Increasing the Efficiency in Non-Technical Losses Detection in Utility Companies

Juan I. Guerrero [#1], Carlos León[#] IEEE Member, Félix Biscarri [#], Iñigo Monedero [#], Jesús Biscarri[*], Rocío Millán[*]

[#] *Electronic Technology Department, School of Computer Science and Engineering, University of Seville*
Av. Reina Mercedes, S/N, 41012 Seville (Spain)
[1] juaguealo@us.es

[*] *Endesa, Automated Metering Management and Field Works Department*
Borbolla Building, Av. Borbolla, S/N, 41092 Seville (Spain)

***Abstract*— Usually, the fraud detection method in utility companies uses the consumption information, the economic activity, the geographic location, the active/reactive ration and the contracted power. This paper proposes a combined text mining and neural networks to increase the efficiency in Non-Technical Losses (NTLs) detection methods which was previously applied. This proposed framework proposes to collect all the information that normally cannot be treated with traditional methods. This framework is part of a research project. This project is done in collaboration with Endesa, one of the most important power distribution companies of Europe. Currently, the proposed framework is in the test stage and it uses real cases.**

I. Introduction

The NTLs represent the non-billed energy due to faults or illegal manipulations in the customers' facility. These provoke incoherences between the registered consumption by the utility company and the real consumption of the customer. This creates important economic losses for the utility company. The detection of NTLs uses a lot of techniques of artificial intelligence, standing out data mining, statistical techniques and the neural networks as methods with more successful in this research field.

Usually, the related papers with the detection of NTLs use features such as the customers' consumption, economic activity, geographic location, the contracted power and the active/reactive ratio. However, in the company databases there exists a lot of information: documentation, inspectors' commentaries, additional information about customers' facility, etc. This paper proposes an additional system to take advantage of the inspectors' knowledge. This allows the reduction of the false positives that could appear in the results of the studies. This knowledge is included in the databases as unstructured information or in natural language. This framework uses text mining, neural networks and statistical techniques to take advantage of the unstructured knowledge.

Endesa is the most important company of power distribution in Spain. It is one of the most important companies of Europe and it is present in South American markets. This company has more than 10 million customers. The databases of the utility companies store great quantities of information about the customers' facilities: documentation, inspectors' commentaries, explanations about the facility, etc. The studies described in [1] and [2] use statistical techniques, data mining and artificial intelligence to detect NTLs in Endesa's databases. This framework is used as part of the Decision Support System (DSS) which is tested with the results of applying the studies described in these references. These are tested with extracted real data of Endesa's databases.

A. Bibliographic revision

The text mining includes a set of techniques derived from data mining and other techniques related to artificial intelligence. The text mining performs the same function as data mining but over unstructured information. Nowadays, it is being used in a lot of research fields, such as: the content generation ([3]), the extraction of information ([4]), etc. In the present paper, the combination of statistical techniques, text mining, neural networks and the inspectors' knowledge are used for increasing the efficiency of campaigns or made studies.

At present, there are no references that use specific techniques to treat the additional information which can be found in the databases of companies. Usually, they are confined to use consumption data and limited information related to the customer.

In [5] a methodology which includes the use of Probabilistic Radial Basis Function (PRBFN) as a Probability Density Function (pdf), with the objective of detecting the normal consumption is applied. In order to do this, they use vectors of 12^{th} dimensions which allow getting the distribution of losses differentiated by voltage level (high, medium and low voltage) and consumption period. Reference [6] uses the customers with almost 25 months of measurements available. A Support Vector Machine (SVM) which uses as a main parameter, the diary average consumption to determine the NTLs, is used. Reference [7] steps up the previous method using genetic algorithms. As an additional feature they use customer's consumption data from electronic-Customer Information Billing System (e-CIBS). In another point of view, reference [8] uses wavelets due to its skill finding local behaviour. These references use the consumption feature as the main indicator of NTL.

Reference [9] proposes the use of data mining combining clustering techniques, classification (as NTL and NON-NTL) and prediction using the WEKA data mining tool. As a feature of the study it uses the customer's consumption and as detection indicator it uses the inspections. In this way, a

978-1-4244-5794-6/10/$26.00 ©2010 IEEE

supervised learning is applied. Reference [10] presents an interesting application of the Rough Set theory about customers using the consumption, the class and connection as conditional attributes and the existence of fraud as attribute of decisions for each customer or example. These references add new features to NTL detection. They emphasize through the use of new features as result of inspection and some other features.

The text mining techniques are used over a lot of research fields, emphasizing the web field which named Web Text Mining ([11]).

In the present paper an additional framework that can be used as an additional step in the process of NTL detection, increasing the efficiency is proposed. The proposed framework uses text mining, neural networks, statistical techniques and inspectors' knowledge to take advantage of more information about the consumers, without confining it only to customer's consumption. In order to do this, it uses the inspectors' knowledge over each one of the facilities of each customer to get a possible explanation to the anomalous consumption of the customer. The development of this framework is made by means of SPSS Clementine, with access to MySQL through ODBC drivers.

II. Concepts

When a customer wishes to contract the electrical supply he/she has to give certain information: personal information, rate, etc. The rate determines, among other features: the band time discrimination, which determines the price of the energy according to the corresponding period of time. Normally they are preestablished and depend on the demand; the billing frequency determines the period of time between one bill and another; and, the power range which is contractable.

Endesa, in low voltage, controls the customer consumption monthly or bimonthly according to the contracted power or power requested by the customer. This control is done by means of personnel who make appointments to take the measure directly from the equipment. In some cases, it is performed by means of telemeasure (the measure equipment has a MODEM which transmits the measure by Power Line Communications or PLC). The measurements by the personnel can include additional information about the customer's facility or observations done about the facilities. When the reading of customer's consumption cannot be taken the billing system of the company or the authorized personnel generate consumption estimation for the corresponding period of time. This estimation is named estimated measure or reading. However, when the estimation is not done the reading is named real. The period of time among measurements is named cycle or billing frequency.

Utility companies define the necessary procedures to make inspections. These procedures are designed to warrant which of the objectives of the inspection is completely done. An inspection is provoked by several reasons: reports, campaigns, suspected fraud, maintenance, cancelling one's membership, registering new customers, etc. There are previously preestablished steps for each one of these procedures. Normally, the inspectors register, in the company databases which actions have been done. In addition, they can register commentaries or observations. When there is the possibility of fraud, fault or administrative error, it must be corrected. In order to do this, a proceeding is started. It stores all the related information with the incident. It registers all the actions and observations done by the inspectors.

The reduction, localization and detection of NTLs in the utility companies are made by means of campaigns of massive inspections. The campaign is one of the most common. The campaign is a group of inspections which carries out a series of conditions, for example: customers which have a contracted power greater than 15 kW and they have three or more consecutive measures with zero consumption. The success of these campaigns is 12 % in the best scenario. These campaigns use a series of conditions to select the consumers for their inspection. These campaigns have a low percentage of success, due to very general conditions used. The proposed framework can be used in these cases to increase the efficiency of the campaigns.

In this paper, as a general reference, any commentaries of Endesa's staff and inspectors are named inspectors' commentaries or inspectors' knowledge.

III. Proposed Framework

The proposed framework needs a learning phase to collect available information and to establish relationships between different research fields.

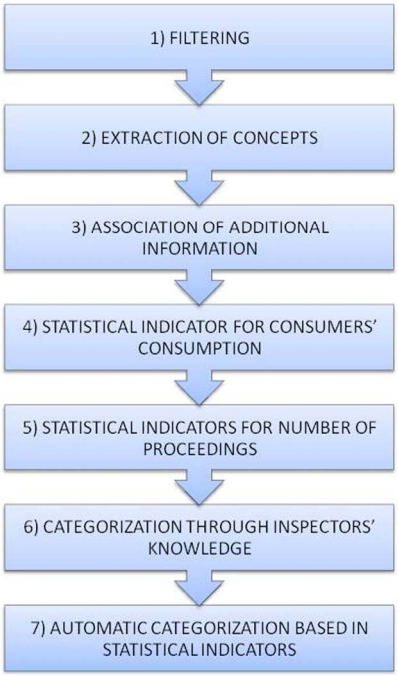

Fig. 1. Learning Process Diagram

A. Learning Process

The learning process is shown in figure 1. This process is done following these steps: 1) filtering of all proceedings or inspections which do not have useful information; 2) extraction of concepts; 3) association of date, source of the concept, and geographic location to a corresponding concept; 4) the consumption in the current, previous and next cycle are stored, in addition, the average, standard deviation, minimum and maximum of consumption are calculated; 5) the average, standard deviation, minimum, and maximum of number of proceedings are calculated; 6) application of inspectors' knowledge to classify them in categories; and, 7) creation of neural network model for automatic categorization of concepts.

B. Testing Process

The testing process of the framework is made with the following steps: 1) introduction of information about the customers to analyse; 2) extraction of concepts, this is the same process that is applied in the learning process, but only applied to the proposed customers to analyse; 3) comparison between learned concepts and extracted concepts; 4) classification in corresponding category; and, 5) results and conclusions. The diagram of this process is shown in figure 2.

Steps 3 and 4 are done using all the categorized concepts: the concepts categorized by inspectors' knowledge and the concepts categorized by the neural network.

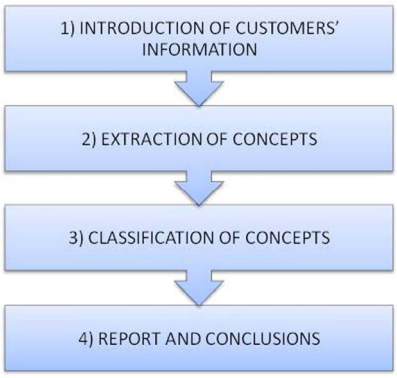

Fig. 2. Testing Process Diagram

IV. Learning Process

The learning process has to use the highest possible quantity of customers. This way, a representative statistical set for each concept is obtained.

A. Extraction of Concepts

A concept can be made up of one or more word. The concept can be a syntagm or a word which represents an entity (action, event, etc.). The process of extraction of concepts is based on the utilization of Natural Language Processing (NLP). NLP is a set of techniques that allows translation between the natural and formal language. This is essential to allow the processing of this kind of information. An interesting review about this technique and utilization in Information System Management is proposed by [12]. In addition, this technology includes the utilization of a series of techniques of normalization which are applied in the extracted concepts: recognition of the punctuation errors to avoid errors which include the inadequate use of punctuation symbols (the tilde, the period, the comma, the point and comma, the dividing bar, etc.); recognition of spelling errors by means of a grouping fuzzy technology to avoid spelling errors (interchange of letters, recurrence words, lack of letters, etc.); utilization of dictionaries to establish the equivalence between synonyms; and, recognition of non-linguistic terms (dates, hours, measures, etc.). Dictionaries are the main tool and it must be configured and updated manually.

B. Adding Information to the Concepts

The information related with the concepts compounds: the date in which the comment is made, the geographic location, and the source of commentary. The date is necessary to establish temporal patterns of appearance of concepts, for example, holiday periods, etc. The geographic location is important to determine guidelines of local behaviour. The source of the commentary allows determining if the commentary is provided by an inspector, staff measuring, etc. This information field determines the concept reliability.

C. Statistics of consumption and number of proceedings

The consumption of the next, current and previous cycles are used in learning process. In addition, the statistic is calculated over the consumption and the number of proceedings. This statistics allow determining: what is the related consumption with a certain concept; and, what is the risk of existence of NTL. In order to do this, each concept of each customer is associated to the consumption of the corresponding cycle when the commentary is typed. Also, the consumption of the next and previous cycle is associated. The average, the maximum, the minimum, and the standard deviation of consumption associated to each concept are calculated in parallel. Following, the average, the maximum, the minimum, and the standard deviation of number of proceedings for each concept of each customer are calculated.

D. Categorization of Concepts by Means of Application of Inspectors' Knowledge

The inspectors' knowledge is used to classify additional information about the customer. The concept is classified in one of the following categories: *information about the consumer's facility*, *consumption characterization*, and *correction of customer's economic activity*. This classification is done through the utilization of rules which implement this knowledge. It is important to extract the concepts process described in *A* section, beforehand because this process applies normalization tools. These tools make the interpretation of natural language easier. In the same way, these tools make the application of rules more efficient. Rules have the IF-THEN structure. The antecedent of these rules is made up of searching conditions of normalized concepts. These rules were created manually.

The *information about customer's facility* category classifies concepts directly related with features of facilities. For example: the existence of a condenser battery, the location of measuring equipment, the possibility of measurement problems, etc. This category has 12 rules that identify certain information about the customer facilities.

The *consumption characterization* category classifies concepts directly related to the consumer's consumption pattern. There are different concepts which identify consumers with null, sporadic, periodical, and seasonal consumption. This category is necessary because there are consumers who have several contracts and only use one of them, due to the fact that the rest are auxiliaries or fire fighting equipment. This category has several subcategories according to the expected consumption suggested by the inspectors. This category has one rule for each sub-category (9 rules). The subcategories are: *seasonal consumption, domestic consumption, periodical consumption, sporadic consumption, zero consumption, low consumption, high reactive range* and *high consumption*.

The *correction of customer's economic activity* category groups the concepts in 95 sub-categories, each of them identifies a different economic activity. This category has one rule for each sub-category (95 rules). When a customer contracts the utility service, the customer has to specify their economic activity but sometimes is not notified or is wrong specified. This feature is very important because it is used in great quantities of studies (e.g. [5], [13]). Normally, the inspectors add information about the economic activity in their commentaries to explain the anomalous consumption patterns. This information can be used to update and to report to the company staff. This report allows the update of customer information and to make new studies.

In this point of view, the proposed framework is used as a Decision Support System (DSS). This framework helps fix and mend the NTL detection techniques used in [1] and [2].

E. Automatic Categorization of Concepts

The previous process of categorization concepts is very difficult, because it must be done manually. The previous method uses the most important and common concepts. The automatic categorization is another method to classify less outstanding concepts. The neural network assigns a trust value for each concept classification depending on initial features. This trust value and the concept are the most important features to make a decision. Additionally, this information is shown in the final report.

The automatic categorization allows classifying the concepts dependent on all the calculated features. This method is used over the rest of the concepts that cannot be classified through inspectors' knowledge (previous step). The implemented categorization in the proposed framework combines neural networks through the inspectors' knowledge. The proposed technique uses the concepts classified through the inspectors' knowledge to make a model to classify the new concepts. In this way, the automatic concept classification stores the pattern of selected features of each category. The features used in this technique are: source of concept, corresponding consumption cycle, consumption of previous cycle, consumption of next cycle, statistic indicators of number of proceedings and consumption, date of the concept, frequency data of the concept appearance, number of real measures, band time discrimination, and number of estimated measures.

The neural network is trained by means of a multiple method. This method creates several neural networks of different topologies depending on the training data. At the end of training, the model with the lowest Root Mean Square (RMS) error is presented as the final neural network. This training method is done by SPSS Clementine. The trained neural network assigns an importance value to each feature. SofMax transfer function ([14]) was used as a punctuation method. The trained neural network has two hidden layers and its structure is 22-28-26-4, due to the quantity of inputs. The first and last layers are the input and output layers, respectively.

V. TESTING PROCESS

The testing process allows determining whether the learning process has worked successfully. For this process a set of different consumers is used. The testing procedure starts with the introduction of all the necessary information, so the data of the unstructured information, consumption, and proceedings fields is collected. The unstructured information fields use process of extraction of concepts, in the same way as in the learning process. The previously trained model and the inspectors' knowledge are applied to determine whether there would be an NTL. The results are shown in a report. This report shows a summary of the found information and the obtained conclusions. The inspector can use this report as part of an NTL detection process. In this sense, this framework is used as part of a Decision Support System (DSS) which currently is in a testing process in Endesa. The report shows the following information: basic information of consumers' identification; classification according to the method of automatic learning; classification according to the inspectors' knowledge, showing information about the localization and state of the measuring equipment, specification of possible errors in economic activity, existence of condenser battery, etc.; and, obtained conclusions. The conclusion determines whether the consumer has a problem.

The testing process is made using several customers of different features.

VI. EXPERIMENTAL RESULTS

The framework is applied over a sample of 300 customers with a possible NTL. This sample is the result of applying the described methods in [1] and [2] over a sample of 51040 customers in Andalusia (Spain). These customers have a contracted power greater than 15 kW.

The framework detected 142 customers as false positive. Although these consumers have an anomalous consumption pattern, the inspectors' commentaries explain what could be the explanation for this anomalous consumption. Also, the framework confirmed 74 customers with some problem or

NTL. The rest of the customers haven't enough or any inspectors' commentaries or unstructured information. In figure 4 this process of experimental test is shown.

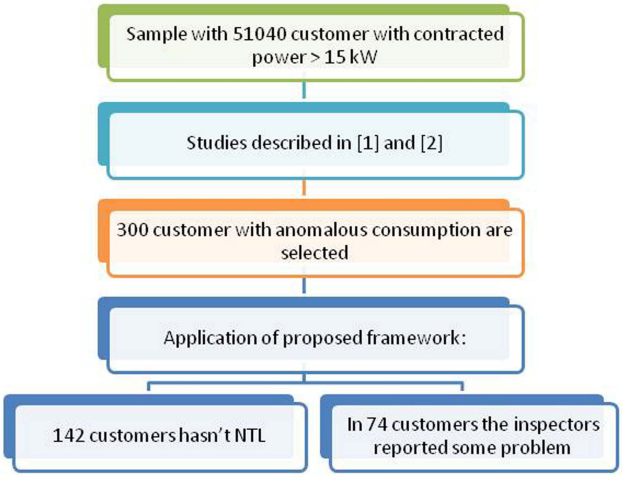

Fig. 3. Process of experimental test.

The customers selected as false positives (142 customers) are determined in the most restrictive way possible. Inspectors' commentaries of each customer can explain the source of an anomalous consumption. The correct identification of each concept is the main problem, because inspectors' commentaries are written in natural language.

The customers selected with some problem (74 customers) are determined in the most restrictive way, too. Inspectors' commentaries of each customer can determine some problem. Mainly, these problems consist of three types: 1) a problem which can be solved without inspection; 2) a problem which is in process of solution; and, 3) an NTL.

Below four specific cases are shown: three cases of false positives and one more that has a measurement problem.

The first case is a consumer of cold-pressed olive oil industry. The consumption of this type of customers is seasonal and is strongly dependent on environmental conditions. This customer had a series of inspectors' commentaries. These commentaries specified that there wasn't any activity during the months with anomalous consumption. In the graph of figure 4 the consumption curve is shown. An increasing of consumption in the beginning of each year can be seen. In 2008 there is a great consumption due to an increase of activity.

The second case presents a null consumption. However, the inspectors' commentaries could determine that this customer only has a motor for an irrigating activity. The consumption in this activity used to have very sporadic consumption. The graph consumption of the figure 5 shows that the customer's consumption is very irregular. But it is a normal behavior, because the land is irrigated more in hot weather or dry months.

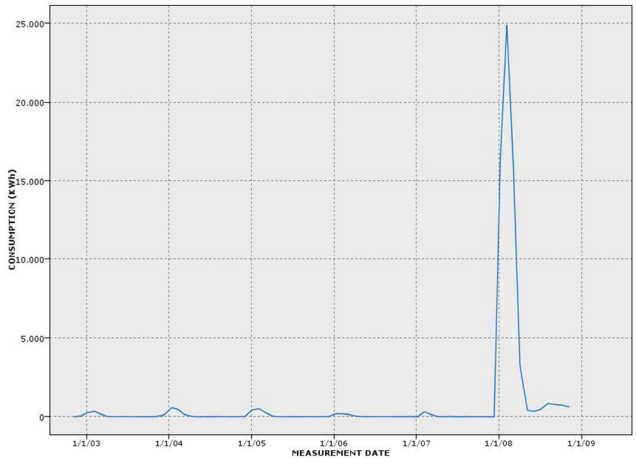

Fig. 4. Cold-pressed olive oil industry consumption graph

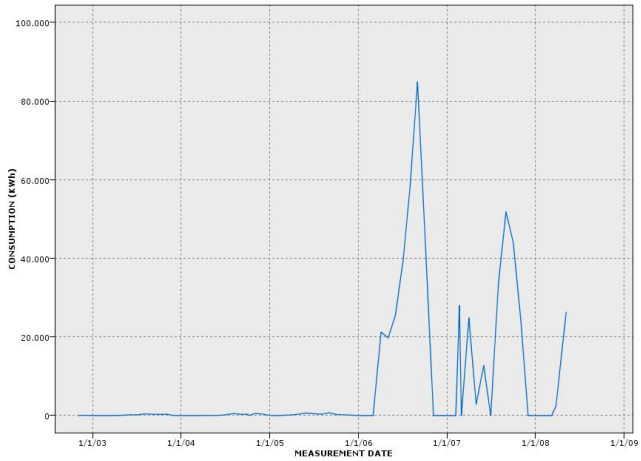

Fig. 5. Irrigating activity consumption graph

The third case is a typical anomalous consumption. In this case, the consumption is always at the same level but, in one or three months, the consumption decreases or, sometimes, is zero. This customer has several inspectors' commentaries. These commentaries specified that there is no activity due to a holiday period. The consumption graph in figure 6 shows this case. This customer has zero consumption in august of 2008 due to a holiday period. This period was notified by staff measuring.

The fourth case is the case of a learned concept by the neural network. The concept has a 100 % trust value that there was a measurement problem. This customer has several inspectors' commentaries. They specified that the staff cannot take the measure due to a padlocked register. This is an anomaly that could represent an NTL in a large time period. The measurement problem is very important because it can hide a fraud. The consumption graph of this customer is a line with zero consumption. This case is inside a set of 74 customers with some problems, previously mentioned.

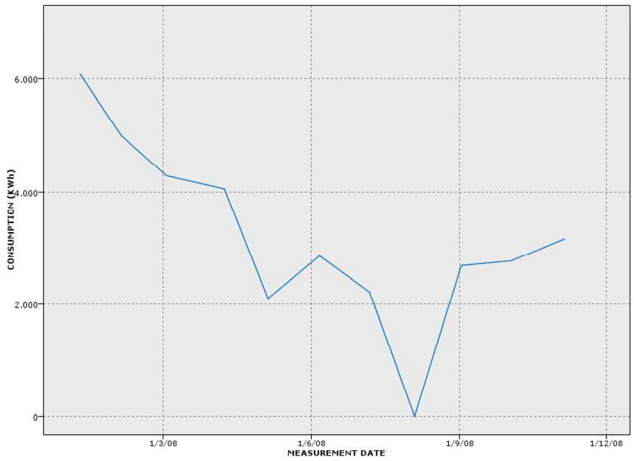

Fig. 6. Holiday period case consumption graph

VII. CONCLUSION

The proposed framework makes its main contribution in the research field: detection of NTLs in utility companies. Usually, the research about this field leads to treatment of customers' consumption and some other features. The proposed framework increases the kind of information used in detection of NTLs and, in the same way, the used techniques. This framework, as a part of a DSS, is applied after the studies of [1] and [2]. This DSS increases the efficiency. This DSS assigns a new classification, having in mind other features. These frameworks are applied to real cases of Endesa's databases.

The application of this framework filtered 47.3 % of customers with no actual has problems. Additionally, they confirm that there are 24.6 % of customers that may have some problems. The rest of customers did not have enough information from inspectors. In this way, the efficiency of campaign is increased due to the reduction of the selected customers, because only 158 customers will be reported to inspectors. In these 158 customers, there are 74 customers (24.6 %) that the inspectors' commentaries specified a series of problems. This information can be added to the report for inspectors.

The successful application of this framework depends on the information available about the consumers. It is necessary that customers' information about the inspections and documentation be available; if the customer hasn't any information, this technique won't have any effect over the conclusion about the customer.

ACKNOWLEDGMENT

We would like to thank the initiative and collaboration of Endesa, in particular Francisco Godoy, Gema Tejedor and the Endesa's inspectors that make the field inspections.

REFERENCES

[1] F. Biscarri, I. Monedero, C. León, Juan I. Guerrero, J. Biscarri, and R. Millán, *A data mining method based on the variability of the customers consumption*, 10th International Conference on Enterprise Information Systems, ICEIS 2008, June 12-16, Barcelona, Spain.

[2] F. Biscarri, I. Monedero, C. León, Juan I. Guerrero, J. Biscarri, and Rocío Millán, *A mining Framework to detect non-technical losses in power utilities*, 11th International Conference on Enterprise Information Systems, ICEIS2009, May 6-10, Milano, Italy.

[3] Hsin-Chang Yang and Chung-Hong Lee, *A text mining approach on automatic generation of web directories and hierarchies*, Expert Systems with Applications 27, pp. 645-663, July 2004.

[4] Nahk Hyun Sung and Yong Sik Chang, *Business information extraction from semi-structured webpages*, Expert Systems with Applications 26, pp. 575-582, December 2003.

[5] J.R. Galván, A. Elices, A. Muñoz, T. Czernichow, and M.A. Sanz-Bobi, *System for detection of abnormalities in fraud in customer consumption*, 12th Conference on the Electric Power Supply Industry, November 2-6, 1998, Pattaya, Thailand.

[6] J. Nagi, A.M. Mohammad, K.S. Yap, S.K. Tiong, and S.K. Ahmed, *Non-technical loss analysis for detection of electricity theft using Support Vector Machines*, 2nd IEEE International Conference on Power and Energy (PECon 08), December 1-3, 2008, Johor Bahru, Malaysia.

[7] J. Nagi, K. S. Yap, S.K. Tiong, S. K. Ahmed, and A.M. Mohammad, *Detection of abnormalities and electricity theft using genetic Support Vector Machines*, TENCON 2008, IEEE Region 10 Conference. 19-21 Nov. 2008, pp. 1-6.

[8] Rong Jiang, Harry Tagaris, Andrei Lachsz and Mark Jeffrey, *Wavelet based feature extraction and multiple classifiers for electricity fraud detection*, IEEE/PES Transmission and Distribution Conference and Exhibition 2002: Asia Pacific. 6-10 Oct. 2002. pp. 2251-2256 vol. 3.

[9] Anisah H. Nizar, Zhao Yang Dong, and Pei Zhang, *Detection rules for non-technical losses analysis in power utilities*, 2008 IEEE Power and Energy Society General Meeting – Conversion and Delivery of electrical Energy in the 21st Century. 20-24 July 2008, pp. 1-8.

[10] José E. Cabral, João Onofre P. Pinto, Edgar M. Gontijo, and José Reis Filho, *Fraud detection in electrical energy consumers using Rough Sets*, 2004 IEEE International Conference on Systems, Man and Cybernetics. 10-13 Oct. 2004. pp. 3625-3629 vol. 4.

[11] Shinqun Yin, Gang Wang, Yuhui Qiu and Weiqun Zhang, *Research and implement of classification algorithm on web text mining*, Third International Conference on Semantics, Knowledge and Grid. 29-31 Oct. 2007. pp. 446-449.

[12] Elisabeth Métais, *Enhancing information systems management with natural language processing techniques*, Data & Knowledge Engineering 41, pp.247-272, 2002

[13] David J. Hand and Gordon Blunt, *Prospecting for gems in credit card data*, IMA Journal of Management Mathematics Vol. 12 (No. 2), pp. 173-200, 2001.

[14] N. Ahmed and M. Campbell, *Multimodal operator decision models*, 2008 American Control Conference, 2008 AACC. Westin Seattle Hotel, Seattle, USA. June 11-13, 2008. pp. 4504-4509.