A Technique for the Quantitative Measure of Data Cleanliness

Mr.Abhijit Wakchaure
School of Electrical Engineering and Computer Science
University of Central Florida,
Orlando, USA
abhijitw@mail.ucf.edu

Dr.Ronald Eaglin
Dept. of Engineering Technology
University of Central Florida,
Orlando, USA
reaglin@mail.ucf.edu

Dr.Bahman Motlagh
Dept. of Engineering Technology
University of Central Florida,
Orlando, USA
bmotlagh@mail.ucf.edu

***Abstract*—With the amount of data that is collected, viewed, processed, and stored today, techniques for the analysis of the accuracy of data are extremely important. Since we cannot improve what we cannot measure, the need for a tangible quantitative measure of data quality is a necessity. This paper focuses on a data-cleanliness algorithm, which makes use of the 'Levenshtein distance', to measure the data quality for a criminal records database. Actual law enforcement name records were used for this research. The results help us arrive at the extent of dirtiness in the data, and also highlight the different types of dirty data. We then go on to show how measuring the data quality not only helps in setting up guidelines for the data clean-up process, but also can be used as a metric for cross-comparing like databases.**

***Keywords*—data quality, dirty data, data cleanliness**

I. Introduction

The volume of data in today's world is growing rapidly. As organizations around the globe realize the increased importance of their data as being a valuable asset in giving them the competitive edge in today's fast-paced business world, more attention is being paid to the quality of this data.

The data explosion phenomenon has created an immense opportunity and the need for methodologies of Knowledge Discovery and Data Mining (KDD) [23]. Data mining is a rapidly growing field and emerging as one of the top key technology areas in information and knowledge management. Organizations are building data repositories and data warehouses so as to effectively mine data and extract meaningful context-based information out of available data. Recent reports by IDC [22] forecast the data warehouse market to grow to $13.5 billion in 2009 at a nine percent compound annual growth rate.

The issue of data quality can never be stressed as more important than in today's world, where data is everything, everywhere. Advances in text mining, web mining, predictive analytics, etc. all depend on accurate data. That one cannot effectively mine data, which is dirty or inaccurate, comes as no surprise. In its 2002 report on data quality, TDWI estimated "that poor quality customer data costs U.S. businesses a staggering $611 billion a year in postage, printing and staff overhead." According to a very recent follow-up study, "Taking Data quality to the Enterprise through Data Governance," by the TDWI conducted in Apr 2006, 53% of the respondents surveyed answered "Yes" to the question: "Has your company suffered losses, problems, or costs due to poor quality data?" [11].

Applications in fields such as health care and medical research [1] are critically data-sensitive. Dogu Celebi, M.D., vice president of clinical affairs and client services at Waltham, Mass.-based IHCIS, when asked about the challenges to data mining says, "Until recently, data quality and supporting technology have been the biggest barriers" [24].

Business intelligence applications also need the data to be of a very high quality, since analysis based on incorrect or inaccurate data leads to losses, such as customer dissatisfaction, increased operational cost, less-effective decision making, reduced ability to make and execute strategy, low employee morale, to name a few [8]. In a January 2006 InformationWeek research survey of business technology professionals and their plans to expand the deployment of business intelligence (BI) tools within their organizations, 51% cited integration issues with existing systems and 45% cited data quality issues, when asked as to why more employees are not currently using BI tools [11]. According to Ted Friedman, research vice president at Stamford, Conn.-based Gartner Inc., "Data quality is a major inhibitor of BI projects, which can cause user distrust and

978-1-4244-1674-5/08 /$25.00 ©2008 IEEE

abandonment of the system. Flawed data can also have dire effects on a business. Bad data truly does breed bad decisions." [9]

A. Definitions of data quality

Data quality has been defined in numerous ways, some of which are 'fitness for use', 'inaccurate, inconsistent, redundant data', data consisting of 'spurious links', to name a few. Most of these definitions encompass various aspects of data quality and broadly address the sources of dirty data. However, Won Kim et al. presents us with a detailed taxonomy of dirty data [2], wherein 33 different types of dirty data have been specified, with a view to use them as metrics for data quality. Dirty data manifests itself in various forms - inaccurate, redundant, duplicate spurious links, incomplete, among others.

According to Claudia Imhoff, PhD, President and Founder of Intelligent Solutions Inc., poor-quality data is data that is inaccurate, incomplete, misleading and one that leads to bad decision-making [10]. As per Informatica, data quality encompasses more than finding and fixing missing or inaccurate data; it means delivering comprehensive, consistent, relevant and timely data to the business regardless of its application, use or origin [12]. Data quality has been considered as a multi-dimensional concept. [13, 14]

II. PREVIOUS WORK

Tamraparni Dasu et al. [17] propose a data quality approach wherein “business rules are implemented as constraints on data in a classical expert system formalism sometimes called production rules”. Their methodology focuses on data quality problems arising due to the lack of accurate and complete documentation of business rules. There also exist practical complications such as frequently changing business rules, and fragmented domain knowledge among various experts, whose opinions do not always converge. According to the authors, it is not uncommon for business databases to have 60% to 90% bad data, which not only forces frequent data audits to maintain database integrity, but greatly affects the company’s performance. Business rules are considered as dynamic constraints on the database, which relate to data flows as per the associated business operations. Thus, their goal is to accurately represent, maintain and update these constraints in order to ensure data usability and reliability, which are two major components of data quality metrics.

Ian Davidson et al. propose the use of data quality matrices [18] in data mining algorithms. The authors mention that routine errors such as non-existent zip codes in an address database, can be detected and corrected by traditional data-cleansing tools, but want to draw attention to undetectable but documentable errors such as say, a particular zip code being mistakenly interchanged with another in the same state. Elizabeth Pierce, an associate professor at Indiana University of Pennsylvania highlights the use of Control matrices [19] as a complementary approach to handle such errors and to link data problems to the quality controls that should help detect and correct these data problems. According to her, the elements of the matrix rate the effectiveness of the quality check at reducing the level of data errors.

Jochen Hipp et al. propose data quality mining [20], the goal of which is to employ data mining methods in order to detect, quantify, explain and correct data quality deficiencies in very large databases. The authors regard their approach to data quality in the context of knowledge discovery in databases (KDD), as new and quite promising. Realizing that poor data quality is a critical problem, when it comes to practical applications of KDD, their definition of data quality mining is a deliberate application of various data mining techniques for the purpose of data quality measurement and improvement.

Bing Tian Dai et al. [21] have presented an interesting approach – column heterogeneity as a measure of data quality. Their novel approach focuses on column heterogeneity, that seeks to quantify the data quality problems that can arise when merging data from various sources.

Many organizations emphasize the use of vendor data quality tools [5, 15] in order to increase the quality of data in a data warehouse environment. Recently, many vendors have felt the need to come out with data quality tools [16]. “By buying Firstlogic, the BI technology provider Business Objects seeks to expand its data integration offerings by adding data quality capabilities. Because data quality is key to success with BI, this acquisition will enable Business Objects to market a more complete solution. Over the past few years, vendors such as IBM, SAS Institute and Informatica have made strategic acquisitions to enter the data quality tool market.” However, the data cleansing tools on the market do not address all types of dirty data [7]. Also, ETL tools typically have little built-in data cleaning capabilities and there is usually no data analysis support to automatically detect data errors and inconsistencies [5].

III. MOTIVATION

The literature reviewed clearly portrays the strong need for data quality assessment/measurement. William McKnight in a white paper by First Logic [4] says "However there has not been a methodology to articulate and improve data quality ROI, until now. You can't improve what you can't measure. So, we need a means for measuring the quality of our data warehouse". According to Business Objects, the first data quality process, as a part of a Successful Data Quality Solution, is measuring the number and type of defects. Claudia Imhoff, President and

Founder of Intelligent Solutions Inc., remarks [10], "If you can't measure it, you can't improve it." Thus, a fundamental part of improving data quality is to be able to measure data quality. Until you have some type of baseline metric, you don't even know where you are." Won Kim et al. mentions the need for metrics for quantifying data quality so as to measure the quality of data sets. Also, Leo L. Pipino et al. [13] stress the importance of usable data metrics to assess how good a company's data quality is. According to them, "Assessing data quality is an on-going effort that requires awareness of the fundamental principles underlying the development of subjective and objective data quality metrics."

First Logic talks about the importance of Data profiling and scoring in its white paper [4]. "By taking account of important data across several tangible factors that can be used to measure data quality, you can begin to translate the vague feelings of dirtiness into something tangible." Data scoring can be then used as a relative measure of conformance to pre-defined data quality rules. Also, " the data quality rules can be arrived at not only by intellectually determining how the data should look like, but at the cost to the function of the system, if the data lacked quality". Mong Li Lee et al. [6] make use of context information between data records to help solve the data quality problem of spurious links, which is a newly discovered class of erroneous data, in which improperly associated multiple links of a real-world entity exist in the database.

Data quality is an on-going process and not just a one-time initiative, say while integrating multiple-source data into a data warehouse. Continuous quality monitoring and assessment is critical. "For example, according to the United States Postal Service, more than 44 million Americans change their addresses each year. This makes address data that was once valid, now incorrect". In a Webinar by FirstLogic, Ms.Cheri Mallory focuses on adopting an 'Information Quality Maturity Model' to define various levels of data quality in an organization. According to her, there is an emergent role of a 'data steward' to manage the ongoing data quality of an organization, wherein the data steward is accountable and takes full responsibility for the company's data.

Thomas Redman, President of Navesink Consulting Group remarks "The science of data quality has not yet advanced to the point where there are standard measurement methods for any of these issues, and few enterprises routinely measure data quality. But many case studies feature accuracy measures. Measured at the field level, error rates range wildly, with reported error rates of 0.5 – 30% [8]. Thus measuring data quality is of the utmost importance, and once we have a tangible measure of the dirtiness of data, one needs to focus on the 'causes' of dirty data and further focus on prioritizing them to achieve a very high level of data quality over a reasonable period of time. "But if you measure data quality and it is too low – let's say it's 80 percent accurate and you need to get to 90 percent accuracy – then that's where things can get complex. It helps to work through that complexity if IT and business users can collaborate to discover where the bad data is coming from." [11]

In our study, we propose to measure the accuracy of data by defining data quality metrics for some of the data elements, and finally arriving at the database cleanliness score for the entire database. Some of the metrics of data quality in a data quality initiative could be the consistency of data quality assessment figures over time, and also the time required in correcting the inaccuracies in the data after each data quality assessment.

A. Data source

The data used in this study came from a criminal records database called FINDER (Florida Integrated Network for Data Exchange and Retrieval).

IV. PROCEDURE

A. Database schema:

One of the challenges in criminal datasets is to identify whether two suspects having 'similar' names are indeed the same person. Often, when crime-related databases are queried against a particular name, the query results return a large number of similar suspect names and it becomes an overwhelming/tedious task to establish all those similar person names with slight name variations, existing as different individuals, as just one suspect. However, it is not at all hard to imagine that by doing so, i.e. – by compressing the results on the basis of individual suspects gives a much more clearer and realistic picture to the crime analysts and makes their job a lot easier.

Since we are working on a criminal dataset, we set our goal to develop an algorithm to identify the extent of dirtiness pertaining to person names, and measure it in a quantifiable manner.

B. Data-Cleanliness algorithm

Our algorithm makes use of the 'Levenshtein edit distance(LD)' [3], which is a measure of the similarity between two strings. It is the minimum number of character deletions, insertions, substitutions, and transpositions required to transform one string into the other.

$$\boxed{LD = f(search\_string, reference\_string, n)}$$

where 'n' is the limit parameter provided, so the function will quickly reject strings containing more than 'k' mismatches.

We have four relevant fields in our algorithm, namely Lastname, Firstname, Date of Birth and the Sexcode. We find the Edit distance for all the fields, except the Sexcode, and arrive at a match percentage for each field. Each field is assigned a weight and then all the individual match percentages are added to arrive at a final match percentage:

$$\boxed{\begin{array}{l}\text{Match \% for each record} = (W_{LN} * LD_{LN}) + \\ (W_{FN} * LD_{FN}) + (W_{DOB} * LD_{DOB}) + (W_{SEX} * LD_{SEX})\end{array}}$$

where 'W_{LN}' & 'LD_{LN}' are the weight and Levenshtein distance for the Lastname, Firstname, DOB and Sex.

The Data-Cleanliness algorithm takes each record from the sample, and calculates the matching records in the entire database, with a matching percentage of 85 (threshold) and higher. Thus all those records resembling the sample record for 'John Smith' say, are essentially records for the same person, since there is a very 'close' match, and we know that these are the slight name or DOB variations for that person.

We thus find out the Lastname and Firstname dirtiness associated with that person, or for each record and finally average it for the entire database.

C. *Basis for 85% threshold*

As we know, a higher match percentage would miss out on the matches, which are actually the same person, but are treated as a different person, thereby increasing the number of false negatives. On the other hand a liberal /less restrictive match percentage, would allow in quite a number of false positives. (Depending upon the commonality of the name.)

In our analysis of the matches, we decided to set 85% as our threshold for establishing the matches above that threshold to be the same person. Some sample test names were used and the sensitivity of the threshold was checked on these names to set our threshold as 85. Thus, we are quite confident that the matched records are of the same person; at the same time we are aware that there is a small chance that some of the matched records do belong to a different person.

D. *How run variations were achieved – significance of weights – justification*

As mentioned in the 'Data-Cleanliness Algorithm' section, each of the four relevant fields has been assigned weights, the combined weight being 1. We try and manipulate the assigned weight for each field in order to see the overall change in the match results. For most of the run variations, the sexcode has been assigned the least weight (0.1), since it is a Boolean and we do not want to penalize the entire record for an easily possibly error in wrongly entering the sex of the person as 'M' instead of 'F', or vice-versa. In one instance, we have eliminated the effect of sex on the match percentage. Here are the run variations, which were carried out, using the following weights assigned weights:

Assigned Weights				Comments
LN	FN	DOB	SEX	
0.3	0.2	0.4	0.1	Higher weight to DOB over LN
0.4	0.2	0.3	0.1	Higher weight to LN over DOB
0.4	0.2	0.4	0.0	Higher weights to both LN and FN
0.3	0.3	0.3	0.1	Equal weights to LN, FN, DOB except SEX

V. RESULTS

The Edit distance or Levenshtein algorithm takes each letter of the passed string and compares it with the existing string, to find out the number of letters which need to be replaced so that the two strings match. In our algorithm, Edit distance is used on the Last name, First name, as well as the Date of Birth. So Edit distance is called three times per record. Also, as mentioned in the Procedure section, the data cleanliness algorithm goes through the sample table, a record at a time, and compares that record with each and every record in the population, to arrive at the match percentage. Again, this is done for all the records in the sample. Hence we decided to use a subset of the entire data as our population, and further took a sample from that subset as our records, for which the dirtiness was to be calculated. Thus our subset was around 50,000 records out of the table size of 1 million records, and our sample size was 5000 a couple of times and 1000 most of the time.

Here is an example, to illustrate how matches are arrived at:

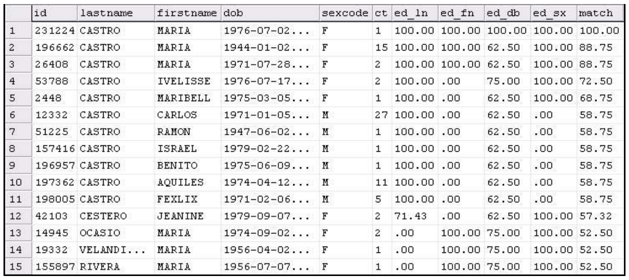
Person details:

LN: CASTRO, FN: MARIA, DOB: 1976-07-02, SEX: F

Weights:

LN (0.3 or 30%), FN (0.2 or 20%), DOB (0.4 or 40%), SEX (0.1 or 10%)

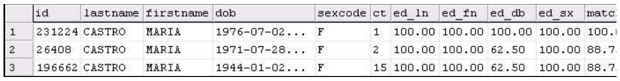
The following figure shows an intermediate table generated by the procedure:



	id	lastname	firstname	dob	sexcode	ct	ed_ln	ed_fn	ed_db	ed_sx	match
1	231224	CASTRO	MARIA	1976-07-02...	F	1	100.00	100.00	100.00	100.00	100.00
2	196662	CASTRO	MARIA	1944-01-02...	F	15	100.00	100.00	62.50	100.00	88.75
3	26408	CASTRO	MARIA	1971-07-28...	F	2	100.00	100.00	62.50	100.00	88.75
4	53788	CASTRO	IVELISSE	1976-07-17...	F	2	100.00	.00	75.00	100.00	72.50
5	2448	CASTRO	MARIBELL	1975-03-05...	F	1	100.00	.00	62.50	100.00	68.75
6	12332	CASTRO	CARLOS	1971-01-05...	M	27	100.00	.00	62.50	.00	58.75
7	51225	CASTRO	RAMON	1947-06-02...	M	1	100.00	.00	62.50	.00	58.75
8	157416	CASTRO	ISRAEL	1979-02-22...	M	1	100.00	.00	62.50	.00	58.75
9	196957	CASTRO	BENITO	1975-06-09...	M	1	100.00	.00	62.50	.00	58.75
10	197362	CASTRO	AQUILES	1974-04-12...	M	11	100.00	.00	62.50	.00	58.75
11	198005	CASTRO	FEXLIX	1971-02-06...	M	5	100.00	.00	62.50	.00	58.75
12	42103	CESTERO	JEANINE	1979-09-07...	F	2	71.43	.00	62.50	100.00	57.32
13	14945	OCASIO	MARIA	1974-09-02...	F	2	.00	100.00	75.00	100.00	52.50
14	19332	VELANDI...	MARIA	1956-04-02...	F	1	.00	100.00	75.00	100.00	52.50
15	155897	RIVERA	MARIA	1956-07-07...	F	1	.00	100.00	75.00	100.00	52.50

Figure 1. Example of an intermediate table generated by the procedure

As can in Figure 1, (first 15 rows displayed) arrives at a match percentage for all the records in the database, and then selects only those matching records with a match percentage greater than 85%, which is the threshold set.



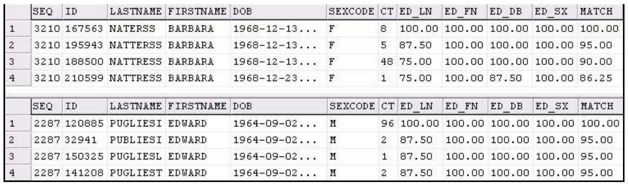
	id	lastname	firstname	dob	sexcode	ct	ed_ln	ed_fn	ed_db	ed_sx	matc
1	231224	CASTRO	MARIA	1976-07-02...	F	1	100.00	100.00	100.00	100.00	100.
2	26408	CASTRO	MARIA	1971-07-28...	F	2	100.00	100.00	62.50	100.00	88.7
3	196662	CASTRO	MARIA	1944-01-02...	F	15	100.00	100.00	62.50	100.00	88.7

Figure 2. Intermediate table depicting variation in date of birth

Thus, for the above person we find three matching records, one of which is the original record, and the other two are matches. We are not concerned as to which of the three records is the right one. We now arrive at the 'LN_dirty' (lastname dirtiness) and the 'FN_dirty' (firstname dirtiness) for this sample, which are both zero, since there is no variation at all in either of the three instances.

We now arrive at the 'LN_dirty' (lastname dirtiness) and the 'FN_dirty' (firstname dirtiness) for this sample, which are both zero, since there is no variation at all in either of the three instances. We further arrive at the 'LN_dirty' and 'FN_dirty' for the rest of the other records in the sample and average them all to find out the overall dirtiness of the database, considering the lastname field, which is the field-level uniqueness.

Some other examples are shown in Figure 3:



	SEQ	ID	LASTNAME	FIRSTNAME	DOB	SEXCODE	CT	ED_LN	ED_FN	ED_DB	ED_SX	MATCH
1	3210	167563	NATERSS	BARBARA	1968-12-13...	F	8	100.00	100.00	100.00	100.00	100.00
2	3210	195943	NATTERSS	BARBARA	1968-12-13...	F	5	87.50	100.00	100.00	100.00	95.00
3	3210	188500	NATTRESS	BARBARA	1968-12-13...	F	48	75.00	100.00	100.00	100.00	90.00
4	3210	210599	NATTRESS	BARBARA	1968-12-23...	F	1	75.00	100.00	87.50	100.00	86.25

	SEQ	ID	LASTNAME	FIRSTNAME	DOB	SEXCODE	CT	ED_LN	ED_FN	ED_DB	ED_SX	MATCH
1	2287	120885	PUGLIESI	EDWARD	1964-09-02...	M	96	100.00	100.00	100.00	100.00	100.00
2	2287	32941	PUBLIESI	EDWARD	1964-09-02...	M	2	87.50	100.00	100.00	100.00	95.00
3	2287	150325	PUGLIESL	EDWARD	1964-09-02...	M	1	87.50	100.00	100.00	100.00	95.00
4	2287	141208	PUGLIEST	EDWARD	1964-09-02...	M	2	87.50	100.00	100.00	100.00	95.00

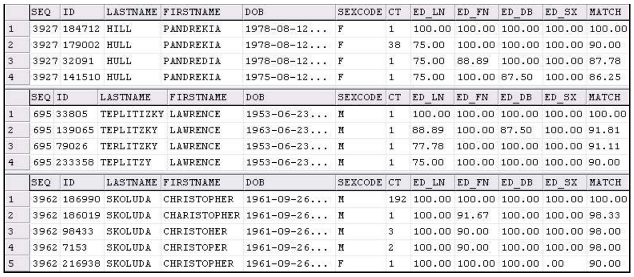
Figure 3. Example of name variation in last name

In the topmost example, we have two distinct last names, namely 'NATERSS' and 'NATTERSS'. Now, one of these last names is correct, whereas the other is incorrect, though, as earlier mentioned, we are not concerned as to which is the right one. Thus, out of four matched records, we have 1 variation of the correct lastname, and thus the LN_dirty = (1/4) = 0.25.

In the second example, we have 4 different last names, for the 4 matched records, and since one of them is the correct one, the other three are incorrect, and thus the LN_dirty = (3/4) = 0.75.

One can use the count ('ct') variable as a possible indicator of the correct lastname in certain instances.

Also, in both the above examples, the FN_dirty=0, since they have the exact same firstname in all the 4 matched records.



	SEQ	ID	LASTNAME	FIRSTNAME	DOB	SEXCODE	CT	ED_LN	ED_FN	ED_DB	ED_SX	MATCH
1	3927	184712	HILL	PANDREKIA	1978-08-12...	F	1	100.00	100.00	100.00	100.00	100.00
2	3927	179002	HULL	PANDREKIA	1978-08-12...	F	38	75.00	100.00	100.00	100.00	90.00
3	3927	32091	HULL	PANDREDIA	1978-08-12...	F	1	75.00	88.89	100.00	100.00	87.78
4	3927	141510	HULL	PANDREKIA	1975-08-12...	F	1	75.00	100.00	87.50	100.00	86.25

	SEQ	ID	LASTNAME	FIRSTNAME	DOB	SEXCODE	CT	ED_LN	ED_FN	ED_DB	ED_SX	MATCH
1	695	33805	TEPLITIZKY	LAWRENCE	1953-06-23...	M	1	100.00	100.00	100.00	100.00	100.00
2	695	139065	TEPLITZKY	LAWRENCE	1963-06-23...	M	1	88.89	100.00	87.50	100.00	91.81
3	695	79026	TERLITZKY	LAWRENCE	1953-06-23...	M	1	77.78	100.00	100.00	100.00	91.11
4	695	233358	TEPLITZY	LAWRENCE	1953-06-23...	M	1	75.00	100.00	100.00	100.00	90.00

	SEQ	ID	LASTNAME	FIRSTNAME	DOB	SEXCODE	CT	ED_LN	ED_FN	ED_DB	ED_SX	MATCH
1	3962	186990	SKOLUDA	CHRISTOPHER	1961-09-26...	M	192	100.00	100.00	100.00	100.00	100.00
2	3962	186019	SKOLUDA	CHARISTOPHER	1961-09-26...	M	1	100.00	91.67	100.00	100.00	98.33
3	3962	98433	SKOLUDA	CHRISTOHER	1961-09-26...	M	3	100.00	90.00	100.00	100.00	98.00
4	3962	7153	SKOLUDA	CHRISTOPER	1961-09-26...	M	2	100.00	90.00	100.00	100.00	98.00
5	3962	216938	SKOLUDA	CHRISTOPHER	1961-09-26...	F	1	100.00	100.00	100.00	.00	90.00

Figure 4. Example of multiple variation

We can see that the 'LN_clean' and 'FN_clean' percentages are quite consistent for the Pawn database, across all the sample runs. The average 'LN_clean' and 'FN_clean' percentages are 99.452 and 99.490 respectively. Though, the percentage of 'total dirty records' for the Pawn dataset is 12.85, whereas that for the Burglary dataset is just 9%, it is interesting to note that the Burglary dataset is dirtier than the Pawn dataset, when it comes to the last name and first name variations for each person in the sample, as can be seen by the 'LN_clean' and 'FN_clean' percentages, which average 97.118 and 97.527 respectively.

The above is an example of measuring the data quality, when it comes to 'record-level uniqueness', since we go through each record, which represents an individual, and then identify the last name and first name variations for that individual, thus arriving at the cleanliness for each record and further averaging it for the database.

We can also measure the data quality field-wise, thus arriving at the field-level uniqueness. Thus, for the last name, we could arrive at the field-level uniqueness percentage for the last name field, simply by calculating the number of last names which were dirty, divided by the total number of last names in our database.

Following is the tabular summary of the sample runs on the Person table in the DSC database, as also the Burglary dataset:

TABLE I. SUMMARY OF RUNS

Run id	Sample size	Last name cleanliness	First name cleanliness	No of dirty records
Pawn:				
A1	1000	99.55	99.57	69
B1	1000	99.45	99.57	84
C1	1000	99.43	99.54	76
D1	1000	99.70	99.59	67
E1	4500	99.47	99.33	364
F1	1000	99.37	99.60	66
G1	1000	99.25	99.33	91
H1	5000	99.35	99.49	395
I1	5000	99.50	99.39	384
Burglary:				
X1	4607	97.093	97.367	519
Y1	4611	97.143	97.687	499

VI. CONCLUSION & FUTURE WORK

The literature review and the documentation studied on data quality clearly states the strong need for having a tangible measure of data quality, since one cannot aim to have a very high data quality without knowing how dirty the data is and the nature of the dirtiness. The proposed Data-Cleanliness algorithm demonstrates how one can arrive at the dirtiness measure of any database, which not only helps in setting up guidelines for the data clean-up process, but also helps in comparing dirtiness measures across different datasets.

The algorithm can be further extended by arriving at cleanliness rules for other fields or tables (for instance, one can compare DOB with AGE, if two such separate fields exist, and check to see if they match, which they should.)

The future work can also include analyses on the dirtiness results, within the database, in terms of a breakdown of the distinct types of dirtiness, which could further help us in understanding the nature of dirty data. Each of these distinct types of dirty data can be associated with a critical dirtiness number, say (based on the severity of the problems it can cause), which would help us in prioritizing the data clean-up process, as also in cross-comparing various datasets not only on the total dirtiness, but also the nature of dirtiness.

REFERENCES

[1] T. Welzer, B. Brumen, I. Golob and M. Družovec, "Medical diagnostic and data quality," Proceedings of the 15 th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002), 1063-7125/02, 2002 IEEE.

[2] W. Kim, B. Choi, E. Hong, S. Kim and D. Lee,"A taxonomy of dirty data," Data Mining and Knowledge Discovery, 7, 81–99, 2003.

[3] "Quick Levenshtein Edit Distance," Website:'http://www.planet-source-code.com/vb/scripts/ShowCode.asp?txtCodeId=584&lngWId=5'.

[4] W. McKnight, "Overall approach to data quality ROI," First Logic white paper.

[5] E. Rahm, and H. Do, "Data cleaning: problems and current approaches," University of Leipzig, Germany.

[6] M. Lee, W. Hsu, and V. Kothari, "Cleaning the spurious links in data," IEEE Intelligent Systems, 1094-7167/04, 2004 IEEE.

[7] W. Kim, "On three major holes in data warehousing today," Published by ETH Zurich, Chair of Software Engineering, Vol. 1, no. 4, September-October 2002.

[8] T. C. Redman, "The impact of poor data quality on the typical enterprise," Communications of the ACM, Vol. 41, No. 2, February 1998.

[9] H. Smalltree, "BI's seven fatal flaws," Website:'http://searchdatamanagement.techtarget.com/originalContent/0,289142,sid91_gci1172024,00.html', 09 Mar 2006.

[10] C. Imhoff, "Poor-quality data… can your company afford the risk?," Teradata white paper, 2006.

[11] M. Schiff, "Data quality first: it's just logical," Business Objects White Paper, 2006.

[12] "Addressing data quality at the enterprise level," Informatica White Paper, October 2005.

[13] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," Communications of the ACM, Vol. 45, No. 4ve, April 2002.

[14] M. Martin, "Measuring and improving data quality, Part II: Measuring data quality," NAHSS Outlook, April 2005.

[15] L. D. Paulson, "Data quality: a rising e-business concern," IT Pro, July/August 2000.

[16] T. Friedman, A. Bitterer, and B. Hostmann, "Focus on data quality in BI motivates Business Objects buy," ID Number: G00137867, Gartner Research, 14 February 2006.

[17] T. Dasu, G. T. Vesonder, and J. R. Wright, "Data quality through knowledge engineering," SIGKDD '03, 1-58113-737-0/03/0008, 2003 ACM.

[18] I. Davidson, A. Grover, A. Satyanarayana, and G. K. Tayi, "A general approach to incorporate data quality matrices into data mining algorithms," KDD'04, 1-58113-000-0/00/0004, 2004 ACM.

[19] E. M. Pierce, "Assessing data quality with control matrices," Communications of the ACM, Vol. 47, No. 2, February 2004.

[20] J. Hipp, U. G¨untzer, and U. Grimmer, "Data quality mining, making a virtue of necessity," DMKD 2001.

[21] B. T. Dai, N. Koudas, B. C. Ooi, D. Srivastava, and S. Venkatasubramanian, "Column heterogeneity as a measure of data quality," CleanDB, 2006.

[22] "Kalido continues rapid expansion and reports record year of growth - Enterprise data Warehouse and master data management provider records best year to date," Website:'http://www.benchmark.com/news/europe/2006/02_21_2006.php', 21 Feb 2006.

[23] "Online data continues to grow at an explosive pace," Website:'http://domino.research.ibm.com/comm/research.nsf/pages/r.kdd.spotlight.html', IBM Research, 2003. J. Moad, "Mopping Up dirty data," Website:'http://www.baselinemag.com/article2/0,1540,1438233,00.asp'.

[24] M. Hagland, "Stronger computer tools allow deeper analysis of medical research, patient care and insurance data," Website:'http://www.healthcare-informatics.com/issues/2004/04_04/hagland.htm', Healthcare Informatics, April 2004.