Risk Assessment of Error-Prone Personal Information in Data Quality Tools

Cihan Varol
Computer Science Department, Sam Houston State University, Huntsville, TX 77341, USA
Email: cvarol@shsu.edu

Coskun Bayrak
Computer Science Department, University of Arkansas at Little Rock, Little Rock, AR, 72204, USA
Email: cxbayrak@ualr.edu

***Abstract*—In order to assist companies dealing with data preparation problems, different approaches are developed to handle the dirty data. However, these firms are not able to predict the final outcome from the customer data, before running all the business process. This gives rise to an extra cost for the company at the end, if the data is highly corrupted. Therefore, in this paper; we propose a framework to estimate the propagation of the error through a data quality tool. Since data quality tools are a variation of the workflows, we have based our modeling on workflow schema. The reliability and the risk propagation parameters are introduced for the sequence, parallel split and conditional type of control properties that can be seen in a data quality tool. At the end, a business model is introduced and an experiment with the proposed model is given as a proof-of-concept.**

***Index Terms*—business process, data cleansing, data quality, risk propagation, workflow**

I. Introduction

Processing customer information in a standardized and accurate manner is known to be a difficult task. Data collection methods vary from source to source by format, volume, and media type. Therefore, it is advantageous to deploy customized data hygiene techniques to standardize the data for meaningfulness and usefulness based on the organization [1].

Because of human errors, data collection methods can often produce incorrect and meaningless results with respect to personal names. The problem of devising algorithms and techniques for automatically correcting words in text has been a perennial research challenge since the 1960s. However, often the use of data administrators or a tool that has limited capabilities to correct the mistyped information can cause many problems. Moreover, most of these techniques are particularly implemented for the words used in daily conversations [2].

Since quality of the information affects the success of a business, data quality (DQ) companies are tending to provide reliable products for their clients. Because of the cost of the operation, understanding the impact on customer satisfaction caused by ill-defined/dirty data (misspelled or mistyped data) is a challenge that needs to be considered by these companies. With that kind of a feature, companies not only predict the outcome of the product, but also monitor or re-design their cleaning structure based on the statistical results. Moreover, they would be able to discuss the estimated success rate with their clients, before running the process. This will remove any possible headaches that they will get from the client when the results are not satisfactory.

Because of the stated facts discussed above, we are introducing a framework which estimates the risk propagation of the ill-defined personal information in DQ tools. First, we developed a hybrid type of misspelling correction technique for personal names and calculated a corresponding confidence level for the corrected error-prone information. Then we have mathematically modeled the estimation of the risk in DQ products by creating the framework using workflow structure. At the end, we introduced a marketing business model where we select the best way to contact the customer (emailing or calling) based on his/her gender. We evaluated the proposed risk estimation technique based on this business model by using the confidence levels achieved from the corrected data.

II. Data Quality Tools And Workflow

Organizations today must digest large amounts of data to drive their decision making processes. While accurate, timely, and complete data cannot guarantee correct decisions, it is clear that decisions based on poor quality data can be costly. Removing the ill-defined data from a system is generally called data cleansing. In many organizations, a data administrator (DA) is the person responsible for cleansing data which often requires complex processes using many different programs and tools. This can lead to the following problems [3].

- They can be error-prone;
- Different selections may be provided for the same job by different DAs;
- A DA may not know to reuse past solutions developed by other DAs; and

doi:10.4304/jcp.6.2.155-161

• The process is labor-intensive. It can take a significant amount of time to produce results.

These and other problems are driving the development of DQ tools (a comprehensive list of the tools can be found at [4, 5]) that are designed to support and simplify the data cleansing process. All of these DQ products are a variation of workflows, which consist of a number of function blocks and input and output fields. At the end, a workflow based business model (i.e. data cleansing operation) is produced by mapping output fields of one function block to another's input fields [3].

The most important features of this DQ business model are:

- Input Dependency: Systems are dependent on customer input
- Resource Flexibility: The resources are flexible (i.e. type of documents)
- Conflict of Interest: The conflict of interest between the organization and its clients supplying basic information or receiving services, and
- Alternatives: Multiple routings and outcomes for different inputs supplied by the customer.

Being able to characterize these DQ processes based on QoS has certain advantages [6]. (A) Monitoring the system from a QoS perspective: While fulfilling costumer expectations by a DQ tool, workflows and adjacent tools must be constantly monitored throughout their life cycles to assure both initial QoS requirements and targeted objectives. When undesired metrics are identified or threshold values are reached, QoS monitoring allows for adaptations of new strategies or aborting the process in order to check the workflow design. (B) Design the system based on QoS: Because the business task can be designed according to QoS metrics, it allows organizations to translate their vision into their business process more efficiently. (C) Selection and Execution based on QoS: It allows for the selection and execution of workflows based on their QoS, to better fulfill customer expectations.

Although QoS has been widely discussed in the areas of real time applications [7] and networking [8, 9], no research has been conducted into DQ products and only a few research teams have made serious attempts to explore the field of workflow or business process automation. So far, most of the research carried out to extend the functionality of workflow systems' QoS has only concentrated on process/service scheduling in order to minimize total execution time. Eder et al. offer heuristics for computing process deadlines and meeting global time constraints [10]. Zeng et al.'s workflow middleware [11], Pegasus [12, 13], Amadeus [14], Askalon [15], and more recently, stochastic modeling approaches [16, 17] exploit grid service technologies, and provide intelligent workflow scheduling techniques in order to meet time constraints. Our system differs from the above since we mainly focus on an accuracy-oriented task by allowing the user to estimate the risk propagation probabilities in their models. Moreover, this system offers the ability to adjust workflow accuracies in such a way that the modified workflow optimizes the QoS constraints.

Estimating this risk propagation probability in a DQ product (Figure 1) is based on the correction technique that is applied to the input file. This correction technique is a pre-process step before running the workflow. To be more specific, assuming that a client wants to run a business process based on the input file he/she possesses. The original input file has to be scanned and corrected for any mistakes or misspelled words that are present in the document. In this process, some of the problems will be corrected while others will not. Since, this pre-process has direct impact on the error distribution in the DQ tools, in the next section; we are going to briefly introduce the Personal Name Recognition Strategy [1] that we created in order to fix the problems in the input data.

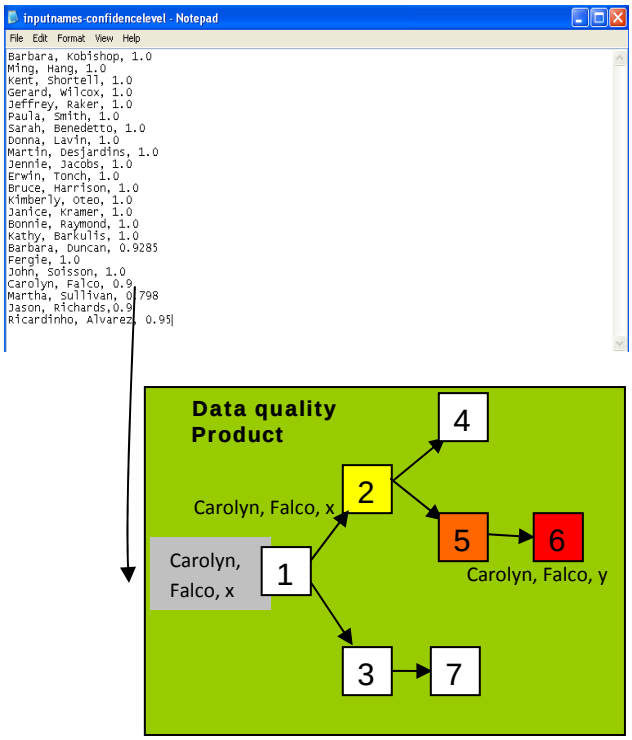

Figure 1. Architectural View of Risk Assessment

III. Personal Name Recognition Strategy

Personal Name Recognition Strategy (PNRS) [1] is based on the results of a number of strategies that are combined together in order to provide the closest match (Figure 2). The near miss strategy and phonetic strategy are used to provide suggestions at the identical time and weight. Since the input data heavily involves names and possible international scope; it is difficult to standardize the phonetic equivalents of certain letters. Once we have a list of suggestions, an edit distance algorithm is used to rank the results in the pool. In order to provide meaningful suggestions, the threshold value, t is defined as 2. In cases where the first and last characters of a word

did not match, we modified our approach to include an extra edit distance. The main idea behind this approach is that people generally get the first character and last character correct when trying to spell a word. Another decision mechanism is needed to suggest the best possible solution for individual names. The rationale behind this idea is that at the final stage it is possible to come up with several potential candidate names for a mistyped word that is one or two edit distances away from the original word. Relying on edit distance often does not provide the desired result. The U.S. Census Bureau study, which has compiled a list of popular first and last names scored based on the frequency of those names within the United States [18], is added as the decision making mechanism. This allows the tool to choose a "best fit" suggestion to eliminate the need for user interaction. Since this part is not the main core of this work, the details of the algorithms are not going to be discussed in this paper. However, extensive information about the techniques can be found in [1].

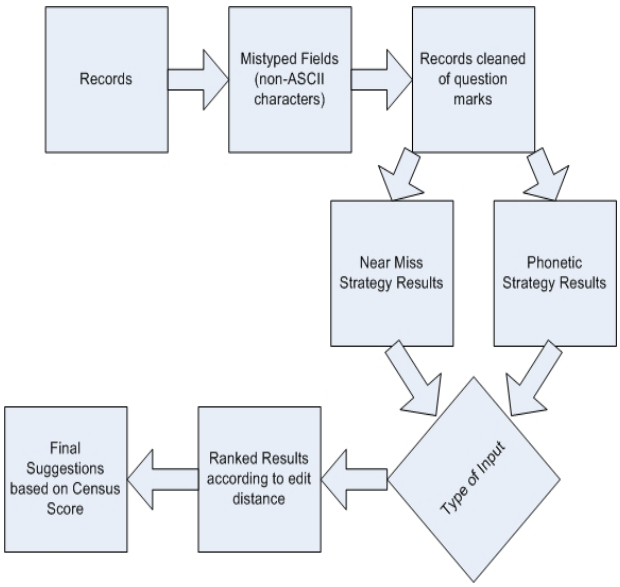

Figure 2. PNRS Strategy

IV. CONFIDENCE LEVEL

The experimental data we used for this risk propagation estimation study is a real-life sample which involves names, addresses (including zip code, city and state) and phone numbers of the individuals. Dirty data is present in a total of 2,812 records, including misspelled names and some non-ASCII characters. PNRS correctly fixed 1,826 records, 163 records were identified as valid names, while 823 records were corrected but produced different names as reflected in Figure 3 and Table I.

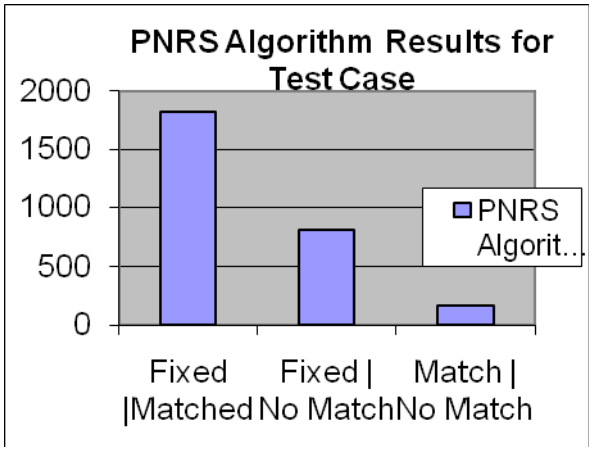

Figure 3. PNRS Correction Rate for Test Case

- Fixed | Matched → Exact correction of the misspelled Name
- Fixed | No Match → Corrections that provide no match with the original name
- Match | No Match → Either the input is accepted as a valid name or the system failed to provide any suggestions.

TABLE I.
DEFINITION OF PNRS RESULTS

Misspelled Name	Original Name	PNRS Suggestion	State of Result
Siteffaannny	Stephanie	Siteffaannny	Match | No Match
Luiz	Luis	Luiz	Match | No Match
Rca	Rice	Ryan	Fixed | No Match
Jaso	Jason	Jason	Fixed | Matched

In order to evaluate the effectiveness of the tool, experiments are conducted not only on the current correction algorithm PNRS, but also on the well known general text spelling correction tools, such as Soundex, Phonex, Phonix, DMetaphone, Levenshtein Edit Distance (ED), LCS, Jaro-Winkler, and n-grams. PNRS averaged 65% correction rate in the test. Since there is no concrete decision mechanism if there is more than one same score with the Soundex, Phonex, Phonix, DMetaphone, Levenshtein Edit Distance, LCS, Jaro-Winkler, and n-grams algorithms, it is a challenge to claim which algorithm performs the best. However, if we look at the minimum and maximum likelihood of full correction rates (Table II) among all these algorithms and PNRS, we would arguably claim the correction rate of PNRS is the best while Jaro-Winkler provided the second best results.

The Jaro-Winkler algorithm calculates the similarity between two words based on the character difference within their original form [2]. The Jaro distance metric (1) states that given two strings s1 and s2, their distance d_j is:

$$d_j = \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right) \quad (1)$$

where, m is the number of matching characters , and t is the number of transpositions.

TABLE II.
MINIMUM AND MAXIMUM CORRECTION RATES OF THE ALGORITHMS

Strategy	Minimum and Maximum Fixed | Matched Percentage for Test Case
Soundex	49.9%-52.4%
Phonex	48.3%-51.1%
Phonix	48.5%-49.1%
DMetaphone	50.9%-51.6%
Edit Distance	42.2%-55.1%
LCS	57.7%-63.3%
Jaro - Winkler	62.1%-63.6%
2-grams	57.2%-59.5%
3-grams	52.2%-58.3%
PNRS	64.9%

The Winkler [2] algorithm (2) improves the Jaro algorithm by applying ideas based on empirical studies which found that fewer errors typically occur at the beginning of names. The Winkler algorithm therefore increases Jaro's similarity measure for agreeing on initial characters (up to four). Given two strings, their Jaro-Winkler distance d_w is:

$$d_w = d_j + (kp(1-d_j)) \quad (2)$$

where:

- d_j is the Jaro distance for the strings
- k is the length of common prefix at the start of the string up to a maximum of four characters
- p is a constant scaling factor for how much the score is adjusted upwards for having common prefixes. The standard value for this constant in Winkler's work is $p = 0.1$

At the end, Jaro-Winkler is a normalized similarity measure between 1.0 (strings are the same) and 0.0 (strings are totally different).

No matter whether we use PNRS or Jaro-Winkler, the attempted correction may totally fix the ill-defined data. However, in most cases, there is a good possibility of having Fixed | No Match type of results at the end which will yield the propagation of error in the DQ tools. Therefore, based on each correction made to the input records, a confidence level should be attached to it. The experiment results demonstrated that PNRS and Jaro-Winkler have their own strengths. For instance, some of the misspelled names were corrected by the PNRS technique while Jaro-Winkler was not able to do so. The opposite is also true. Therefore, our system uses a confidence level coefficient achieved by the average score of both of these algorithms when calculating the risk propagation in the DQ product.

V. ASSESSING THE MISSPELLING OF PERSONAL NAMES

Before discussing the definition and details of the risk propagation, we need to define basic control properties of the workflow that is being used in DQ tools and the associated reliability parameter in order to assess the risk through the workflow. In other words, we will define three basic control properties of workflow and each corresponding reliability parameter will guide us when calculating the risk propagation in the DQ tools [19].

A. Basic Control Properties of the Workflow

Sequence: An activity in a workflow process is enabled after the completion of another activity in the same process. The typical implementation involves linking two activities with an unconditional control flow arrow (Figure 4). However the sequence can be reduced to a single state and the reliability will be calculated as (3):

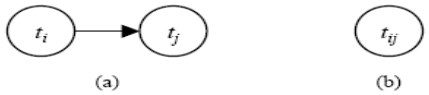

Figure 4. Sequence

$$R(t_{ij}) = R(t_i) * R(t_j) \quad (3)$$

where R refers to the reliability of the reduction.

Parallel Split and Synchronization: A point in the workflow process where a single thread of control splits into multiple threads of control, which can be executed in parallel so allowing the activities to be executed simultaneously or in any order, can be defined as parallel split. The reduction of the system is illustrated in (Figure 5).

The reliability of t_{1n} is calculated as (4):

$$R(t_{1n}) = \prod_{1 \le i \le n} R(t_i) \quad (4)$$

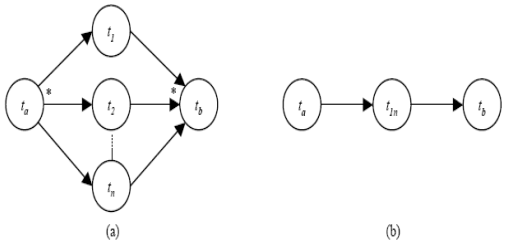

Figure 5. Parallel Split and Synchronization

Conditional System: The ability to converge two or more branches such that the first activation of an incoming branch results in the subsequent activity being triggered and subsequent activations of remaining incoming branches are ignored (Figure 6).

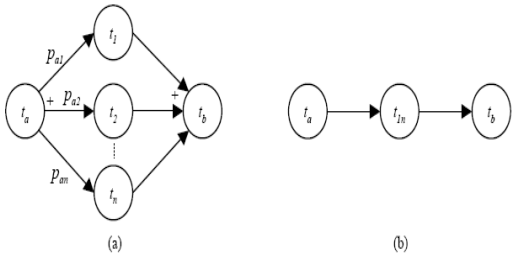

Figure 6. Conditional System

The reliability of t_{1n} is calculated as (5):

$$R(t_{1n}) = \sum_{1 \le i \le n} p_{ai} * R(t_i) \tag{5}$$

B. *Risk Propagation Probabilities*

In the next subsection, we first introduce and then discuss the feature of *risk propagation* in the DQ product using the workflow specifications, where each component (business action) is called a function block.

1) Definition of Risk Propagation in Workflow:

We will consider two function blocks, say D_i and D_j, of an architecture, and let us assume Z is the connector that carries information from D_i to D_j. The specific form of function blocks D_i, D_j and the connector Z is not important for the purposes of our discussion; we can assume they are functions and lines that map an output field of one function block to another's input field by the discussed connector.

The *risk propagation probability* from function block D_i to function block D_j is denoted by RP(D_i, D_j) and defined by (6):

$$RP(\mathrm{D_i}, \mathrm{D_j}) = \mathbf{P}([\mathrm{D_j}](z) \neq [\mathrm{D_j}](z') \mid z \neq z') \tag{6}$$

where $[D_j]$ denotes the purpose of function block D_i, and z is transported data elements of the connector Z from D_i to D_j. We construe $[D_j]$ to capture all the effects of executing function block D_j, including the effect on the function block D_j as well as the effect on any outputs produced by D_j.

Since the outcome of executing D_j will be affected by the error in D_i, we construe RP(D_i, D_j) as the probability that an error in D_i is propagated by D_j. Also, we let RP(D_i, D_i) be equal to 1, which is the probability that an error in D_i causes an error in D_i. Given architecture with N function blocks, we will let RP be an N×N matrix such that the entry at row D_i and column D_j be the risk propagation probability from D_i to D_j.

When defining the estimation of risk propagation in the DQ Tools, the environment can be seen as a collection of components (function blocks) $D_{i,j}$=1,…,N. We used the confidence level coefficients value (obtained from Section IV) $\boldsymbol{\mu}_z^{ij}$ for every interface element z∈Zi and every other function block D_j, j≠i.

- $\boldsymbol{\mu}_z^{ij}$ = [0,1] if the interface element *z* provided by D_i is required by D_j which means that any error in function block D_i associated with interface element z will propagate to function block D_j.
- $\boldsymbol{\mu}_z^{ij}$ = 0, otherwise.

Therefore we used the following (7) for estimating the risk propagation probability RP(D_i, D_j), for every pair of function blocks D_i and D_j, i≠j, based on the confidence level coefficients.

$$RP(\mathrm{D_i}, \mathrm{D_j}) = \frac{1}{|Z_i|} \sum_{z \in Z_i} \mu_z^{ij} \tag{7}$$

Note that the definition of the risk propagation given above uses the concept of *conditional* probability. In other words, we calculate the probability that a risk propagates from D_i to D_j, if D_i actually transmits a message to D_j. However it would be more realistic if we calculated the probability that an error propagates from D_i to D_j not conditioned based on the event that D_i sends a message to D_j. Function $R(D_i, D_j)$ is dependent on $RP(D_i, D_j)$, but it also involves the probability that D_i sends a message to D_j. Therefore, we should consider the *transmission probability matrix* $\boldsymbol{T}$ where the $\boldsymbol{T}(D_i, D_j)$ reflects the probability with which the connector gets activated during D_i sending a message to function block D_j. It is reasonable to assume that $\boldsymbol{T}(D_i, D_i) = 0$ for all function blocks D_i.

The *absolute risk propagation* can be obtained as the product of the conditional risk propagation probability with the probability that the connector between function block D_i and D_j is activated (8),

$$\mathrm{R}(D_i, D_j) = \mathrm{RP}(D_i, D_j) \text{ X } \mathrm{T}(D_i, D_j) \tag{8}$$

However, it is not always possible to compute the transition matrix in a DQ business model. Therefore, we simply omit the transition matrix when calculating the risk probabilities.

2) Multi-Step Risk Propagation in the Workflow:

So far we have focused our attention on single step risk propagation from function block D_i to D_j. However, for the given workflow example (Figure 7), the probability of an error in function block t_a propagates to function block t_b in a number of steps starting in t_a and ending in t_b. This will yield to the *multi-step error rate (MSER) in workflow* from t_a to t_b and defined as (9):

Figure 7. Workflow Sample

$$\mathrm{MSER}(t_{ab}) = \mathrm{R}(t_a) * EP(t_a, t_{1n}) = \frac{1}{|X_i|} \sum_{x \in X_i} \mu_x^{ij} * EP(t_{1n}, t_b) = \frac{1}{|Y_i|} \sum_{y \in Y_i} \mu_y^{ij} \tag{9}$$

where the set X_i and Y_i are the interface elements of the provided functions of t_a *and* t_{1n}.

VI. Test Case and Results

For the proof of concept, we have developed a primitive business model using Talend Open Studio 3.0, which is an open-source ETL, data integration and DQ tool. In the designed business model, since we are willing to limit our sales geographically to a particular area, the first step is to find the home address of the person to whom we would like to send our product's brochure. After collecting the home addresses, based on their sex, we are aiming to contact them via phone or email. According to our market research, contacting females via phone tends to result in more product sales than using email or mail as the communication device, whereas male buyers tend to use email more often and respond more positively. The high level architecture of this particular business model can be seen in Figure 8.

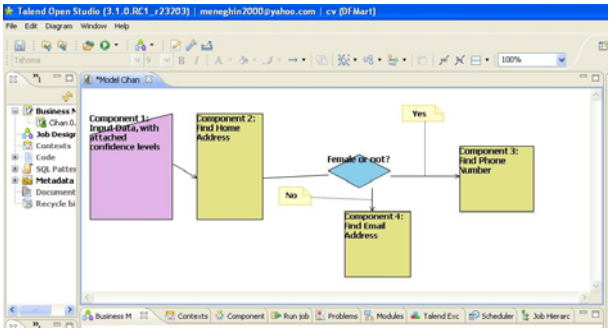

Figure 8. High Level Architecture View of the Business Model

In detail, the Component 1 holds the input data. This input data contains personal names and a confidence level attached next to each record, where 1.0 represents full confidence, others represent the confidence level on the proposed name (a portion of it is shown in Figure 9). Component 2 is the process where the system identifies home addresses. Component 3 and Component 4 are the processes to find phone numbers and email addresses respectively. As this business model reflects, both Component 3 and Component 4 are dependent on Component 2. In the mean time, both of them are also dependent on component 1 as well.

After designing the business model as workflow diagrams, the model is exported as an XML file (Figure 10). Then we estimated the risk propagation probabilities based on the confidence levels and gathered the risk propagation metrics as shown in Table III. For instance, the estimated risk propagation from Component 2 to Component 3 is calculated as the total confidence level of the input file divided by the total number of records carried on that particular connector, which is 0.969. In a multi step, such as the risk propagation from Component 1 to Component 3 is estimated by the reliability of the input file, and risk propagation probabilities from Components 1 to 2 and 2 to 3, which yields a score of 0.947. The numbers close to one indicate higher confidence levels for the task.

As we have experimented with other business models from local industry, we have realized that higher dependency yields more error to the final product. Therefore embedding this estimation model to predict the confidence level of the outcome before processing the job is crucial for companies.

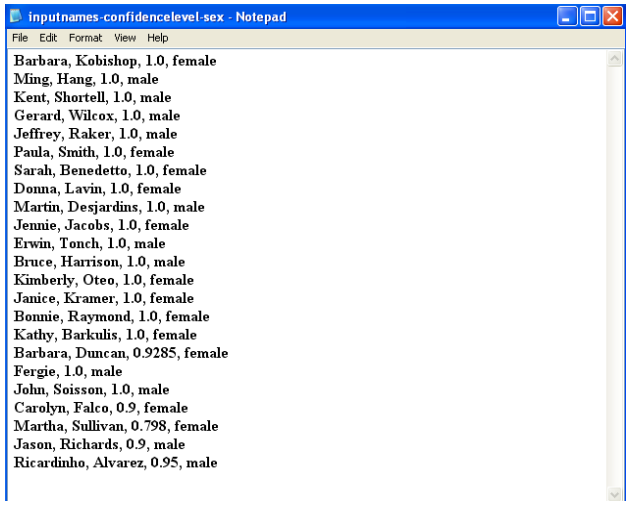

Figure 9. Input File Used in Business Model

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
  - <!-- ArgoUML -->
  - <!-- Created by ArgoUML 0.28 at 10:48:23 PM on
      Wednesday, July 18, 2009 -->
  - <!-- Export Option = 3 -->
- <model id="{Model}" name="Business Model 7 ">
 <filedirectory>C:\Documents and Settings\ Cvarol\
      Desktop\Business Model \Talend\</filedirectory>
 <filename>Business Model 7</filename>
 <filespec>C:\Documents and Settings\ Cvarol\
      Desktop\ Business Model \Talend\Business Model
      7 </filespec>
+ <objects count="4">
+ <data count="23">
+ <associations count="4">
+ <diagrams count="1">
   </model>
```

Figure 10. XML Structure of the Business Model

Table III.
Estimated Risk Propagation Probabilities

Components	1	2	3	4
1	1	0.977	0.947	0.963
2	0	1	0.969	0.986
3	0	0	1	0
4	0	0	0	1

VII. Conclusion

Monitoring and re-designing the system before running all the processes is crucial for business purposes. With this kind of feature, companies can predict the possible outcome of the business process, also monitor or re-design their cleansing structure based on the statistical results. Therefore, we have designed a model in order to estimate the risk propagation through a DQ tool.

Obviously, the cleaner the data we have the better the results we will get from the business process. Therefore,

our system relies on the best performing techniques to correct the ill-defined data.

Since it is the first study in its area, some issues still need to be addressed in the near future, such as applying the techniques not only in terms of personal names, but also addresses and other preferences as well. Moreover, another challenge would be to integrate this framework into all DQ products.

REFERENCES

[1] C. Varol, C. Bayrak, R. Wagner, and D. Goff. "Application of Near Miss Strategy and Edit Distance to Handle Dirty Data". Data Engineering: Mining, Information and Intelligence, pp 91-101, Springer, ISBN: 978-1-4419-0175-0

[2] C. Varol and C. Bayrak. "Personal Name-based Pattern and Phonetic Matching Techniques: A Survey". ALAR Conference on Applied Research in Information Technology, February 13, 2009, Conway, Arkansas, USA

[3] C. Varol and C. Bayrak. "Measuring Reliability Component for Quality of Service (QoS) in Business Process Automation", The 14th International Conference on Distributed Multimedia Systems, September 4-6 2008, Boston, Massachusetts, USA.

[4] V. Goasdoué, A. Nugier, D. Duquennoy, and B. Laboisse. "An evaluation framework for data quality tools". Proceedings of the International Conference for Information Quality (ICIQ'07).

[5] J. Barateirio, H. Galhardas. "A survey of data quality tools", Databank-Spectrum, 2005

[6] J.A. Miller, M. Fan, S. Wu, I.B. Arpinar, A.P. Sheth, and K.J. Kochut. "Security for the Meteor workflow management system". UGA-CS-LDIS Technical Report, University of Georgia. 1999

[7] D. Clark, S. Shenker and L. Zhang. "Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism". Proceedings of ACM SIGCOMM. pp. 14-26. 1992

[8] R.L Cruz. "Quality of service guarantees in virtual circuit switched networks". IEEE Journal of Selected Areas Communication, 13(6): 1048-1056. 1995

[9] L. Georgiadis, R. Guerin, V. Peris, and K. Sivarajan. "Efficient Network QoS Provisioning Based on Per Node Traffic Shaping". IEEE ACM Transactions on Networking 4(4): 482-501. 1996

[10] J. Eder, E. Panagos, and M. Rabinovich. "Time constraints in workflow systems". Lecture Notes in Computer Science, 1626:286, 1999

[11] L. Zeng, B. Benatallah, A. H. Ngu, M. Dumas, J. Kalagnanam, and H. Chang. "Qos-aware middleware for web services composition". IEEE Transactions on Software Engineering, 30(5):311–327, 2004.

[12] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. C. Laity, J. C. Jacob, and D. S. Katz. "Pegasus: A framework for mapping complex scientific workflows onto distributed systems". Scientific Programming, 13(3):219–237, 2005.

[13] Y. Gil, V. Ratnakar, E. Deelman, G. Mehta, and J. Kim. "Wings for pegasus: Creating large-scale scientific applications using semantic representations of computational workflows". In Proceedings of the 19th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI), Vancouver, British Columbia, Canada, July 22- 26, 2007.

[14] I. Brandic, S. Benkner, G. Engelbrecht, and R. Schmidt. "Qos support for time-critical grid workflow applications". E-Science, 0:108–115, 2005.

[15] M. Wieczorek, R. Prodan, and T. Fahringer. "Scheduling of scientific workflows in the askalon grid environment". SIGMOD Rec., 34(3):56–62, 2005

[16] A. Afzal, J. Darlington, and A. S. McGough. "Qos-constrained stochastic workflow scheduling in enterprise and scientific grids". In GRID, pages 1–8, 2006.

[17] W. Wiesemann, R. Hochreiter, and D. Kuhn. "A stochastic programming approach for qos-aware service composition". The 8th IEEE International Symposium on Cluster Computing and the Grid (CCGRID' 08), pages 226–233, May 2008.

[18] Census Bureau Home Page. www.census.gov 1990

[19] J. Cardoso, A.P. Sheth, J.A. Miller, J. Arnold, and K. Kochut, "Quality of service for workflows and web service processes" J. Web Sem. 1(3), 281-308 (2004).

Cihan Varol is an Assistant Professor of Computer Science at Sam Houston State University.

He received a BS degree in computer science from Firat University, Elazig, Turkey in 2002 and a masters degree from Lane Department of Computer Science and Electrical Engineering at West Virginia University, Morgantown, WV, USA in 2005 and a Ph.D. degree from the Department of Applied Science at the University of Arkansas-Little Rock, AR, USA in 2009.

His research interests are data quality, entity resolution, workflow systems, software engineering and software metrics. He has published over 20 journal and conference papers including a book chapter in Data Engineering: Mining, Information, and Intelligence published by Springer (2010).

Coskun Bayrak is a Professor in the department of Computer Science at the University of Arkansas at Little Rock.

He received a BS degree from Slippery Rock University, a MS from Texas Tech University, and a Ph.D. from Southern Methodist University in Computer Science.

His primary research is in the intersection of software engineering component based development, data mining, and Biomedical Engineering. However, he also has interest in modeling and simulation, cellular automata, and monitoring and control. Dr. Bayrak has published over 50 research articles in scientific conferences and journals, two book chapters, given tutorials at major conferences, and served on program committees for numerous international conferences and symposiums.