# ICT for Basic Service Delivery - Case Study: Water Sector in Uganda

by

**IKAE CATHERINE OMAL**

E-mail: ikae.catherine@cit.mak.ac.ug, cikae_3@yahoo.com, Tel: +256 772 695998

**Abstract**

Mobile information and communication technologies (ICT) and crowdsourcing has been used to generate large volumes of data in a short time and effectively. This kinds of quick means of data collections has enabled quick response in disaster management situations in developed countries but it has not yet been used in public service provision especially in the developing countries due to the quality of the data collected in such countries with poor infrastructure and limited knowledge on use of some of the Mobile ICT devices. This research therefore proposes identifying quality control techniques likely to work well for each data obtained from different sources(sms, web, voice, text- or image-based) so as to validate and modify them for the context of least developed countries since classifier performance depends greatly on the characteristics of the data to be classified. The task will be using data obtained from the water sector with the view of improving on the quality of that data which will later be used to improve basic service delivery in the water sector in Uganda.

## 1. Background, research hypotheses and objectives

In the recent past, mobile information and communication technologies (ICT) and

crowdsourcing for generating, collecting, processing, storing and disseminating data have been advancing rapidly; these has enabled organizations and individual to provide timely information especially as a result of Internet use, mobile phones, email and sms**??**. Within the various sources of data are the possible errors incurred during mobile data collection. If the quality of the data is improved it would generate new and useful knowledge to support allocation of resources, improvement of performance, and respond to emergencies**?**. Hence leading to better service delivery especially in the water sector to facilitate provision of safe drinking water to the communities in Uganda.

Crowdsourcing is a process that involves outsourcing tasks to a distributed group of people**?**. This has resulted into the collection of huge volumes of data within a short time period. This has brought in new opportunity to research to carry out research in ways of ensuring data quality. Crowdsourcing has its own obstacles, the quality of the data cannot be controlled manually because of its volume**?**.

Its almost certain crowdsourcing will dramatically change the nature of work and creativity in the future of public service provision. As crowdsourcing continues to make previously scarce resources such as information become much more abundant, the attitude towards service provision will change dramatically since the service consumer have more power in their hands in form of information**?**. This will likely have very far reaching implications in areas such as improved public service provision. It is therefore important to ensure data quality of decentralized and voluntary information collection for this resource to have its desired effect.

Mobile devices have shown great promise for improving the efficiency and effectiveness of data collection in resource-poor environments as already applied in the water and sanitation sector in Uganda**?**. The advantage of mobile devices is that they offer

immediate digitization of collected data at the point of survey. This allows for fast and automated data aggregation. It is therefore critical to ensure the accuracy of data entered for purposes of improving service provision.

This research starts by recognizing that mobile information and communication technologies (ICT) and crowdsourcing are available for the collection of data from end users by service providers in an attempt to improve on the services provided**???**. It also looks into the sources of data such as text, sms, emails and mobile phone in an attempt to determine the most appropriate source. In order to determine the source of data that provides greater accuracy in the collected data, several Machine learning algorithms will be used to test accuracy of the different sources of data and modified to determine the best technique that works well for a particular data source. The best technique will then be used to automate the quality control of the data collected from that particular source.

Mobile ICT and crowdsourcing have been successfully used and deployed in developing countries for things like disaster management but they have not yet been used in the area of improvement of public service provision due to some technical hurdles which among them is ensuring data quality of decentralized and voluntary information collection to be used in this area.

Research into automated data quality control is very important because the volume of data collected using mobile ICT and crowdsourcing is growing rapidly and has exceeded the ability to be analyzed manually**?**. It is also important because the interpretation of this kind of data depends greatly on its quality.

Machine learning algorithms have been applied in some areas to try and automate the control of data quality but still more work is needed to detect interesting features with

machine learning algorithms as well as to search for quality problems to improve on data quality**?**. Since machine learning algorithms have successfully been used to detect anomalies in SMS collected**?**, but has not yet been applied to web, voice, text- or image-based data. There is need to train these classifiers with web, voice, text- or image-based data to compare there performance and determine the technique that suits each data source. The training of this classifiers will help in the validation and modification of this classifiers for the context of least developed countries such as Uganda.

For any improvement in basic service delivery, interpretation of data obtained from the service consumer is critically dependent on the quality of data obtained. The quantity of the data obtained has also significantly increased by the introduction of Mobile ICT and crowdsourcing in data collection. Hence there is a need to determine the best technique for each data source(sms, web, voice, text- or image-based) and the overall best technique to be used for automated quality control of the large volumes of data to ensure that quality information is derived from the collected data.

The general objective of the research is to train classifiers for automated Quality control of ICT and Crowdsourced sms, web, voice, text- or image-based data from the water sector using Machine learning algorithms so as to validate and modify them. To Design , develop and Evaluate the improved classifiers . To determine the best Algorithm for automated quality control for each of the data source. To determine the best classifier for automated quality control of ICT and crowdsourced data.

These will be achieved by: Investigating the types of data obtained by the different sources (sms, web, voice, text- or image-based), investigating Machine learning algorithms in relation to the attributes and characteristics specific to the different types of data, Training classifiers using Machine learning algorithms for automated quality control of the types of data, determining the best classifier for each data source and

determining the best classifier for automated quality control of ICT and crowdsourced data.

The sms, web, voice, text- or image-based data will be obtained from the water sector in Uganda because this sector might highly benefit from mobile ICT technologies as demonstrated by the recent ICT projects targeting the water sector**?**.

Research questions

(a) what are the appropriate machine learning algorithms suitable for automated quality control of the data obtained from the different sources?

(b) What is the most appropriate Machine learning algorithm for automated quality control of ICT and crowdsourced data?

The research will cover development of classifiers for automated quality control of ICT and crowdsourced data from the water sector in Uganda. It will also determine the appropriate algorithm for each source of data as well as the best algorithm for quality control measure. Validation of the classifiers will be limited to comparing performance of different techniques on data form the different sources.

Due to the high penetration of mobile phone usage in Uganda and the desire of many organization to get feed back from the service consumers, there has been an increase in the amount of data collected with the combination of ICT and crowdsourcing. For this data to be relevant for decision making, it is important that its of high quality with minimum errors so that the response can lead to better service delivery in case of the water sector in Uganda.

This research could lead to better data quality control. It will also provide an option of researchers to test for data quality. The classifiers will be able to assist software developer to analyze and identify errors in less time and efficiently and is cost-effective

in terms of time and manpower. This would eventually lead better service provision due to the improved quality control of data.

2. **State of research in the field**

This section looks at Machine learning in general, use of machine learning in data quality control, the classification techniques that have already been used in data quality control. It also looks at some classification techniques that have not yet been used in data quality control.

Machine learning is a branch of artificial intelligence in which a computer generates rules based on raw data that has been fed into it. It detects patterns in data and adjust program actions accordingly. Machine learning and computer vision have significant overlap since they are both based on learning algorithms for future outcomes**??**.

Machine learning techniques can be used for quality assessment by considering the required technical characteristics for specific data quality problem. Each of this problems can therefore be addresses by applying the various machine learning techniques**?**. All of these with the aim of providing data with an appropriate level of quality in a timely and cost-effective manner.

Naive Bayes Classification

A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. It allows classification of a new object based on the current existing object. It is derived from Bayes Theorem, which states that the evidence confirms the likelihood of a hypothesis only to the degree that the evidence would be more probable with the assumption than without it.

Bayes' formula is: $prob(X|Y,I) = prob(Y|X,I)prob(X|I)/prob(Y|I)$, that is, the posterior probability is proportional to the product of the likelihood function and the prior

probability. In supervised Bayes classification, we begin with a training set where each evidence (data) vector E has been assigned to a class C. Training consists of computing the probability density function for each combination $E_i$ and C, i.e., $Pr(E_i|C)$, and the overall probability for each class. The forward application of the classifier computes the probability for each possible class C as: $Pr(C|E) = PPr(E_i|C)Pr(C)/Pr(E)$ (In practical applications, the denominator is often bypassed by normalizing over all the classes.) Naive Bayes is often used as a baseline in text classification because it is fast and easy to implement**?**.

Probabilistic Networks

Probabilistic networks (e.g., Bayesian networks) are graphs consisting of nodes and arcs which are used to represent assertions and causal belief such as the probability that a value of one node causes a given value in another node. They are therefore likely to be good to be applied to data quality assessment in cases where there are a number of known factors that determine the resulting quality of a set of data. A special case of this is when corroboration is needed to confirm a data quality problem: the nodes could represent features, events, or quality-related factors extracted from several data sources, and the arcs could represent judgments about data quality in light of the additional evidence**?**.

Neural Networks

Neural networks perform classification using nodes that perform a simple transfer function connected either in a non-recurrent structure (i.e., without feedback, such as feed-forward/back propagation networks also known as supervised) or a recurrent structure (e.g., Kohonen self-organizing maps which are unsupervised classifiers). Neural networks work well with continuous data, which can be an advantage for data quality assessment in Earth science archives**?**. Neural networks can also handle categorical data, by mapping each category value to a separate node and representing the value

itself as 0 (not present) or 1 (present). Neural networks can be used on temporal data through the use of "sliding window" techniques or recurrent network structures**?**.

Support Vector Machines

Support Vector Machines performs classification by considering each object as an $n$ dimensional feature space, $n$ being the number of features used to describe the objects numerically. Each object is also assigned a binary label, let us assume the labels are "positive" and "negative". During learning, the algorithm tries to find a hyperplane known as "support vectors" in that space, that perfectly separates positive from negative objects**?**. For the above reason, Support Vector Machines can be used to perform complex mappings of data into a feature space where a hyperplane can be used to separate the classes**?**. In terms of handling categorical and continuous data, they are similar to neural nets, and thus would likely have similar applicability to various data quality assessment problems**?**.

## 3. Methodology

The study proposes selecting appropriate quality control technique for sms, web, voice, text- or image-based data, the methodology to be used will investigate the structure of the data obtained through sms, web, voice, text- or image-based, Examining the quality control theories in relation to attributes and characteristics of each of the data obtained through sms, web, voice, text- or image-based. Training the different classifiers with sms, web, voice, text- or image-based data. The classifiers will be validated and modified to work with the locally generated data obtained from the water sector in Uganda. The performance of the different techniques will then be compared in order the choose the appropriate technique for each data. Evaluation of the classifiers will enable the selection of the most appropriate technique that can be adopted for automated data quality control.

Data collection

- Data Collection from the water sector through ICT and crowdsourcing.
- Data Characterization that is extracting from each data set those characteristics thought to be correlated with data quality and also to transform variable to a state that can be used in machine learning.
- Reference Sub setting, which is capturing and storing reference data sets.
- Data Corroboration, is used to distinguish between unusual (but real) features and quality artifacts.

Training of classifiers

In the training phase, the training set will be used to decide how the parameters ought to be weighted and combined in order to separate the various classes of objects, the weights determined in the training set will then be applied to a set of objects that do not have known classes in order to determine what their classes are likely to be. The training phase is decisive on the final results. The larger the training set, the better the classifier. At this stage Important features are given a high weight, while unimportant features may not be used at all.

Testing of classifiers

Testing of all the trained classifiers will be done on all data obtained from the different sources (ie sms, web, voice, text- or image-based data obtained from the water sector in Uganda). Once a potentially useful classifier has been constructed, the accuracy of the classifier must be measured. Knowledge of the accuracy is necessary both in the application of the classifier and also in comparison of different classifiers. Five-fold cross-validation will be used to measure the accuracy of classifiers. This will be done by dividing the training set into five randomly selected subsets having roughly equal numbers of objects. The classifier is then trained five times, excluding a single subset

each time. The resulting classifier is tested on the excluded subset. Each training session must be completely independent of the excluded subset of objects; one cannot results of an earlier training session as a starting point for a new training session. The advantage of cross-validation is that all objects in the training set get used both as test objects and as training objects. This ensures that the classifier is tested on both rare and common types of objects.

Design of improved classifiers

The modified classifiers will be designed with the aim of improving their performance in terms of data quality control. By adjusting the classification threshold, adjusting number of votes required to classify positive, varying the probability threshold for classifying as positive and varying margins for positive and negative examples.

Development of improved classifiers

The development of the improved classifier will be carried out through the training process using the collected data to check changes in the performance as compared to previous unmodified classifiers.

Evaluation of improved classifiers

The performance the classifiers will be evaluated by calculating the error rates, accuracy rates, Precision, Recall, Sensitivity and specificity. The values calculated will be compared to determine the best classifier appropriate for data quality control in the context of the data in use.

4. **Timeframe and Milestones**

5. **Organization**

6. **Expected results**

This research will deliver an appropriate automated quality control technique. Other deliverables will include:

(a) A working prototype of the selected automated data quality control technique.

(b) PhD thesis.

(c) Publication of at least two journal papers on the following topics:

- Automated data quality control: technical, theoretical, and methodological issues;
- Selecting appropriate data quality control techniques.

7.

| **Phase** | **Description** | **Duration (months)** | Period | Deliverables |
|---|---|---|---|---|
| 1. | **Literature Review**<br>Quality control theories<br>Data characteristics and attribute | | | draft proposal |
| 2. | **Literature Review**<br>Quality control tasks<br>Quality control techniques | | | Chap 1, 2 and 3 |
| 3. | **Data collection and preprocessing**<br>Data characterization<br>Reference sub setting | | | Data in a desired format |
| 4. | **Development of classifiers**<br>Training the classifiers<br>Assigning weight to different features | | | A prototype |
| 5. | **Testing and validation of classifiers** | | | Tested methods |
| 6. | **final report write-up** | | | Final thesis |
| 7. | **Preparation for Research Defense** | | | Defense preparation |
| 8. | **Corrections after defense** | | | Corrected final report |
| 9. | **Submission** | | | Final submission |
| 10. | | | | |

Table 1: Timeframe and Milestones