A FUZZY SEARCH MODEL FOR DEALING WITH RETRIEVAL ISSUES IN SOME CLASSES OF DIRTY DATA

Olufade, F. W. ONIFADE
University of Ibadan, Ibadan, Oyo State, Nigeria.
olufadeonifade@ieee.org

Oladeji, P. AKOMOLAFE
University of Ibadan, Ibadan, Oyo State, Nigeria.
akomspatrick@yahoo.com

Abstract: Potential capital losses and heightened exposure are inherent in the usage of poor data quality management. Existing efforts like treating data as products; capturing metadata to manage data quality; statistical techniques; source calculus and algebra; data stewardship and dimensional gap analysis all failed in inculcating the contextual factors which a fuzzy in nature. The conventional manner of using information requires discrete values which are precise and devoid of ambiguity, however, this is not realizable as human being employs imprecise expression with high level of uncertainty or no clear boundaries to describe a situation e.g I am very hungry, it is going to be cloudy today. The bulk of the challenges to dirty data can be seen to stem from the "not missing, but wrong data". These result from different data across database, ambiguous data, use of abbreviation or incomplete text and non-standard data which engulf different representation of compound data. This research employs fuzzy model to facilitate retrieval despite these myriads of dirty data problems.

Keywords: Dirty data, Fuzzy search, Fuzzy string matching, Data quality

1. INTRODUCTION

Data and information quality engulf highly complex and huge magnitude of data quality issues based on the comprehensive context and the organizational context. Data quality problems are usually not in isolation, they comprise accumulated, lengthy, and hidden process, and signal root constituting data consumers' experience of difficulties with usage of data (Lee, et al., 2002). It is worth noting that the existence of data quality problem is not limited to automated computer environments, thus attempt at improving data quality problem must consistently and carefully diagnose and improve not only the data but alongside the enabling environment in the specific context. Data environment refers to issues related to collection, storage and usage. Alongside the above are the database systems, information systems infrastructure, related task process mechanism, rules, methods, actions, policies and culture representing a typical organizational composition.

The notion of quality has been described as been "fungible" – the same information can be used by different consumers with widely variant purposes and grossly dissimilar domain of interest (Bovee, et al. 2002). This factor necessitates a high level of flexibility and consistency in the definition. Juxtaposing the above two, we have information quality which has suffered from multiplicity of definitions and views with vivid examples as found in FASB, (1993), Wang, et al. (1995), Wang & Strong (1996).

Potential capital losses and heightened exposure are inherent in the usage of poor data quality management (Even & Shankaranarayanan, (2005). Existing efforts like treating data as product; capturing metadata to manage data quality; statistical techniques; source calculus and algebra; data stewardship and dimensional gap analysis all failed in inculcating the contextual factors (Pipino, et al., 2002). It was thus opined that, once the concept of contextual perspective is generally accepted, there might be need to re-evaluate the current data quality assessment methods. Conducting review research on the impact of data quality on decision performance, Jung, (2004) evaluated the contextual, representational and accessibility of data quality and their influences on decision making. The rationale is such that high quality decision is based on access to information which is complete and relevant to the scope under consideration. Distinction was made amongst data, information and knowledge, from which the relationship between data quality and decision making was established. Contextually, the requirement is that data quality must be viewed in the context of the task at hand, i.e. data must be relevant, timely, complete, and appropriate in terms of amount to be able to add value. Representational and Accessibility data quality implies the importance of the role of the system i.e. the system must be accessible but secured, and present data in a way that they are interpretable, easy to understand, and represented concisely and consistently. In the rest of this work, we reviewed different submission on the concept of data quality leading to the presentation of semiotic analysis of data quality dimensions. Section three presents the fuzzy model for dirty data while section four has the fuzzy search model and its conceptual diagram with operation in real life scenario. We conclude the work in section five.

2. RELATED WORKS

A consistent and accurate chronology of the work on information quality might be difficult to present because of the diversity in focus, or more appropriately the mode of achieving quality presented by different author. This becomes more cumbersome due to the varieties and the clumsy manner with which reference is made to the various dimension of information quality. Having considered a number of them, we settled and expanded the submissions of Tejay, et al., (2006) as our guide. It gives a summarized format for most of the considered quality dimensions and went ahead to deduce a semiotic analysis of the concepts to arrive at four levels: empirics, syntactic, semantics and pragmatics as depicted in the table below.

Semiotic levels	**DQ Dimension**	**Work**
Empirics	Accessibility	Delone, et al., (1992), Goodhue, (1995), Miller, (1996), Wang, et al., (1996), Bovee, (2001)
	Timeliness	Ballou, et al., (1985), Caby, et al., (1995), Fox, et al., (1994), Goodhue, (1995), Hilton, (1979), Miller, (1996), Wang, et al., (1996), Zmud, (1978), Wand & Wang, (1996) Goodhue, (1995)
	Locatability	Caby, et al., (1995)
	Portability	Miller, (1996), Wang, et al., (1996)
	Security	

Syntactic	- Accuracy	Ballou, et al., (1996), Wang, et al., (1996), Caby, et al., (1995), Fox, et al., (1994), Goodhue, (1995), Hilton, (1979), Miller, (1996), Zmud, (1978), Delone,, et al., (1992), Doernberg, et al., (1980), Norman, (2002)
	- Appearance , Comparability, Freedom from bias, Precision, Redundancy, Uniqueness, Usable	Delone, et al., (1992) Zmud, (1978)
	- Arrangement, Readable	Goodhue, (1995)
	- Clarity, Ease of use, Presentation	Miller, (1996), Redman, (1996) Goodhue, (1995), Miller, (1996)
	- Coherence, Format	Caby, et al., (1995)
	- Compatibility	Delone, et al., (1992), Wang, et al., (1996)
	- Composition	
	- Flexibility, Robustness, Conciseness	Ballou, et al., (1985), Caby, et al., (1995), Fox, et al., (1994), Wang, et al., (1996),
	- Consistency	Wand, et al., (1996) Wang, et al., (1996)
	- Correctness	Brodie, (1980) Caby, et al., (1995), Goodhue, et al., (1995)
	- Ease of operation, Objectivity	
	- Integrity	
	- Level of detail	
Semantics	- Ambiguity	Doernberg, et al., (1980), Wand, et al., (1996)
	- Believability, Understandabilty	Wang, et al., (1996) Delone, et al., (1992)
	- Content, Informativeness	Zmud, (1978)
	- Factual, Reasonable	Wang, et al., (1996), Caby, et al., (1995), Bovee, (2001)
	- Interpretability	Goodhue, (1995), Wand, et al., (1996)
	- Meaningful	Brodie, (1980), Delone, et al., (1992), Goodhue, (1995), Zmud, (1978)
	- Reliability	Miller, (1996)
	- Validity	
Pragmatics	- Appropriate amount of data, Reputation, Value-added	Wang, et al., (1996)
	- Appropriateness	Caby, et al., (1995)
	- Completeness	Ballou, et al., (1985), Caby, et al., (1995), Fox, et al., (1994), Miller, (1996), Wang, et al., (1996), Wang, et al., (1996), Doernberg, et al., (1980), Norman, (2002)
	- Relevance	Delone, et al., (1992), Hilton, (1979), Miller, (1996), Wang, et al., (1996), Bovee, (2001), Norman, (2002)
	- Importance, Sufficiency, Usefulness	Delone, et al., (1992)

Table 1: Semiotic Analysis of Data Quality Dimension (Adapted from Tejay, et al., 2006)

Table 1 is a presentation of what is referred to as the semiotic analysis of data quality dimension as opposed to sets of data quality attributes that represent a single aspect or construct of data quality (Wang & Strong, 1996). Data quality problems range from its definition, measurement, analysis, and improvement to tools, methods and processes (Wand, et al., 2001). Teyjay, et al., (2006) defined semiotic interpretation of data quality dimension to address the definition, measurement and analysis aspect of data quality. Alongside, the improvement aspect is implicitly mentioned. One fact stressed was that, dealing with quality attributes such as metrics will ultimately lead away from the main goal into the field of networking which is a deviation from the initial objective. Semiotic broadens the understanding of the interdependencies amongst data, information and knowledge vis a vis data quality.

The semiotic analysis depict that the pragmatic level is associated with knowledge, semantic level is associated with information while only the syntactic level is associated with data (Tejay, et al., 2006). The diagram below further buttresses this submission.

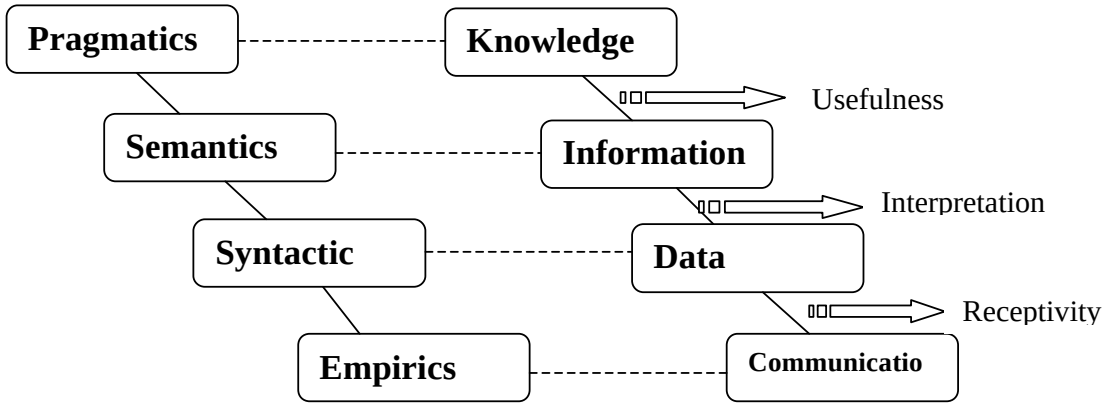

Figure 1: Semiotics, data-information-knowledge and the gap (Tejay, et al., 2006)

Figure 1 thus depicts that the dimensions operating at the pragmatic, semantic and the syntactic level pertains to knowledge quality, information quality and data quality respectively. Consequently, attempt to improve the data quality will focus attention on dimensions operating at the syntactic level. However, it is almost impossible to neglect dimensions associated with knowledge quality and information quality. The trio need be properly harnessed to arrive at a laudable conclusion.

It is important to note that by consolidating data from disparate sources into a “central” position (warehouse) facilitates running of data analysis across application to obtain information that are strategic and tactical towards taking cogent decisions (Inmon, 1999). It is however unfortunate that most of the data kept in the data warehouses for strategic decisions are ‘dirty’. By dirty data we imply that data is either missing or wrong, or it is in a non-standard representation (Williams, 1997). The concept of missing data is encountered in every information retrieval system in existence today. Generally, as the complexity and size of data increases, the issue of missing data becomes expedient.

There are several reasons why the data may be said to be missing the reasons stem from both human

related faults and machine malfunctions. Examples are: data may be missing because equipment malfunctioned, the weather was terrible, people got sick, or the data were not entered correctly.

2.1 Missing Completely at Random

In the above cases, the data are said to be *missing completely at random (MCAR).* When we say that data are missing completely at random, we mean that the probability that an observation (X_i) is missing is unrelated to the value of X_i or to the value of any other variables. Thus data on family income would *not* be considered MCAR if people with low incomes were less likely to report their family income than people with higher incomes.

Similarly, if Whites were more likely to omit reporting income than African Americans, we again would not have data that were MCAR because missingness would be correlated with ethnicity. However if a participant's data were missing because he was stopped for a traffic violation and missed the data collection session, his data would presumably be missing completely at random. Another way to think of MCAR is to note that in that case any piece of data is just as likely to be missing as any other piece of data (Dunning, & Freedman, 2008).

Notice that it is the value of the observation, and not its "missingness," that is important in this regard. If people who refused to report personal income were also likely to refuse to report family income, the data could still be considered MCAR, so long as neither of these had any relation to the income value itself. This is an important consideration, because when a data set consists of responses to several survey instruments, someone who did not complete the Beck Depression Inventory would be missing all BDI subscores, but that would not affect whether the data can be classed as MCAR. This nice feature of data that are MCAR is that the analysis remains unbiased. We may lose power for our design, but the estimated parameters are not biased by the absence of data.

2.2 Missing at Random

Another dimension considered by Howell, (2009) is known as *missing at random*. Often data are not missing completely at random, but they may be classifiable as missing at random (MAR). For data to be missing completely at random, the probability that Xi is missing is unrelated to the value of Xi or other variables in the analysis. But the data can be considered as missing at random if the data meet the requirement that missingness does not depend on the value of Xi after controlling for another variable.

2.3 Missing Not at Random (MNAR)

If data are not missing at random or completely at random then they are classed as *Missing Not at Random (MNAR).* For example, if we are studying mental health and people who have been diagnosed as depressed are less likely than others to report their mental status, the data are not missing at random. Clearly the mean mental status score for the available data will not be an unbiased estimate of the mean that we would have obtained with complete data. The same thing happens when people with low income are less likely to report their income on a data collection form (Dunning, & Freedman, 2008).

When we have data that are MNAR then, the problem is significant. The only way to obtain an unbiased estimate of parameters is to model missingness. In other words we would need to write a model that accounts for the missing data. That model could then be incorporated into a more complex model for estimating missing values. This is not a task anyone would take on lightly.

Kim, et al., (2003) provide hierarchical refinement approach to arrive at their taxonomy of dirty data. Their taxonomy was based on the premise that dirty data manifest itself either as missing data, not-missing but wrong, and not-missing and not wrong but unusable. The hierarchy decomposes these manifestations and also represents dirty data resulting from more than one type of dirty data. A comprehensive listing of their taxonomy is presented in Onifade, (2010).

- Missing data
  - Missing data where there is no Null-not-allowed constraint
  - Missing data where Null-not-allowed constraint should be enforced
- Not-missing, but
  - Wrong data, due to
    - Non-enforcement of automatically enforceable integrity constraints
      - Integrity constraints supported in relational database systems today
        - User-specificable constraints
          - Use of wrong data type (violating data type constraint, including value range)
          - Dangling data (violating referential integrity)
          - Duplicated data (violating non-null uniqueness constraint)
          - Mutually inconsistent data (action not triggered upon a condition taking place)
        - Integrity guaranteed through transaction management
          - Lost update (due to lack of concurrency control)
          - Dirty read (due to lack of concurrency control)
          - Unrepeatable read (due to lack of concurrency control)
          - Lost transaction (due to lack of proper crash recovery)
      - Integrity constraints not supported in relational database systems today
        - Wrong categorical data (e.g., wrong abstraction level, out of category range data)

        - Outdated temporal data (violating temporal valid time constraint; e.g., a person's age or salary not having been updated)
        - Inconsistent spatial data (violating spatial constraint; e.g., incomplete shape)

- Non-enforceability of integrity constraints
    - Data entry error involving a single table/file
        - Data entry error involving a single field
            - Erroneous entry (e.g., age mistyped as 26 instead of 25)
            - Misspelling (e.g., principle instead of principal, effect instead of affect)
            - Extraneous data (e.g., name and title, instead of just the name)
        - Data entry error involving multiple fields
            - Entry into wrong fields (e.g., address in the name field)
            - Wrong derived-field data (due to error in functions for computing data in a derived field)
    - Inconsistency across multiple tables/files (e.g., the number of Employees in the Employee table and the number of employees in the department table do not match)

- Not wrong, but unusable data
    - Different data for the same entity across multiple databases (e.g., different salary data for the same person in two different tables or two different databases)
    - Ambiguous data, due to
        - Use of abbreviation (Dr. for doctor or drive)
        - Incomplete context (homonyms; and Miami, of Ohio or Florida)
    - Non-standard conforming data, due to
        - Different representations of non-compound data

- Algorithmic transformation is not possible
  - Abbreviation (ste for suite, hwy for highway)

As regards the categorization of Kim, et al., on the concept of dirty data, there has been however various methods for treatment of dirty data (Dunning, & Freedman, 2008, Inmon, 1999, Kim, et al., 2003). Among these are commercial software tools for creating data warehouses or transforming data for multidimensional analysis or data mining with several ways to replace missing data in a field with mean arithmetic values. The effectiveness of this 'guess' could sometimes be detrimental. Incomplete text, use of abbreviation and other forms of missing data can be handled in a better manner.

3. FUZZY MODEL FOR DIRTY DATA

It is not uncommon for system data to degrade rapidly, this can commence with customers information, for example names, addresses and missing information. The rate at which errors like these accumulate can be in matter of days, few weeks or might even take longer time. Unfortunately information from such database becomes unreliable. Error is not limited to the size of a database or the organization, even in professionally designed, implemented and operated with strict data control, there exist errors which constitute risks inimical to the organization. The focus of this research is not in detection and removal or what is referred to as data cleaning, but improve search operation despite the level of dirtiness of the database.

We referred to dirty data as a term employed to refer to information/data that is *misleading, incorrect or without generalized formatting*, that has been collected by any data-capture means. This could be in form of *spelling mistake or punctuation, incomplete or outdated data*, or even data that has been *duplicated* in the database (Mike, 2009).

The taxonomy presented in Kim, et al., (2003) is based on the premise that manifestation of dirty data comes in three broad ways: *missing data, not missing but wrong data, and not missing and not wrong but unusable*. The last occurrence is more pronounced whenever there is database integration or when representation standards are not consistently pursued in inputting data. The taxonomy also represents dirty data that are manifested based on the combination of more than one type of dirty data (e.g. wrong order in data concatenation, misspelling – "Amos David" instead of "David Amos").

This taxonomy is aimed at providing a framework for understanding the origins of a complete spectrum of dirty data and the impact of dirty data on data mining, and sheds light on techniques for dealing with dirty data and for defining a metric for measuring data quality. In resonance with the submissions of Kim, et al., (2003), we represent below in simple hierarchical structure some classes of dirty data which our model addressed.

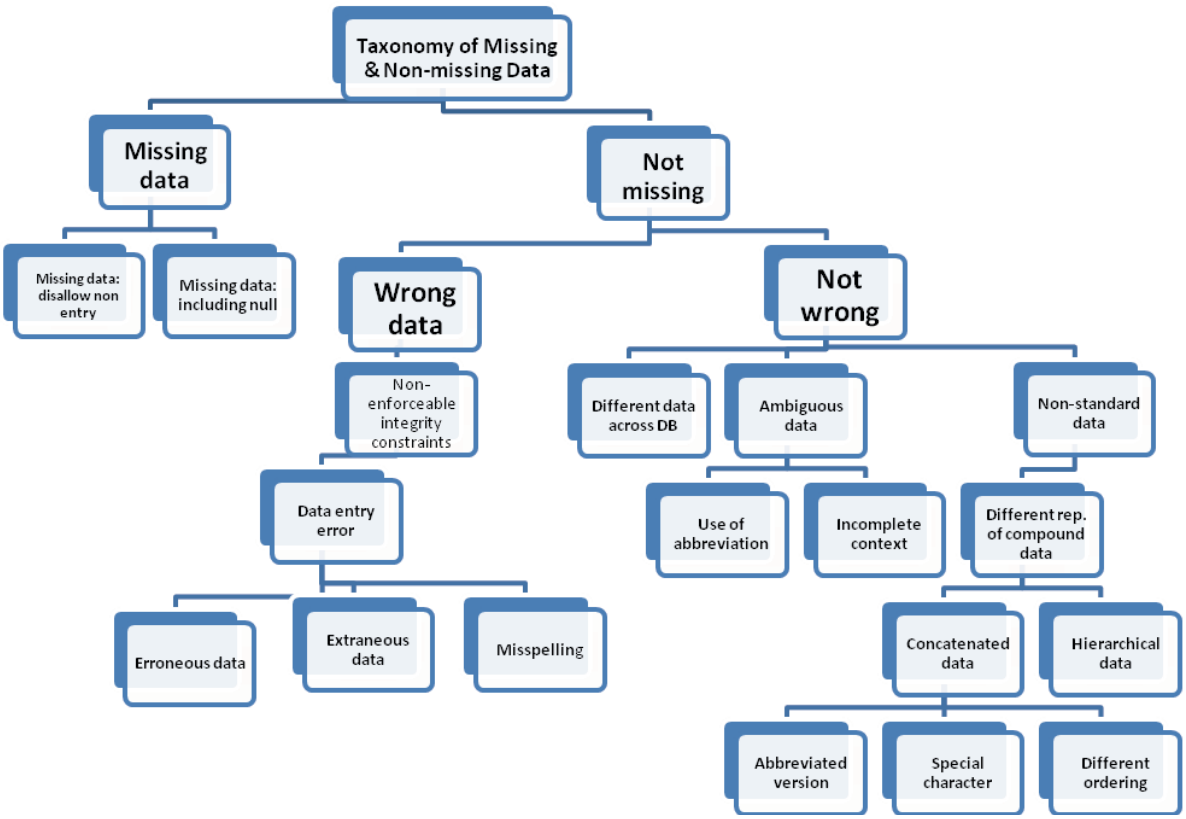

Figure 2: A tree structure for Data Missingness & Non-missingness (Adapted from Kim, et al., 2002)

The life cycle of data includes its capture, storage, update, transmission, access, archive, restore, deletion, and purge. The focus of our research is on the access aspect by a user or application that operates correctly. *As such, we say that data is dirty if the user or application ends up with a wrong result or is not able to derive a result due to certain inherent problems with the data.* The sources of dirty data include data entry error by a human or computer system, data update error by a human or computer system, data transmission error by a computer system, and even bugs in a data processing computer system. In the following section, we present our model for dealing with some classes of dirty data towards a more meaningful information retrieval process. This is sequel to the fact that its also possible apart from the above listed factor to have dirty data problems based on other exogenous factors rather than the state of the data in its entirety.

4. FUZZY SEARCH MODEL

Fuzzy searching is much more powerful than exact searching when used for research and investigation based on its accommodation for imprecision and ambiguity. Fuzzy searching technique comes in handy when researching unfamiliar, foreign-language, or sophisticated terms, the proper spellings of which are not widely known or asserted. Fuzzy searching can also be used to locate individuals based on incomplete or partially inaccurate identifying information in an attempt to deal with dirty data. A fuzzy search is done by means of a fuzzy matching program, which returns a list of results based on likely relevance even though search argument words and spellings may not exactly match. Exact and highly relevant matches thus appear near the top of the list while subjective relevance ratings, usually are expressed in percentages form.

Search operation involves a lot of string manipulations. Their efficiency is thus closely linked to the performance of the algorithm upon which they are implemented. We employed the Transfer function (Tanino, 1984) employed in fuzzy preference ordering in group decision making to evaluate each of the sub-strings with the alternatives toward the matching of a query as given in eqn. [1].

$$p_{ij}^{k} = f(x_i^k, x_j^k) = \frac{1}{2}(1 + (x_i^k \ominus x_j^k)) \qquad [1]$$

p_{ij}^{k} characterizes the match-preference degree between alternative sub-strings a_i and a_j expressed via $\mu_f(x)$ and Θ is the subtraction operation on two fuzzy sets.

Again, from the principle of Pseudo-Order Preference Model (POPM) for determining the preference of one or more pseudo-criteria, (Wang, et al., 2006) three fundamental preference relations in classical preference structure suffices. The last was manipulated to have the following: Strict match ($\mathcal{M}$), Weak match ($\mathcal{W}$) and Fuzzy match ($\mathcal{F}$) which is capable of generating inferences even if the order/arrangement of the sub-string confuses the retrieval system. To this end, instead of generating no match-found, the fuzzy match operates based on the predefined membership function and the comparison of the sub-strings. We adapt these to arrive at the following equations:

Strict match relation ($a_i \mathcal{M} a_j$)

$$M_{ij}^{k} - m_{ji}^{k} > m \qquad [2]$$

Weak match relation ($a_i \mathcal{W} a_j$):

$$\mathcal{W} < M_{ij}^{k} - m_{ji}^{k} \leq m \qquad [3]$$

Fuzzy match relation ($a_i \mathcal{F} a_j$):

$$|M_{ij}^{k} - m_{ji}^{k}| \leq \mathcal{W} \qquad [4]$$

(where k = 1, …, m; i, j = 1, …, n)

Equation 4 represents the major drive away from popular search engines. The fuzzy match relation permits us to accommodate a high level of ambiguity which would have hitherto generated a ‘no match found’ to users query as a result of dirty data. Following, we present the conceptual diagram depicting the operational sequence of the fuzzy system.

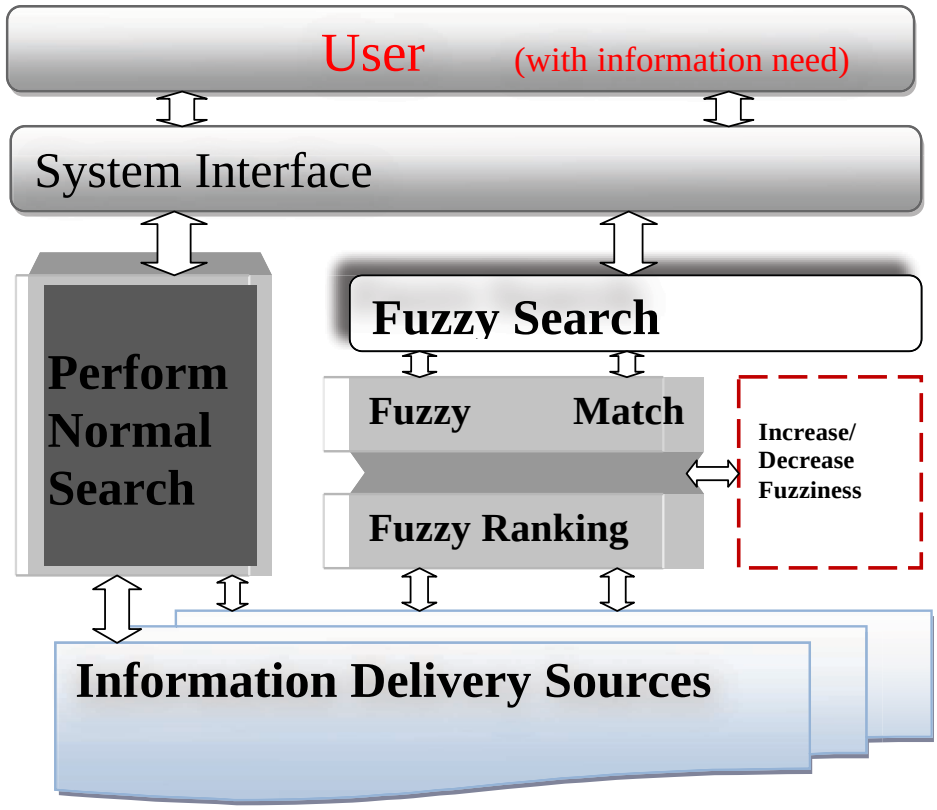

Figure 3: Conceptual Diagram for Fuzzy Search model

The process of information retrieval can either be implicit or explicit based on the decision maker's "understanding of the problem", "familiarity with the terrain" and the "size of the problem" amongst others. Decision making requires explicit information search usually bears various forms of risk which makes accessibility to qualitative information an important factor in decision making. Sadly, with exponential growth in the volume of information in digital format, existing information retrieval algorithms, methods, technologies and tools are too constricted for adaptive and robust information retrieval activities which continue to render the mode of handling dirty data a herculean task. This factor has rubbished many attempts to inform because of inability to access adequate information timely. On the part of the users, incomplete and partial understanding exist making the process cumbersome with imprecision and ambiguities. On the part of the system, lack of flexible representation of queries and documents exist. These factors constitute high risk to decision problem resolution. The following excerpt from what is tagged as the "systemic failure of intelligence" in the December 25th saga of the Detroit bound plane will form the basis of our discussion next.

President Obama reiterated that "once again, it is a failure of the US intelligence agencies that we are told, and are to be blamed. The report found out that the **US government did have sufficient information** to disrupt the Christmas day attack. But that **information was scattered around databases.** It was never pulled together to **present a coherent picture of threat.** A series of **human errors occurred,** apparently someone **misspelled Umar Farouk Abdulmutallab's name** as entries in the database and that was why no one realised he had a US visa" – CNN, BBC News, (2010).

The wrong spelling of Umar Farouk Abdulmutallab's name serves as one of the basis for dirty data that we addressed in this research. The effect of which could have been devastating if successful in the case presented above. Its therefore imperative that flexible means of operation on these dirty data should be weaved around the major design of our information delivery systems. We researched the internet to determine the possible

representation of the name. Our findings depict that there were more than 7 different representation of the same name leading to gross inconsistency that could be very costly.
The conceptual fuzzy search shown in figure 3 is powered by different modules and functions among which is the fuzzy string matching (Onifade, et al., 2010). The aim of this is to ameliorate the erroneous ambiguity resulting from human errors that are detrimental to the sourcing and usage of information. Below, we present the fuzzy string-matching model and we discuss its operation as regards the problem of "systemic failure of intelligence".

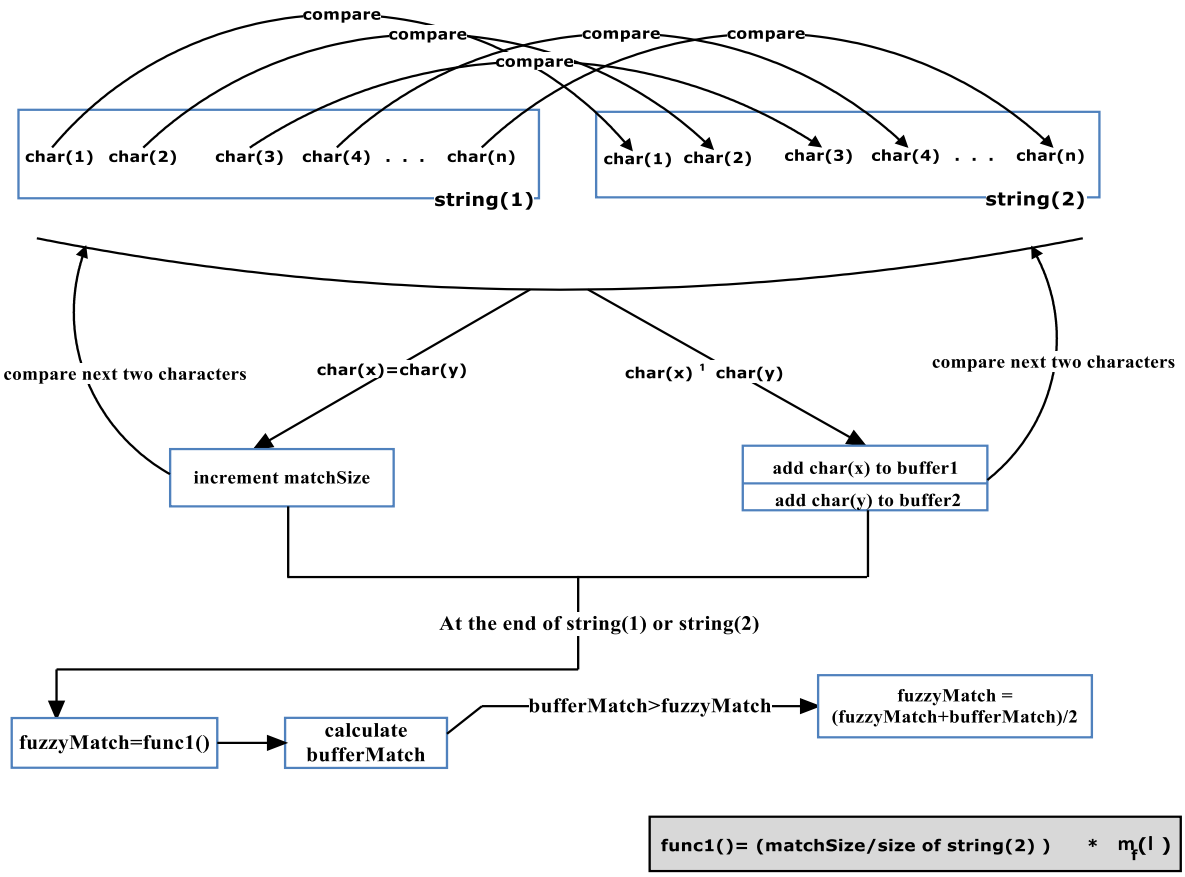

Figure 4: Fuzzy String Matching Model

The systemic failure of intelligence resulted principally from human errors, basically, the president reported that someone misspelled the name and the following are our findings for some of the representation of the same name: Abdul Farouk; Umar Farouk Abdul; Umar Farouk; Umar.Farouk.abdul; Farouk Abdul; Abdul Mutallab and Abdulmutallab. The bulk of the challenges to dirty data can be seen to stem from the "not missing" arm of the hierarchical structure presented in figure 2. We could easily see the following: different data across database, ambiguous data (use of abbreviation or incomplete text), and non standard data which engulfs different representation of compound data (hierarchical data, concatenated data (abbreviated version, special character and different ordering)).

Taking a critical look at the names cropped from the internet, it is possible for human being to make a meaning or deduce interrelationship amongst the seven different names, but this is not the same with database systems because of their inability to handle ambiguities. Thus we can first and foremost see the first sub-class – different representation across databases, some form of concatenation, different representation of compound data, abbreviated

versions and different ordering as the source of dirty data that could have cost the lives of around 256 passengers in the Detroit bound plane.

In other to favourably and concurrently compare the user's string during search and the database content, two dynamic buffers were created at the commencement of the operation. One holds the unmatched characters of the user sub input *'buffer1'* and the other holds the unmatched characters of the database substring *'buffer2'*. The algorithm then scans the character content of the two strings concurrently. When the characters are similar, the variable indicating how many characters were matched is incremented. If the characters are dissimilar, the two characters are stored in *buffer1* and *buffer2* respectively. After all the characters might have been compared, it gets to the end of one of the strings (in the case where the size of the two strings are not the same), the fuzzy match value is calculated based on the level of containment or belongingness (via fuzzy membership function) of the *matched character size* and the *size of the database substring*. The above operation does not do away with the unmatched characters, instead they are considered to generate some other entries to be displayed alongside the retrieved entries. A full discussion of the above can be found in Onifade, 2010.

Differences amongst Abdul Farouk; Umar Farouk Abdul; Umar Farouk; Umar.Farouk.abdul; Farouk Abdul; Abdul Mutallab and Abdulmutallab are seemingly apparent and detectable to human eyes. This follows in real life that taking strategic decision involves resolving ambiguities posed by these various facets of dirty data which are detrimental if not properly handled. Evaluating the match-preference degree between several alternatives sub-strings a_i and a_j expressed via $\mu_f(x)$ and θ as presented in equ. 1 was employed in figure 3 to determine the relationship amongst various entries representing the name. the fuzzy model is capable of achieving different level of fuzziness (figure 1) which serves well in dealing with the stated classes of dirty data.

5. CONCLUSION

Dirty data consist in impossible phone number, nonexistence postal code or future birth date amongst others as example of invalid data. This type can easily be fixed than other types of dirty data. Detecting incomplete data is more difficult than invalid data, however inconsistent type may prove much more difficult to detect since it requires more inside knowledge (substitute "rules" or "metadata"). The most worrisome of these is the incorrect data. This is sequel to the fact that it is valid, complete and consistent, yet it is just wrong. Thus it will not be detected by validation, completeness, or consistency check. They are almost intractable. This work has thus presented a model aimed at solving this problem via the use of soft-computing methodology.

REFERENCES

[1] Bouyssou, D. (1989) : "Modeling Inaccurate determination, Uncertainty, Imprecision using multiple criteria". In A.G. Lockett and G. Islei, editors, *Improving Decision Making in Organisations*, LNEMS 335, pages 78-87, 1989. Springer-Verlag, Berlin. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/1243 on 01/10/2009

[2] Bovee, et al., M. (2001): "A Conceptual Framework and Belief-Function approach to Accessing Overall Information Quality". Proc. Of 6th International Conference on Information Quality (ICIQ' 01), pp 311-324.

[3] Brin, S., Page, L., Motwani, R., Winograd, T.: The PageRank Citation Ranking: BringingOrder to the Web, Technical report, Stanford University, 1998

[4] Buche, P., Dervin, C., Haemmerle, O. & Thomopoulos: "Fuzzy Querying of Incomplete, Imprecise & Heterogeneously Structured Data in the Relational Model using Ontologies & Rules". IEEE transaction on Fuzzy Systems, vol 13, no 3, June 2005, pp 373-383.

[5] *Dunning, T., & Freedman, D.A. (2008) Modeling section iffects. in Outhwaite, W. & Turner, S. (eds) Handbook of Social Science Methodology. London: Sage*

[6] Even, A. & Shnakaranarayanan, G. (2005): "Value-Driven Data Quality Assessment In Proceedings of the International Conference on Information Quality". ICIQ-05, Cambridge, M.A, Nov, 2005.

[7] *Howell, D. C., (2009): "Treatment of Missing Data"*

http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html
http://choo.fis.utoronto.ca/fis/respub/aisis99/.

[8] Inmon, W.H. 1999. Data Warehouse Performance. New York: John Wiley.
Jung, W. (2004): "A Review of Research: An Investigation of the Impact of Data Quality on Decision Performance". International Symposium on Information & Communication Technologies (ISITC'04), pp 166-171

[9] Kim, W., Choi, B-J., Homg, E,-K; Kim, S-K and Lee, D.(2003) "A Taxonomy of Dirty Data". Data Mining and Knowledge Discovery, 2003, Kluwer Academic Publishers. 7.2003. pp. 81-99

[10] Lee, Y. W., Strong, D. M., Kahn, B. K. & Wang, R. Y. (2002): AIMQ: A Methodology for Information quality assessment. Elsevier Information and Management 40 (2002) 133 – 146.

[11] Mike, (2009): "The problem of dirty data" http://www.articlesbase.com/print/1111299 Accessed on 05/11/2009
Navarro G (2001). "A guided tour to approximate string matching". *ACM Computing Surveys* **33** (1): 31–88. doi:10.1145/375360.375365.

[12] Onifade O.F.W., (2010) "Intelligent Retrieval Tool for Strategic Information Risk Management" Verlag Dr. Muller (VDM) Aktiengesellschaft & Co. Kg. dudweiler Landstr. 99, 66123, Saarbrucken, Germany.

[13] Onifade, O. F. W., Thiéry, O., Osofisan, A. O. & Duffing, G.: "Dynamic Fuzzy-String Matching Model for Information Retrieval Based on Incongruous User's Queries". Presented paper at World Congress on Engineering Conference (WCE'10), London, U.K., 30 June - 2 July, 2010.

[14] Onifade O.F.W., Thiéry O., Osifisan, A.O. & Duffing G.: "A Fuzzy Model for Improving Relevance Ranking in Information Retrieval Process". Presented paper for presentation at the 2010 International Conference on Artificial Intelligence & Pattern Recognition (AIPR -10), Florida, USA, July, 2010.

[15] Pipino, L.L., Lee Y. W. & Wang, R.Y. (2002): "Data Quality Assessment". The Communication of the ACM, April 2002, vol. 45, no 4e, pp 211-218.

[16] Robertson, S.E., Sparck Jones, K.: Relevance weighting of search terms, Journal of the American Society for Information Sciences, 27(3), 1976, pp. 129–146

[17] Tang Kai Yin, C. S. George Lee, (1995) Fuzzy Model-Reference Adaptive Control, IEEE Transactions on Systems, Man, and Cybernetics, vol. 25, no. 12, December 1995.

[18] Tanino, T. (1984). "Fuzzy Preference Ordering in Group Decision Making", fuzzy sets and systems, vol. 12, pp117 – 131

[19] Tayi, G. K., & Ballou (1998) "Examining Data Quality". In communications of the ACM, 41, 2, Feb. 1998

[20] Tejay, G., Dhillon, G., & Chin, A. G.: "Data Quality Dimension for Information system security: A theoretical exposition" – Invited paper. In Security management, Integrity & Internal Information Control in Information Systems, Springer Books, Boston, pp 21 – 39, 2006.

[21] Wang, R.Y., Reddy, M.p., & Kon,(1995) H.B Decision Support Systems, 13, 349-372.

[22] Wagner, R. A. & M. J. Fisher. (1974). The string-to-string correction problem. Journal of the Association for Computing Machinery, 21(1):168-173, January 1974

[23] Wand, Y. & Weber, R.: "On the Ontological Expressiveness of Information Systems Analysis and Design Grammars". Journal of Information Systems (1993) 3, Pp. 217 – 237.

[24] Wang, H., Johnson, T. R. & Zhang, J. (1998): "UEcho: A Model of Uncertainty Management in Human Abductive Reasoning". In proc. Of 20th Annual Meeting of the Cognitive Science Society, 1998, Hillsdale, NJ

[25] Wang, R. & Strong, D. M.: "Beyond Data Accuracy: What Data Quality Means to Data Consumers". Journal of Management and Information System, Springer 1996, Vol. 12, No. 4, Pp. 5 – 34.

[26] Wang, P., Chao, K., Huang, C., Lo, C. & Hu, C.: "A Fuzzy Decision Model of Risk Assessment through Fuzzy Preference Relation with Users' Confidence Interval". Proc of 20th Int'l conference on advance Information Network and Application, (AINA'06), IEEE Society, 2006, Pp. 889 – 893.