yields Table 3.2 shown below.

A\B	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>	B <sub>4</sub>	B <sub>5</sub>	B <sub>6</sub>	Sums
A <sub>1</sub>	2	4	0	0	0	0	6
A <sub>2</sub>	0	0	9	0	0	1	10
A <sub>3</sub>	0	0	0	1	2	0	3
A <sub>4</sub>	0	0	1	1	2	2	6
Sums	2	4	10	2	4	3	25

**Table 3.2**

Based on these counts

$$x = 1 + 6 + 36 + 1 + 1 + 1 = 46$$

$$y = 15 + 45 + 3 + 15 - 46 = 32$$

$$z = 1 + 6 + 45 + 1 + 6 + 3 - 46 = 16$$

$$w = 300 - 46 - 32 - 16 = 206$$

$$\text{Rand Index} = (46+206)/(46 + 32 + 16 + 206) = 0.84$$

$$\text{Adjusted Rand Index} = (46 - (78*62)/300)/((78 + 62)/2 - (78*62)/300) = 0.5546$$

By contrast,

$$T-W \text{ Index} = 0.24$$

An important aspect of the preliminary research is to determine which one, or possible which combination, of these indices provides an appropriate level of discrimination in comparing the partitions actually generated by Customer Recognition applications involving large volumes of transactions.

## 4.0 CUSTOMER RECOGNITION QUALITY METRICS

Given that Customer Recognition system outcomes can be represented as partitions, and that an appropriate index has been selected to assess the degree of difference between partitions, the next step is to investigate the use of the index to create data quality metrics relevant to customer recognition systems. Having measurements appropriate for critical touch points in a data process flow is an important aspect of any total data quality strategy [2]. For purposes of this discuss, we will simply refer to it as the “similarity index”. The following suggests how a partition similarity index could be applied.

### *Metric for Customer Recognition Consistency*

The following describes three contexts in which a similarity index could provide a type of consistency metric. The first is a comparison between two different recognition systems, and the second is an assessment of changes to a single recognition system. In both cases we hold the transaction set fixed. Experiments 4.1 and 4.2 illustrate these two applications, respectively. A third example (Experiment 4.3) considers the case where the engine is held fixed and the transaction set changes in quality.

#### **Experiment 4.1, Different Engines:**

In this experiment, the first recognition system R is a CDI product based on traditional “merge/purge” approximate string matching technology, and the second system S is a newer CDI product using both matching and a knowledge base of external information about occupancy associations. Both R and S are used as the recognition engine in Customer Recognition applications. T is a fixed set of ordered customer transaction.

The following is a comparison of the partitions A and B created by R and S respectively.

Statistic	A	B
Record Cnt	673,003	673,003
Class Cnt	175,527	136,795
Single Cnt	112,857	62,839
Avg Class	3.83	4.92
Max Class	110	80

Similarity Index Results

Index	Value
T-W	0.4339
Rand	0.9998
Adj Rand	0.8104

In this experiment, the second partition B shows more grouping in that it has fewer partition classes than the partition A created by engine R that relies entirely on string matching. On average the partition classes created by the knowledge-assisted engine S are larger and there are fewer singleton classes. These all indicate that the knowledge-assisted recognition engine S creates groups more transactions. Presumably this can be attributed to the additional knowledge that allows some of the “match only” classes of R to be consolidated into a single class using external knowledge. For example, partition A may contain two classes, one with two transactions, {“John Jones, 123 Main”, “J. Jones, 123 Main”}, and another with one transaction {“John Jones, 345 Oak”}. However if external knowledge indicates that “John Jones” has moved from “123 Main” to “345 Oak”, then these three transactions would be in the same class of partition B, i.e., the class {“John Jones, 123 Main”, “J. Jones, 123 Main”, “John Jones, 345 Oak”}.

Although this may be an expected result, the indices only indicate the degree to which R and S generate different partitions, with the profile showing that R makes fewer associations (on average) than S. The measurement does not indicate which, if either, makes more correct associations. Furthermore, the three indices vary widely on the degree of similarity with the Rand indicating rather strong similarity, the T-W a fairly strong difference, and the Adjusted Rand somewhere in the middle.

**Experiment 4.2, Changes to the Same Engine:**

Having a way to measure the impact of changes to the Recognition Engine can also be very useful in assessing recognition quality, especially in the initial phases of a system implementation. In this scenario, the input transactions are held fixed, and the grouping is performed twice, once before the change (R), and once after the change (S). The similarity index provides a metric for assessing the change in groupings that can be attributed to the change in the recognition engine.

In this experiment, R is the April release of a knowledge-based CDI product that is released monthly and used in customer recognition applications. S is the May release of the same product. T is a fixed set of ordered customer transaction.

The following is a comparison of the partitions A and B created by R and S, respectively.

Statistic	A	B
Record Cnt	17,778	17,778
Class Cnt	3,218	3,223
Single Cnt	1,271	1,222
Avg Class	5.53	5.52
Max Class	63	63

Similarity Index Results

Index	Value
T-W	0.9972
Rand	0.9999
Adj Rand	0.9989

Although the partition of the new release (B) shows increase clustering in terms of fewer singleton classes and fewer classes overall, the average class size has slightly decreased. This would be an expected result if we believe that in a knowledge-based approach, knowledge about the entities in a fixed set of transactions increases over time, i.e., there is a time-latency in knowledge gathering. Under this assumption, and given that the transactions are held fixed in time, one could expect that knowledge about these transactions (customers) will increase over time, and that the engine’s ability to connect transactions for the same customer will improve. In this particular measurement, all three indices point to a very high degree of similarity (consistency) between the partitions produced by the two releases, and that the second release brings together slightly more transactions. However this measurement only points to stability between the two releases, and does not prove that the second release is more or less accurate in grouping than the first.

**Experiment 4.3, Changes in Input Quality:**

Here the Recognition Engine is held fixed and the transaction set is intentionally degraded in quality. For experimental purposes, the change (error) can be introduced at a fixed rate.

In this experiment, R is a knowledge-based CDI product used in Customer Recognition applications is held fixed. R identifies individual customers based on name and address (occupancy). First, R processes the ordered transaction set T to create the partition A. Next, the quality of T is deliberately degraded by removing all vowels from the names in 800 of the 17,788 transaction records (4.5%), and R processes the degraded transactions to create the second partition B.

The following is a comparison of the partitions A and B created by R and S respectively.

Statistic	A	B
Record Cnt	17,788	17,788
Class Cnt	3,218	3,332
Single Cnt	1,271	1,675
Avg Class	5.53	5.34
Max Class	63	60

**Similarity Index Results**

Index	Value
T-W	0.6665
Rand	0.9998
Adj Rand	0.8782

In this scenario, the effect of quality degradation is evident. Even though more classes are created from the degraded transactions, the number of singleton classes has increase dramatically. These represent records that were formerly integrated into larger classes, but due to degradation cannot be matched and become outliers. The average size of the classes has also decrease significantly. Again, the T-W index is the most sensitivity to this change, whereas the Rand indicates almost complete similarity.

***Metric for Customer Recognition Accuracy***

The “touch stone” metric would be the use of a similarity index to measure the accuracy of customer recognition. If A and B are both partitions of the same ordered transaction set T, and if A represents the “correct partition of T” (i.e., is a benchmark), and B represents the partition of T imposed by some recognition system R, then the similarity index can provide an objective measure of the accuracy of the

recognition system  $R$ . Because all of the indices described above have the characteristic that they take on the value of 1 when the partitions are identical, and values less than 1 as the partitions become dissimilar, then the value of the similarity index times 100 (or some normalized transformation of the similarity index) can be used as an accuracy metric.

In the case that the correct partition of  $T$  is known, the similarity measure can also be compared to measures developed to assess the effectiveness and efficiency of duplicate detection and information retrieval in general, such as, Precision and Recall Graphs [1].

Even though it is evident how one could create an accuracy metric for customer recognition using a similarity index, it is less obvious how to create the benchmark of correct groupings. In practice, this can be very difficult to do. The authors have experience in using the following methods to create a benchmark.

In the case of recognition systems that rely only on matching, it is possible to create correct grouping by manually inspecting the records and making an expert judgment about which records belong in each class. The primary limitation of this method is the effort required to create a benchmark of any significant size. In addition, experts don't always agree, and this method may require some type of arbitration, such as a voting scheme.

However in the case of knowledge-based recognition systems, manual inspection is not enough. For example, the mere inspection of two consumer records, such as "Jane Smith, 123 Oak" and "Jane Jones, 456 Elm", cannot establish if they should or should be in the same class without knowing if these represent the same customer who has married and moved to a new address. In this situation, creating a benchmark requires accurate information about changes in addresses and changes names that is best obtained from the customer's themselves. Such a benchmark can be both expensive and difficult to create, even for a relatively small sample [6]. Even attempts to create these by having company employees volunteer this information have been largely abandoned due to privacy and legal concerns.

The authors are currently exploring a third method that is somewhat of a hybrid of the two just described. It is based on the observation that most transactions into identification classes based on simple matching. If the transactions are first grouped according to a conservative, agreed upon match algorithm, then these classes can be "subtracted" from the overall partition as being "correct" without further analysis. Hopefully this leaves a much smaller number of transactions to be analyzed, and if necessary, investigated to establish their correct grouping.

## **5.0 CONCLUSIONS AND FUTURE WORK**

The algebraic approach of characterizing Customer Recognition systems as partitions of ordered transaction sets is proving to be useful in creating metrics for quality assessment. In addition to providing an easily understood model, it also opens the door to utilizing the research literature already available related to cluster analysis.

Although the preliminary experiments indicate that the T-W Index provides even more discrimination than the Rand or Adjusted Rand Indices, and is easier to calculate, further testing on a broader range of recognition outcomes needs to be done before abandoning these or other techniques. Because the T-W Index is not weighed in terms of the degree of overlap, there may cases where the T-W Index does not perform as well as other methods.

Future work plans include:

- Investigation of alternative or modified approaches to an accuracy metric that reduces the cost and effort to obtain a benchmark. For example, the approach described above that eliminates consideration a set of “given” or “assumed” agreements, such as exact or close matching, and focuses effort on validating the exception to this rule, or even a sample of the exceptions.
- Further refinement of metric definitions, such as, confidence intervals for accuracy, and tests for significance of Index differences.
- A wider range of experiments on the sensitivity of recognition outcomes to the quality of input data that include different types of quality problems.
- Further validation of preliminary results based on measurements of actual customer data.

## **BIBLIOGRAPHY**

- [1] Bilenko, M. and Mooney, R.J. “On Evaluation of Training-Set Construction for Duplicate Detection.” *Proceedings: ACM SIGKDD-03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, 2003.
- [2] Campbell, T. and Wilhoit, Z. “How’s Your Data Quality? A Case Study in Corporate Data Quality Strategy.” *Proceedings: International Conference on Information Quality*, MIT, 2003.
- [3] Huang, K., Lee, Y.W., and Wang, R.Y. *Quality Information and Knowledge*, 1999, Prentice Hall.
- [4] Hubert, L. and Arabie, P. “Comparing Partitions.” *Journal of Classifications*, 1985, pp. 193-218.
- [5] Rand, W.M. “Objective criteria for the evaluation of clustering methods.” *Journal of the American Statistical Association*, 1971, **66**, pp. 846-850.
- [6] Talburt, J.R. “Shared System for Assessing Consumer Occupancy and Demographic Accuracy”, *Proceeding: International Conference on Information Quality*, MIT, 2003.
- [7] Talburt, J.R., Wang, R.Y., et.al. “Customer-Centric Information Quality Management”, MITIQ White Paper, May 24, 2004,  
<http://mitiq.mit.edu/Documents/CCIQM/CCIQM%20White%20Paper.pdf>. [8] Wang, R.Y., Ziad, M., and Lee, Y.W. *Data Quality*, 2001, Kluwer Academic Publishers.
- [9] Yeung, K.Y. and Ruzzo, W.L., Details of the Adjusted Rand Index and Clustering Algorithms, Supplement to the paper “An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data”, *Bioinformatics*, 2001, 17 (9) pp. 763-774,  
<http://faculty.washington.edu/kayee/pca/supp.pdf>