Deciding Confidence for Identity Resolution in Closed and Open Universes of Entity Identity Information

Fumiko Kobayashi, John R. Talburt
Department of Information Science
University of Arkansas at Little Rock
Little Rock, AR, USA
fxkobayashi@ualr.edu, jrtalburt@ualr.edu

***Abstract*- Entity Identity Information Management (EIIM) systems provide the information technology support for Master Data Management (MDM) systems. One of the most important configurations of an EIIM system is identity resolution. In an identity resolution configuration, the EIIM system accepts a batch of entity references and returns the corresponding entity identifiers. However, these batch EIIM systems lack in providing easily accessible identity information. To address this, the EIIM system is extended with an Identity Management Service (IMS) [1] to decouple identity resolution from batch EIIM processing which allows it to move into an interactive realm. Due to the uncertainty of the information provided by the requester, the system may match the reference to many different Entity Identity Structures (EIS), and the identifier returned by the system may not be the correct one. To assist the client, the IMS provides a second output called a confidence rating. The confidence rating is a measure of the likelihood that the returned identifier is accurate. The confidence rating is related to the match score, but must also consider other factors including whether the entity identity information resides in a closed or open system. This paper discusses a model for generating confidence ratings based on an assessment of the differences between the match scores for competing EIS in either a closed or open universe. The paper also includes some preliminary results from experiments performed on a production IMS system supporting student MDM.**

Keywords- Entity Identity Information Management; Identity Resolution; Master Data management; Open Universe; Closed Universe; Confidence rating

I. INTRODUCTION

Identity Resolution (IR) is the process of determining if an entity reference refers to the same entity as one of the Entity Identity Structures (EIS) under management in an Entity Identity Information Management (EIIM) system. IR is sometimes called "entity recognition" because the system is being asked if the input entity reference can be recognized as one of the entities already under management.

ER is the process of determining whether two references to real-world objects in an information system are referring to the same object, or to different objects [2]. Real-world objects are identified by their attribute similarity and relationships with other entities. Some examples of attributes for person entities are First Name, Last Name, and Social Security Number (SSN). For place or location entities the attributes might be Latitude, Longitude, Description, or postal address. ER has also been studied under other names including but not limited to record linkage [3], deduplication [4], reference reconciliation [5], and object identification [6].

ER is a key task in data integration where different systems or data sources provide information for a common set of entities. ER has its roots in Customer Relationship Management (CRM) where it is often referred to as Customer Data Integration (CDI) [7]. The need for accurate and efficient ER is a necessity with the amount of data that is able to be collected and stored with current levels of technology. ER research is also driven by the need to share entity identity information across independently governed organizations in many areas such as education, healthcare, and national security.

Previous Information Quality (IQ) research [8] [9] has extended ER into the larger context of Entity Identity Information Management (EIIM) that includes the creation and maintenance of persistent data structures to represent the identities of external entities [2]. The overall goal of EIIM is to allow the ER system to achieve entity identity integrity, a state in which two conditions hold [10].

1. Each identity structures corresponds to one, and only one, real-world entity
2. Distinct identity structures correspond to distinct real-world entities.

Entity identity integrity is another way of stating the Fundamental Law of Entity Resolution [2] which requires that two entity references should be linked if, and only if, they are equivalent where equivalence means both reference the same real-world entity.

In the current model of EIIM, the configurations to maintain the EIS operate primarily in an offline batch mode. In general the EIIM model is focused on the processes necessary to achieve and maintain entity identity integrity of the EIS under management in systems Identity Knowledgebase (IKB). EIIM provides the tools to support the complete life cycle of identity information.

In the EIIM model, the only configuration that does not modify the IKB is the IR configuration. In an IR process each entity reference input into the system is resolved

 ISBN: 978-1-61208-395-7

against a previously defined set of identities represented as some type of EIS. When the pre-defined identities represent master data of an organization such as employees, customers, or products, IR becomes an important component of Master Data Management (MDM). MDM and IR in a broad set of applications including business, health care, education, law enforcement, and the military. In a business context of MDM of customer information, IR is sometimes called "customer recognition" [2].

The IR operation is intended to provide access and use for information. The idea of decoupling IR from the batch EIIM system [1] allows more robust access to the entity identity information. The decoupling is done through an interactive Identity Management System (IMS) to interactively access the IKB maintained by the EIIM system.

In this paper, Section I introduces the terminology and concepts required to understand, define, and discuss the goals and approaches used in this research. In Section II, existing research and limitations addressed by this paper are discussed along with the importance of this research for business intelligence. Section III defines the specific problem addressed in this paper and the method used. In Section IV, the concept of candidate selection is explained and an extension to standard probabilistic scoring algorithms is defined which provides improved selection. Section V defines the differences between closed and open universes of information and how assumptions inherent in each universe alter the application of confidence rating for an EIS. In Section VI, the formula for calculation of and the method for applying δ (delta) are defined. Section VI also provides the final algorithms for calculating the confidence rating in closed and open universes of information. In Section VII presents the experimentation performed to gauge the accuracies of the new unique ratio (UR) score algorithm and the δ application. Section VIII concludes and summarizes the paper and proposes future research.

II. EXISTING RESEARCH

EIIM systems are designed in such a manner to provide a robust framework for the maintenance and management of entity information over time in a single batch system [11]. EIIM focuses mainly on the capture, update, and store phases on the CSRUD (capture, store and share, resolve and retrieve, update, and dispose) MDM lifecycle [9]. For practical use, the Resolve and Retrieve Phase is the most important of all the of the CSRUD MDM life cycle phases. Resolving an entity reference to its correct entity (EIS) is the primary use case for MDM. It's this resolve and retrieval that provides actual value to the client systems.

Quantifying the reliability of a resolved entity identifier is an important problem that is not addressed by EIIM. The reliability of identification will vary from inquiry-to-inquiry depending upon the depth, breadth, and context of the match to the EIS in then identity knowledgebase. In order to provide guidance to the inquiring client system, the IMS should compute a confidence rating for each inquiry providing the client system with an estimate of the likelihood a resolved entity identifier is correct.

This research provides methods that fill the void that was left in regards to design for the Resolve and Retrieve Phase of the CSRUD MDM life cycle. The increased accuracy of resolution and the confidence rating are vital for the area of Business Intelligence. This is because it provides a gauge when selecting meaningful and useful information from an IKB for business analysis purposes.

III. PROBLEM DEFINITION

There are many factors that contribute to the accuracy of an IMS. The main factor is the quality and reliability of the input reference. However, the number of input attributes provided, the domain of the information being searched, and others factors also have a bearing. While certain factors cannot be controlled, their impact on accuracy can be mitigated through observational testing and matching algorithms. This research focuses primarily on a model for calculating the confidence in the accuracy of the entity identifier returned by an IMS. The model takes into consider two broad categories of IMS, open universe IMS and closed universe IMS.

When IR is reconfigured into an interactive mode, the user still expects the system to provide the same type of results that a batch system provides. This expectation is that for every input, there should be one decisive identifier returned. The problem of determining the correct identifier to return revolves around two issues: the uncertainty of the user provided input and the universe of the entity identity information.

Imagine the situation in which the entity identity information comprises records with 14 attribute values. When a user provides a subset of these attributes values as input, the goal of the system is to provide the most accurate response possible. This is difficult to gauge when the user only provides a small percentage of the attribute values, i.e. in this situation, 3 or 4 attributes out of 14. In addition, there are no constraints or validations applied to the input data so it is often "dirty" input data. The issue of low-quality input data is further addressed through the introduction of the delta range discussed later.

Not only the entity identity information itself but also the universe in which that information resides is an important consideration. In this research, the concept of a closed and open universe have been applied to the entity identity information to allow for baseline assumptions to be made regarding model. These assumptions help determine the best response.

Through the application of the closed and open universe concept, this paper shows how it is possible to

 ISBN: 978-1-61208-395-7

provide the requestor with an additional piece of information alongside the response that will allow them to know to which degree they can accept the response as fact. This is being referred to as the confidence rating.

The overall method performed in the research consists of four steps that must each be understood to accurately define a confidence. These are as follows:

1. Scoring and Candidate Selection
2. Universe Assumption application
3. δ (delta) application
4. Confidence application

Each of these is detailed in the next few sections.

IV. CANDIDATE SELECTION AND SCORING

In an ER system, it is impractical to perform matching on all the reference in an IKB against an input reference. To reduce the number of comparisons that are required custom indexing is used as a form of blocking [2]. This index allows the attributes values in the input reference to be used to quickly populate a candidate list [2] against which the more complex matching algorithms can be applied.

Once the candidate list is selected, the matching is applied in the form of a scoring algorithm. Scoring is the process of assigning a numerical value (normalized between 0 and 1) that predicts the probability that a reference in the IKB matches the input reference. The score is expensive in terms of processing but can be run on all of the references that were specified as possible matches by the index. After scores are calculated for each reference, any scores that meet the predefined threshold can then be used as the final set on which confidence is calculated.

In an interactive IR system, it was found that standard score algorithms [12] were not suitable. For more accurate score generation, the algorithm needs the ability to use attribute data that exists in the candidate list to skew the final results. To accomplish this, the unique ratio (UR), a basic profiling statistic, was found to provide large gains in accuracy when factored into the score algorithm.

The idea behind the modified score algorithm is that in a result set, exact matches with a higher unique ratio (UR) should hold a higher weight than an exact match with a low unique ratio (UR). A unique ratio is calculated as:

$$UR_i = \frac{UC_i}{RC} \quad (1)$$

Where

- UC = number of unique attribute values in a given attribute (inclusive of blanks)
- RC = record count in the candidate set

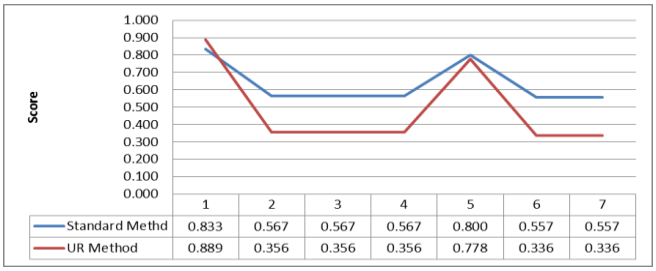
The following is the modified score algorithm with the UR factored in:

$$URScore(\mathrm{IR}, \mathrm{CR}) = \frac{\sum_{i=1}^{n} LED(IR_i, CR_i) * UR_i * AF_i}{\sum_{i=1}^{n} UR_i * AF_i} \quad (2)$$

The attribute flag (AF_i) is calculated for each reference that is processed by the system. The attribute flag stores a simple Boolean value (1 or 0) and is calculated for each attribute as:

$$AF_i = \begin{cases} 1 \; if \; i \neq \emptyset \\ 0 \; if \; i = \emptyset \end{cases} \quad (3)$$

The purpose of this modified score algorithm is to value unique differences in a search over repetitive values. It can only be applied to the final result set to allow the optimal result to be selected. The algorithm uses the result set values as a group to skew the final calculation instead of allowing the records to be valued independently.



	1	2	3	4	5	6	7
Standard Methd	0.833	0.567	0.567	0.567	0.800	0.557	0.557
UR Method	0.889	0.356	0.356	0.356	0.778	0.336	0.336

Figure 1. UR vs. Standard Scoring.

Figure 1 shows the results of a request that has 7 references in the final candidate list. This shows how the UR score method more clearly classifies matches (or possible matches) from other false positives by widening the gap in their match score results.

Figure 1 shows that the Standard Method placed both EIS 1 and 5 into the same bucket. When the UR score algorithm was applied however, the gap in the scores for EIS 1 and 5 became profound enough that they now fall within 2 separate buckets. This clearly labels EIS as the winner.

V. UNIVERSE OF ENTITY IDENTITY INFORMATION

In terms of entity identity information, the universe defines the type of information that is being accessed. That is, if the entity identity information is in a controlled environment in which the data entering the system is known, or if the entity identity information being entered into the system has little or no oversight beyond structure. These two types of entity identity information are classified into closed universe information and open universe information respectively. Depending on which a set of entity identity information falls into, a certain level of assumption can be applied to the data to assist in the decision making.

A. *Closed Universe*

In an Interactive IR system, a closed universe defines a set of entity identity information in which the requestor can be certain that the entity they are searching for exists within

the IKB and that the entity is unique. This means that the following assumptions can be applied to entity identity information that falls within this category:

- The record being searched for is known to exist within the universe
- There is always one and only one perfect match
- The data and EIS are controlled and updated only by an knowledge expert

Due to the type of data, these assumptions are valid and allow for leaps to be taken in the assignment of a final confidence rating. Specifically, with these assumptions it can be inferred that even if the top score generated is very low, it is still the best match and could be assigned a high confidence rating.

An example of a closed universe could be that of a university enrollment system. In such a system, a professor could be certain that a student attending their class will exist in this universe and that there should only be a single entity for that student. If the professor needs to look up information on said student, they should have confidence that the entity identity information retrieved from their search should be the entity they are looking for.

This means that for professor's class A, and student registration system B:

$$A \subseteq B \tag{4}$$

B. *Open Universe*

In an Interactive IR system, an open universe defines a set of entity identity information in which the requestor is uncertain if the entity they are searching for exists. The following assumptions can be made regarding entity identity information that falls within this category:

- The contents of the entity identity information in the open universe contain no oversight or restrictions beyond the particular layout of the attribute data.
- An entity may or may not be present when searching.

These two assumptions introduce doubt to the user of the information as they cannot be certain that any results they receive are the correct result. These assumptions and the corresponding doubt require confidences in an open universe to endure an additional consideration. A hard cut off for confidences is applied in the form of a threshold. This means that unlike the closed universe, if the top score is very low, no assumption can be made and this reference will be assigned no confidence.

An example of an open universe could be that of a criminal IKB. Many organizations could contribute information to the IKB ranging from local to federal level. However, even though the IKB contains vast amounts of information, when information of a suspect is being processed through the system there is no guarantee that the information exists within the IKB. If information is found, it may consist of multiple records that match the search criteria but aren't actually the correct EIS.

In both closed and open universe, the overall goal is to mitigate some of the uncertainty in results provided to the requestors. This can be accomplished through a confidence rating.

VI. CONFIDENCE RATING

A confidence rating is the primary focus of this research. It is a numerical value between 0 and 1 that is returned to a requestor along with the final match result for the request. The confidence rating informs the requestor of the likelihood that the match result was the correct EIS for the request. The confidence rating can be used by the requestor to either accept the response or make another request providing more information. As noted previously, the main difference between determining confidence in a closed and open universe is the threshold that is applied to the open universe under which no confidence can be assigned to a reference.

The calculation of a confidence rating is the last step of reference selection in an interactive IR system but in order to accurately calculate the confidence rating a δ (delta) calculation must be applied to account for uncertainty in the selection.

A. *δ (delta)*

A naïve approach to confidence rating calculation is to simply assign a confidence of 1 to the top match scored value. However, there is not always a single top scored candidate. Also depending on the universe model, it is not always reasonable to assume the top score is the best match. From this insight, it was decided to build "buckets" and assign the match scores to the buckets.

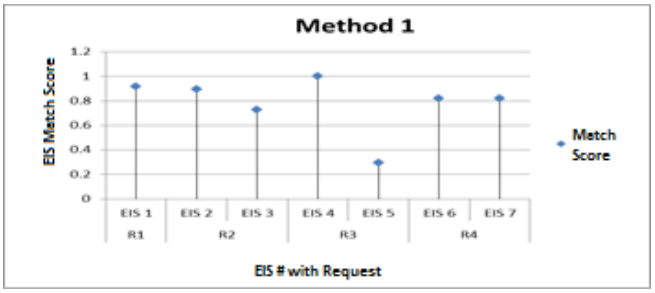


Figure 2. Methods 1 – Buckets.

The information is Figure 2 is as follows:

- Every dot represents an EIS
- Each EIS noted has at least one matching record
- Rows represent a corresponding score

In this method, the EISs in the highest bucket were selected as the candidate and assigned a confidence as a ratio of the number of EIS.

$$M1Con = \frac{1}{EC} \tag{5}$$

 ISBN: 978-1-61208-395-7

In the Method 1 Confidence (M1Con) calculation, EC is the count of EIS in the top used bucket. However, this is inefficient because an EIS with a match score of 99.9% would be assigned the same confidence as a match score of 90.2% with no regard to the number of attributes or other considerations.

This problem can be addressed by applying a δ (delta) value. The δ is a number used as a sliding window (bucket) for the top candidate selection when calculating confidence.

By applying δ, like confidences could be assigned to close scored references even if they fall within different fixed buckets. To illustrate this concept, Figure 3 shows that δ=A would assign equal confidence ratings to EIS 1 and 4, δ=B would assign equal confidence ratings to EIS 1, 4, and 7, and δ=C would assign equal confidence ratings to EIS 1, 2, 4, and 7. This leads to the question of how to determine a δ in a systematic and accurate way.

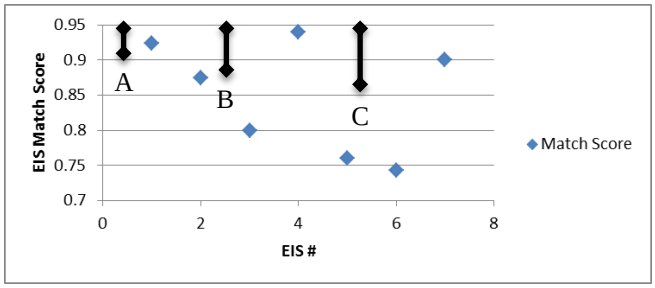


Figure 3. δ (delta) Range.

The method for calculating δ is based on the number of input attributes. The value of δ varys based on the attribute count of the IKB and the attributes contained in the request being made. This is to accommodate requests of different sizes and accuracies, i.e. if a user only provides 2 attributes, the δ would be larger as not enough information was provided to generate an accurate response. For a user that provides 12 attributes (out of 20) the δ should be much smaller as the user provided more information and gets a more accurate decision from the system. The formula for δ is as follows:

$$\delta = 0.1 * \frac{\frac{B}{A * A - 1} + (A - 1) * 0.8}{B * 2 + A} \quad (6)$$

Where:

- A= total attribute count for a given IKB
- B= Total attribute count for a given input request into the IKB

When δ is calculated for every combination of request attribute count and total attribute count and then plotted, the resulting is a δ curve which visually illustrates the expectations for δ values. Figure 4 shows the δ curve for an IKB consisting of 99 attributes.

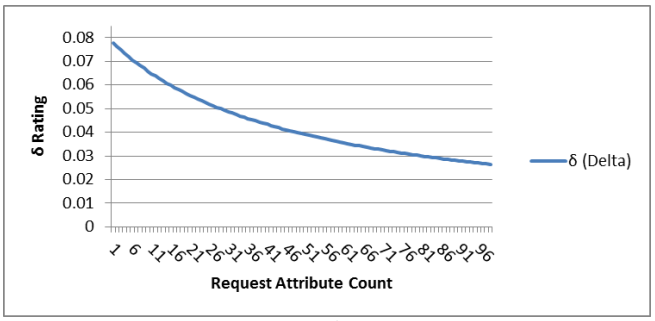


Figure 4. δ Curve.

Once the δ is calculated for a request, it can be applied to more accurately group references when calculating the confidence rating.

B. *Calculating the Confidence Rating*

The last step in the process is to calculate the confidence rating for the result based on the information compiled in the previous steps. In a closed universe, this is done by the following:

1. Select the EIS that has been assigned the highest score based on the UR score function
   a. Use this as the upper bound of the δ range
2. Subtract the calculated δ from this score value.
   a. Use this as the lower bound of the δ range.
3. Count the number of EIS that meet the following
   Upper Bound≥EIS Score≥Lower Bound
   a. Assign this count to variable EC
4. Assign a confidence equal to 1⁄EC to each of the EIS that fall within the δ range
5. Return the EIS identifier and the corresponding confidence rating for the EIS with the highest score value to the requestor.

In an open universe, the threshold for confidence must be considered. The following is the modified method:

1. Compare the highest UR score (urs) to the threshold
   a. If urs<threshold
      i. Assign a confidence rating of 0 to top scored EIS and return the EIS identifier and confidence to requestor.
   b. If urs≥threshold
      i. Select the EIS that has been assigned the highest score based on the UR score function
         1. Use this score as the upper bound of the δ range
      ii. Subtract the calculated δ from this score value.
         1. Use this as the lower bound of the δ range.
      iii. Count the number of EIS that meet the following
         Upper Bound≥EIS Score≥Lower Bound
         1. Assign this count to variable EC

 ISBN: 978-1-61208-395-7

iv. Assign a confidence equal to 1/EC to each of the EIS that fall within the δ range
v. Return the EIS identifier and the corresponding confidence rating for the EIS with the highest score value to the requestor.

VII. EXPERIMENTATION

The testing of the modified rating algorithm in a closed universe was done on real data for a student identity system management system. This data consisted of approximately 1 million EIS each containing 39 attributes. From this data three random sets of EIS were pulled making 3 closed universes consisting of 100, 200, and 1,000 EIS. A truth set for each of these was created and each of the search results was compared against this for accuracy.

On each of these three IKBs, 840 searches were performed and the resulting EIS and its confidence were checked against the truth set. These 840 searches consisted of 70 requests for 1 to 12 attributes per request. The reason for this was to identify the point at which accuracy gains stopped outweighing the need for additional attributes.

It is important to note that in an ER system, match decision are applicable if and only if the attributes used for the decisions are classified as identifying attributes. For the data tested, 12 of the 39 attributes were selected for testing once they were identified as information that could be used to accurately identify a match.

The results were recorded for each set of 840 searches and then the results were averaged to generate an accuracy estimate for each attribute level for each IKB. The averages are shown in TABLE I.

TABLE I. ALGORITHM ACCURACY AVERAGES

Attribute Count	**Cluster Count**			
	100	200	1000	Average
1	11.429%	7.143%	1.429%	6.667%
2	14.286%	10.000%	7.143%	10.476%
3	21.429%	22.857%	21.429%	21.905%
4	30.000%	28.571%	37.143%	31.905%
5	47.143%	45.714%	58.571%	50.476%
6	84.286%	87.143%	84.286%	85.238%
7	94.286%	94.286%	91.429%	93.333%
8	95.714%	94.286%	92.857%	94.286%
9	97.143%	95.714%	95.714%	96.190%
10	97.143%	95.714%	95.714%	96.190%
11	98.571%	97.143%	97.143%	97.619%
12	98.571%	98.571%	97.143%	98.095%

When these averages are graphed as shown in Figure 5, it is evident that the accuracy of the algorithm grows as more attribute values are provided. This result was expected.

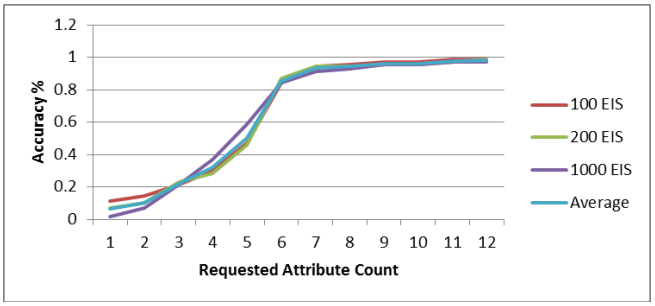


Figure 5. Algorithm Accuracy.

The validation of the modified rating algorithm is that with 6 attributes, no matter the size of the universe, the algorithm made an accurate decision returning the correct EIS 89% of the time.

VIII. CONCLUSION/SUMMARY

It was found that once Identity Resolution is moved into the interactive realm, additional complications are encountered during processing which requires a more accurate method of deciding a singular "best match". This best match is expected by the user of the system as they assume an interactive IR system should return results comparable to that of a batch system. With the introduction of a confidence rating, the system can provide the requestor with the most confident EIS in relation to the candidate list of matches for a request. The confidence rating is a relative measure from the initial match score.

During testing, it was found that the domain of the information must be considered as confidences should act differently in both and closed and open universes due to accepted assumption about the two realms of entity identity information. To address this issues a modified score algorithm was created to utilize the data contained in the final match set to augment the final scores and provide a more accurate confidence decision. Through this modified score algorithm in conjunction with the δ range, it was identified that there is a threshold on data at which the accuracy will peak beyond 90%. Depending on the number of identifying attributes in the IKB, this number may vary but for the experimentation done it required was 40% of the attributes to be provided to achieve accuracy of 90% or above.

The experimentation performed during this research showed that the system was able to determine the correct match with a high level of accuracy. The accuracy of the UR Score provided an almost 10% gain is accurate selection when compared to other standard scoring algorithms. When the UR Score was combined with the δ range, the resulting confidences were determined to accurately represent the trustworthiness of a returned EIS. This increase in accuracy and the ability to rate and return a confidence has many

 ISBN: 978-1-61208-395-7

applications in the Business Intelligence domain. These including that of increased trust in the results of business analysis performed on entity identity information requested from an IMS system.

Future research into the confidence rating will focus on an open universe. The expanded research will consider the use of neural networking and other graph theory concepts to approve selection amongst match candidates.

REFERENCES

[1] F. Kobayashi and J. R. Talburt, "Decoupling Identity Resolution from the Maintenance of Identity Information," in *International Conference on Information Technology (ITNG)*, Las Vegas, NV, 2014.

[2] J. R. Talburt, Entity Resolution and Information Quality, Burlington, MA: Morgan Kaufmann, 2011.

[3] H. B. Newcombe, J. M. Kennedy, S. J. Axford and A. P. James, "Automatic Linkage of Vital Records," *Science*, vol. 130, pp. 954--959, October 1959.

[4] S. Sarawagi and A. Bhamidipaty, "Interactive deduplication using active learning," in *KDD '02*, Edmonton, Alberta, Canada, 2002.

[5] X. Dong, A. Halevy and J. Madhavan, "Reference reconciliation in complex information spaces," in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, Baltimore, Maryland, 2005.

[6] S. Tejada, C. A. Knoblock and S. Minton, "Learning object identification rules for information integration," *Inf. Syst.*, vol. 26, pp. 607--633, dec 2001.

[7] J. Dyché, E. Levy, D. Peppers and M. Rogers, Customer Data Integration: Reaching a Single Version of the Truth, Wiley, 2006.

[8] R. Y. Wang and D. M. Strong, "Beyond accuracy: what data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, pp. 5--33, mar 1996.

[9] Y. Zhou and J. Talburt, "Entity identity information management (EIIM)," in *2011 International Conference on Information Quality (IQIC11)*, Australia, 2011.

[10] Y. Zhou and J. Talburt, "The Role of Asserted Resolution in Entity Identity Management," in *2011 International Conference on Information and Knowledge Engineering (IKE'11)*, Las Vegas, Nevada, 2011.

[11] Y. Zhou, "Modeling and Design of Entity Identity Information in Entity Resolution Systems", Ph.D. dissertation, Dept. Info. Sci., Univ. of Arkansas at Little Rock, Little Rock, AR, 2012.

[12] F. Kobayashi and J. R. Talburt, "Probabilistic Scoring Methods to Assist Entity Resolution Systems Using Boolean Rules," in *Proceedings of the 2013 International Conference on Information and Knowledge Engineering (IKE'13)*, Las Vegas, NV, 2013.

 ISBN: 978-1-61208-395-7