2017 IEEE 24th International Conference on Web Services

Fuzzy QoS Modeling of IT Maintenance Tickets

Suman Roy*, Durga Prasad Muni*, Adrija Bhattacharya†
, Dipanjan Dutta* and Navin Budhiraja*
*Infosys Ltd.,Bangalore 560 100, India
{Suman_Roy, DurgaPrasad_Muni, Dipanjan.Dutta, Navin.Budhiraja }@infosys.com
†Dept of Computer Science and Engineering, University of Calcutta, Kolkata, India
adrija.bhattacharya@gmail.com

***Abstract*—Ticketing system is an example of a Service System (SS) which is responsible for handling huge volumes of tickets generated by large enterprise IT (Information Technology) infrastructure components, and ensuring smooth operation. An issue is captured as summary on the ticket and once a ticket is resolved, the solution is also noted down on the ticket as resolution. Further the system maintains the provision of recording the time when a ticket is opened, acknowledged to user, resolved and/or closed, from which different QoS parameters could be obtained. For example, Resolution Time can be computed as the difference of resolution date and opening date of the ticket. QoS parameters are used to measure the performance of different aspects of a service. In case of impreciseness of observations of these parameters fuzzy sets seems to be an optimal tool to model them. To ensure better operation for services based on these QoS values we propose a two-stage analysis framework for QoS prediction of incoming tickets which includes fuzzy clustering of incident tickets based on QoS values and building a fuzzy regression model using this categorization and the textual contents of tickets. Further we carry out a fuzzy correlation analysis of different categories (clusters) of QoS parameters. Lastly we report on our experimental results.**

***Keywords*-Quality of servies (QoS), service system, tickets, fuzzy set, Response Time, Resolution Time, fuzzy clustering, Fuzzy C-means algorithm (FCM), non-negative least squares (NNLS), fuzzy correlation.**

I. Introduction

Ticketing system, an example of SS, is used as one of the inputs for Information Technology Infrastructure Library (ITIL) services such as problem management and configuration management. Dozens of tickets are raised by the users on the ticketing system for the purpose of resolving their problems while using different support systems. Normally a ticket is recorded on the system with a summary which describes the related issue. The maintenance team examines the ticket, responds to the user with an acknowledgment with possibly seeking more explanation of the issue and indicating the possible solution and subsequently, undertakes remedial actions on the ticket, after writing a short note on the resolution steps and closes the ticket in the end. The maintenance team also records other details about the ticket like, the opening date, the response date, the resolution date and the closing date for the ticket. These parameters could be used to compute different Quality of service (QoS) parameters like Response Time (difference between response date and open date), Resolution Time (difference between resolution date and open date), Closure Time (difference between closure date and open date) etc.

It makes sense to model different QoS parameters arising in ticketing application services using Fuzzy sets instead of normal (crisp) sets, because Resolution Time for example, could be different from person to person handling the same ticket. Also Resolution Time would depend on the description of the summary of the ticket written by different users although the issue may be more or less the same. These QoS parameters are intrinsically fuzzy. Therefore, fuzzy sets are an optimal tool for modeling these imprecise parameters. In line with that, a more appropriate way to describe the Resolution Time for a new ticket is to say that it is high, or medium, or low. These phrases 'high', 'medium' and 'low' can be regarded as fuzzy quantities. This kind of categorization of ticketing services also helps people especially those with almost no technical experience, to understand reasonable QoS values as provided by the service.

Normally, statistical techniques are employed for estimation and inference which is based on the computation of deviations between the observed values and the estimated values, which are supposed to occur due to measurement errors or random variations. In QoS analysis these variations may happen due to imprecise observed data or the indefiniteness of the system structure. As the indefiniteness is due to fuzziness rather than randomness, regression analysis on fuzzy data is done with techniques of fuzzy regression. Further heterogeneity of observations is a common phenomenon in ticketing systems, *e.g.,* Resolution Time may vary from 1 hour (less than a day) to 300 days. It is difficult to handle heterogeneity problem in regression analysis. The heterogeneity in the data may happen because of the hidden presence of different clusters of observations. If we categorize the tickets (based on their QoS values) into clusters and use their class membership as the dependent variable and construct suitable feature vectors for tickets based on their textual summaries to be used as independent variable then linear regression enables us to overcome to the

978-1-5386-0752-7/17 $31.00 © 2017 IEEE
DOI 10.1109/ICWS.2017.55

heterogeneous problem in model fitting.

Motivated by this we use a two-stage procedure for predicting QoS values for new tickets. First, we choose Fuzzy C-means algorithm to cluster ticket data using their different QoS values and compute the membership values of tickets with respect to each cluster. Then we use these membership values and feature vectors of tickets with TF*IDF values to set up a linear regression model. This model helps us predict QoS values for incoming tickets. Like correlation of crisp sets where the joint relationship of two variables is examined by measuring the interdependence of the two variables, one can also study fuzzy correlation between fuzzy data. In our case, we can treat each cluster (category) for each QoS as a fuzzy set with membership values for tickets and find out the correlation between categories using fuzzy correlation.

II. A primer on Fuzzy sets

Fuzzy sets were introduced by Lotfi A. Zadeh [1] in 1965 as an extension of the classical sets. Fuzzy set theory allows membership assessment of an element described with the aid of a membership function valued in the real interval $[0.1]$. Thus this set theory can be seen as a generalization of the popular set theory since the characteristic (indicator) functions of the classical sets are special cases of membership functions of fuzzy sets. In fuzzy set theory, classical sets with bivalent membership of elements are called crisp sets.

Let S be a non-empty set. A fuzzy set A in S is characterized by its membership function $\mu_A : S \to [0, 1]$. It is easy to see that A is completely determined by the set of tuples: $A = \{(z, \mu_A(z)) \mid z \in S\}$.

If $S = \{z_1, \ldots, z_n\}$ is a finite set and A is a (discrete) fuzzy set in S then the following notation is often used

$$A = \mu_1/z_1 + \cdots \mu_n/z_n,$$

where the term $\mu_i/z_i, i = 1, \ldots, n$ signifies the grade of membership of z_i in S and '+' denotes the union. For an elaborate exposition on fuzzy sets the reader is advised to consult [2].

III. Schema of ticket data

We consider incident tickets with similar schema which are frequent in IT maintenance. These tickets usually consist of two fields, fixed and free form. There is no standard value for free-form fields. These fields can contain call description or summary of ticket which capture the issue behind the creation of such tickets on the maintenance system. This can be just a sentence that summarizes the problem reported in it, or it may contain a short text describing the incident. Fixed fields are customized and inserted in a menu-driven fashion. Example of such fields are the ticket's identifier, employee number of the user raising the ticket etc, priority of the ticket, the time the ticket is raised, responded to or closed on the system or, if a ticket is of incident or request in nature. Various other important information are also captured through these fixed fields such as category of a ticket, sub-category, application name (application under which the ticket was raised), incident type etc. A small part of ticket data is shown in Figure 1.

A. Feature vector creation from Ticket Data

We assume a free field of a ticket to contain a succinct problem description associated with it in the form of a summary (or call description). We extract features from the collection of summaries of tickets using light weight natural language processing. As a pre-processing we remove the tickets which do not contain a summary. In the beginning we lemmatize the words in the summary of tickets. Then we use Stanford NLP tool to parse the useful contents in the summary of the tickets and tag them as tokens. Next we set up some rules for removing tokens which are stop words. We compute document frequency (DF)[1] of each lemmatized word. We discard any word whose DF is smaller than 3. By discarding words with DF less than 3, we can remove the rare words (like the name of persons raising the tickets) which do not contribute to the content of ticket summary. In this way, the feature vector size could be reduced significantly. We perform some other pre-processing and select bi-grams and tri-grams as terms (keyphrases), the details of which are described in [4]. Further specific to this work, we perform PCA (Principal Component Analysis) on the ticket-term TF*IDF matrix for reducing the feature vectors of tickets.

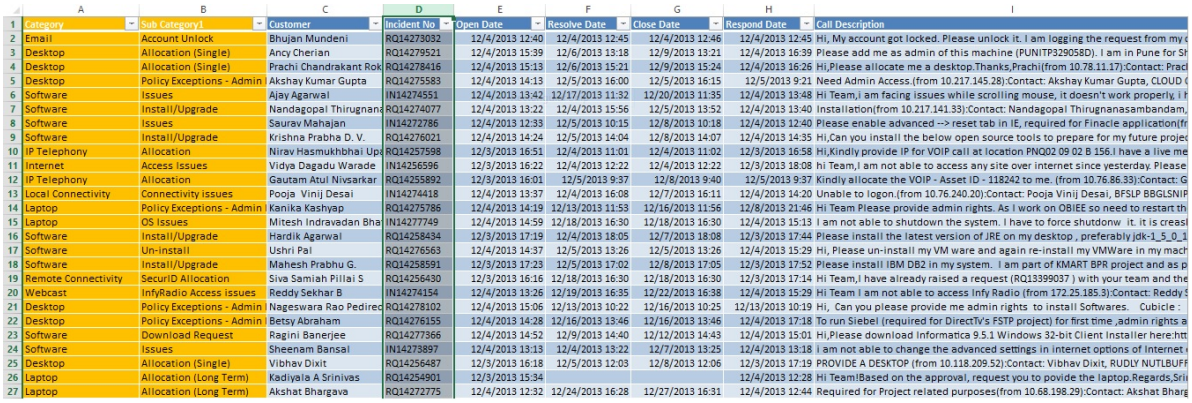
We model a ticket T as a row vector $T(x_1 \ \ldots \ x_p)$, where each element x_i represents the importance or the weight of a term/keyphrase w_i with respect to the ticket and each feature vector contains p keywords. We shall use TF*IDF of a word as its weight[2].

B. Grouping the tickets *wrt* tuples

The fixed field entries of a ticket can be represented using a relational schema. For that we shall consider only a limited number of fixed fields of a ticket for choosing attributes that reflect its main characteristics (the domain experts' comments play an important role in choosing the fixed fields), for example the attributes can be, - application name, category and sub-category. They can be represented as a tuple: Ticket(application name, category and sub-category). Each of the instances of tuples corresponding to entries in the fixed fields in the ticket can be thought of an instantiation of the schema. Examples of rows of such schema can be, (AS400 - Legacy Manufacturing, Software, Application

[1]Document frequency of a word is the number of tickets (ticket summaries) containing the word in the data set (corpus) [3]

[2]TF*IDF is a popular metric in the data mining literature [3]



	A	B	C	D	E	F	G	H	I
1	Category	Sub Category1	Customer	Incident No	Open Date	Resolve Date	Close Date	Respond Date	Call Description
2	Email	Account Unlock	Bhujan Mundeni	RQ14273032	12/4/2013 12:40	12/4/2013 12:45	12/4/2013 12:46	12/4/2013 12:45	Hi, My account got locked. Please unlock it. I am logging the request from my
3	Desktop	Allocation (Single)	Ancy Cherian	RQ14279521	12/4/2013 15:39	12/6/2013 13:18	12/9/2013 13:21	12/4/2013 16:39	Please add me as admin of this machine (PUNITP329058D). I am in Pune for SI
4	Desktop	Allocation (Single)	Prachi Chandrakant Rok	RQ14278416	12/4/2013 15:13	12/6/2013 15:21	12/9/2013 15:24	12/4/2013 16:26	Hi,Please allocate me a desktop.Thanks,Prachi(from 10.78.11.17):Contact: Pracl
5	Desktop	Policy Exceptions - Admin	Akshay Kumar Gupta	RQ14275583	12/4/2013 14:13	12/5/2013 16:00	12/5/2013 16:15	12/5/2013 9:21	Need Admin Access.(from 10.217.145.28):Contact: Akshay Kumar Gupta, CLOUD (
6	Software	Issues	Ajay Agarwal	IN14274551	12/4/2013 13:42	12/17/2013 11:32	12/20/2013 11:35	12/4/2013 13:48	Hi Team,I am facing issues while scrolling mouse, it doesn't work properly. i h
7	Software	Install/Upgrade	Nandagopal Thirugnana	RQ14274077	12/4/2013 13:22	12/4/2013 15:56	12/5/2013 13:52	12/4/2013 13:40	Installation(from 10.217.141.33):Contact: Nandagopal Thirugnanasambandam,
8	Software	Issues	Saurav Mahajan	IN14272786	12/4/2013 12:33	12/5/2013 10:15	12/8/2013 10:18	12/4/2013 12:40	Please enable advanced --> reset tab in IE, required for Finacle application(fr
9	Software	Install/Upgrade	Krishna Prabha D. V.	RQ14276021	12/4/2013 14:24	12/5/2013 14:04	12/8/2013 14:07	12/4/2013 14:35	Hi,Can you install the below open source tools to prepare for my future projec
10	IP Telephony	Allocation	Nirav Hasmukhbhai Upa	RQ14257598	12/3/2013 16:51	12/4/2013 11:01	12/4/2013 11:02	12/3/2013 16:58	Hi,Kindly provide IP for VOIP call at location PNQ02 09 02 B 156.I have a live me
11	Internet	Access Issues	Vidya Dagadu Warade	IN14256596	12/3/2013 16:22	12/4/2013 12:22	12/4/2013 12:22	12/3/2013 18:08	hi Team,I am not able to access any site over internet since yesterday. Please
12	IP Telephony	Allocation	Gautam Atul Nivsarkar	RQ14255892	12/3/2013 16:01	12/5/2013 9:37	12/8/2013 9:40	12/5/2013 9:37	Kindly allocate the VOIP - Asset ID - 118242 to me. (from 10.76.86.33):Contact: G
13	Local Connectivity	Connectivity issues	Pooja Vinij Desai	IN14274418	12/4/2013 13:37	12/4/2013 16:08	12/7/2013 16:11	12/4/2013 14:20	Unable to logon.(from 10.76.240.20):Contact: Pooja Vinij Desai, BFSLP BBGLSNIP
14	Laptop	Policy Exceptions - Admin	Kanika Kashyap	RQ14275786	12/4/2013 14:19	12/13/2013 11:53	12/16/2013 11:56	12/8/2013 21:46	Hi Team Please provide admin rights. As I work on OBIEE so need to restart th
15	Laptop	OS Issues	Mitesh Indravadan Bha	IN14277749	12/4/2013 14:59	12/18/2013 16:30	12/18/2013 16:30	12/4/2013 15:13	I am not able to shutdown the system. I have to force shutdonw it. it is creasl
16	Software	Install/Upgrade	Hardik Agarwal	RQ14258434	12/3/2013 17:19	12/4/2013 18:05	12/7/2013 18:08	12/3/2013 17:44	Please install the latest version of JRE on my desktop , preferably jdk-1_5_0_1
17	Software	Un-install	Ushri Pal	RQ14276563	12/4/2013 14:37	12/5/2013 13:26	12/5/2013 13:26	12/4/2013 15:29	Hi, Please un-install my VM ware and again re-install my VMWare in my mach
18	Software	Install/Upgrade	Mahesh Prabhu G.	RQ14258591	12/3/2013 17:23	12/5/2013 17:02	12/8/2013 17:05	12/3/2013 17:52	Please install IBM DB2 in my system. I am part of KMART BPR project and as p
19	Remote Connectivity	SecurID Allocation	Siva Samiah Pillai S	RQ14256430	12/3/2013 16:16	12/18/2013 16:30	12/18/2013 16:30	12/3/2013 17:14	Hi Team,I have already raised a request (RQ13399037) with your team and the
20	Webcast	InfyRadio Access issues	Reddy Sekhar B	IN14274154	12/4/2013 13:26	12/19/2013 16:35	12/22/2013 16:38	12/4/2013 15:29	Hi Team I am not able to access Infy Radio (from 172.25.185.3):Contact: Reddy S
21	Desktop	Policy Exceptions - Admin	Nageswara Rao Pedirec	RQ14278102	12/4/2013 15:06	12/13/2013 10:22	12/16/2013 10:25	12/13/2013 10:19	Hi, Can you please provide me admin rights to install Softwares. Cubicle :
22	Desktop	Policy Exceptions - Admin	Betsy Abraham	RQ14276155	12/4/2013 14:28	12/16/2013 13:46	12/16/2013 13:46	12/4/2013 17:18	To run Siebel (required for DirectTv's FSTP project) for first time ,admin rights a
23	Software	Download Request	Ragini Banerjee	RQ14277366	12/4/2013 14:52	12/9/2013 14:40	12/12/2013 14:43	12/4/2013 15:01	Hi,Please download Informatica 9.5.1 Windows 32-bit Client Installer here:htt
24	Software	Issues	Sheenam Bansal	IN14273897	12/4/2013 13:13	12/4/2013 13:22	12/7/2013 13:25	12/4/2013 13:18	I am not able to change the advanced settings in internet options of Internet
25	Desktop	Allocation (Single)	Vibhav Dixit	RQ14256487	12/3/2013 16:18	12/5/2013 12:03	12/8/2013 12:06	12/3/2013 17:19	PROVIDE A DESKTOP (from 10.118.209.52):Contact: Vibhav Dixit, RUDLY NUTLBUFF
26	Laptop	Allocation (Long Term)	Kadiyala A Srinivas	RQ14254901	12/3/2013 15:34			12/4/2013 12:28	Hi Team!Based on the approval, request you to povide the laptop.Regards,Srii
27	Laptop	Allocation (Long Term)	Akshat Bhargava	RQ14272775	12/4/2013 12:32	12/24/2013 16:28	12/27/2013 16:31	12/4/2013 12:44	Required for Project related purposes(from 10.68.198.29):Contact: Akshat Bharg

Figure 1. A sample of ticket data

Errors), (AS400 Legacy - Retail, Software, Application Functionality Issue) etc. The relation key can vary from 1 to number of distinct tuples in the schema. One such key can hold several Incident IDs, that is, it can contain several tickets with different IDs. This way we can partition the tickets using their tuples.

C. Possible QoS parameters for ticket data

Given this ticket schema one can formulate the following QoS paramaters, Response Time $\mathcal{R}$, Resolution Time $\mathcal{S}$, and Closure time $\mathcal{C}$. As the open, close, response and resolve dates are recorded for each ticket we can set the following parameters: *Response Time* $\mathcal{R}$ as the difference between response date and open date, *Resolution Time* $\mathcal{S}$ as the difference between resolve date and open date and *Closure time* $\mathcal{C}$ as the difference between closure date and open date etc. All these parameters are calculated in days. Note resolve date and closure date can be same for some of the tickets, however in some cases it is different. In these situations the system administrators need to check other issues before closing the tickets although resolutions have been already provided for them earlier.

IV. Clustering tickets based on QoS parameters

Instead of clustering tickets based on the actual value of QoS parameters (with a suitable distance metric like "l_1" norm) we prefer to categorize tickets based on fuzzy procedures for facilitating inference mechanism generated from either expert knowledge or data. Also fuzzy clustering provides a more suitable partition of the set of objects than crisp clustering do. For this purpose we adopt the well-known fuzzy clustering method, *viz,* Fuzzy C-Means (FCM) [5] which groups a set of data instances into a number of different clusters. The algorithm assigns each data point to each cluster with a membership degree, that is, each data point may belong to more than one cluster with different degrees of membership. Let $\{z_1, z_2, ..., z_n\}$ be the set of n data points and κ be the number of clusters. FCM clusters the data points into κ fuzzy groups and iteratively finds a cluster center in each group following minimization of an objective function. The details of the steps of FCM are described in [6], [5]. At any iteration one can compute the degree of membership, μ_{ij}, of each jth data point z_j *wrt* cluster center $c_i, 1 \leq i \leq \kappa$ as

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{\kappa} \left(\frac{d_{ij}}{d_{kj}}\right)^{2/m-1}}, \tag{1}$$

where $d_{ij} = ||c_i - z_j||$. Here m is the index of fuzziness. Typically the value of m is taken as 2. The cluster centers are updated as:

$$c_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m z_j}{\sum_{j=1}^{n} \mu_{ij}^m} \tag{2}$$

To evaluate obtained clusters, there are various cluster validity methods [7]. One such method uses the Xie-Beni (XB) validity index [5] that uses membership values and the data set to evaluate fuzzy clusters. Lower the value of V_{XB}, better the clustering solution is. The optimal number of clusters can be determined by taking different number of clusters and then by computing the validity index.

V. A Regression model for predicting QoS paramaters for tickets

We shall use a linear model for multiple regression analysis for predicting QoS parameters. As in our case the fuzzy memberships are defined on discrete sets, we choose to work with different linear regression models for these individual discrete fuzzy sets. Other fuzzy linear regression models (*e.g.*, model of Tanaka *et al* [8]) cannot used in this case since they assume fuzzy sets on continuous domain in the form of triangular/trapezoidal fuzzy numbers or fuzzy numbers with α-level set defined on real numbers.

A. Multiple regression analysis

We sketch the method of solving the problem of estimating the regression parameters using a collection of

sample data. For that we consider a sample of n data points, observing values $\tilde{y}_1, \ldots, \tilde{y}_n$ of the response variables and $\{(x_{1,1}, \ldots, x_{1,p}), \ldots, (x_{n,1}, \ldots, x_{n,p})\}$ of the predictor variables, where $x_{i,j}$ denotes the jth predictor variable measured for the ith observation. We choose $b_0, b_1, \ldots, b_p$ as regression parameters. For each observed response y_i, with corresponding predictor variables $x_{i,1}, \ldots, x_{i,p}$, (all non-negative) $1 \leq i \leq n$, we would like to fit the following equation to predict the response variable $\tilde{y}_i$:

$$\tilde{\mathbf{y}} = \mathbf{X}\mathbf{b} \tag{3}$$

where $\tilde{\mathbf{y}} = [y_1 \cdots y_n]^T$, also

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$$

In order to estimate $\mathbf{b}$, we follow an approach based on least squares by minimizing the sum of the squared distances of each observed response to its fitted value under the assumption of non-negative values of $\mathbf{b}$ (more popularly known as non-negative least square (NNLS) method).

$$\mathbf{b}^* = \underset{\mathbf{b} \geq 0}{\operatorname{argmin}} F(\mathbf{b}), \text{where } F(\mathbf{b}) = \frac{1}{2} \left\| \left(\tilde{\mathbf{y}} - \mathbf{X}\mathbf{b} \right) \right\|^2 \tag{4}$$

We solve this typical constraint optimization problem using Lagrange multipliers. The matrix $\mathbf{X}^T\mathbf{X}$ is symmetric and $\mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b} = \|\mathbf{X}\mathbf{b}\|_2^2 \geq 0$. Hence $\mathbf{X}^T\mathbf{X}$ is positive semi-definite. Hence the objective function $F(\mathbf{b})$ is convex. Also the constraints $\mathbf{b} \geq 0$ define a convex feasible set. Hence the minimization problem becomes a convex programming problem for which the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for the optimal solution. Any update algorithm for solving NNLS can be used to solve this (convex programming) minimization problem.

B. Prediction of QoS parameters

Let us assume a QoS parameter Q modeled on ticket data, *e.g.*, Response Time or Resolution Time or Closure Time. Let us also assume that the Fuzzy C-means algorithm described in Section IV, partitions the ticket data with respect to Q into κ clusters with $c_i, 1 \leq i \leq \kappa$, as cluster centers. Let us take a domain for Q as the collection of cluster centers, that is $\Gamma = \{c_1, \ldots, c_\kappa\}$. Then for a ticket T_i we can find the membership value of QoS parameter Q as $\mu_{ij}, 1 \leq j \leq \kappa$. That is, for each ticket we can define a fuzzy set for the QoS Q as $\mu_{i1}/c_1, \cdots, \mu_{i\kappa}/c_\kappa$ with the constraint $\sum_{j=1}^{\kappa} \mu_{ij} = 1$.

Now let us formulate the prediction problem for the QoS parameter Q. We have seen that each ticket T_i can be modeled using a feature vector $(x_{i,1}, \cdots, x_{i,p})$ where $x_{i,j}$ denotes the TF*IDF of term $w_j, 1 \leq j \leq p$, for ticket T_i. The responder variables can be expressed as membership values for the jth cluster as follows, $\mathbf{y^j} = [y_{1,j}(= \mu_{1,j}) \cdots y_{n,j}(= \mu_{n,j})]^T, 1 \leq j \leq \kappa$. Hence using Equation 3 the regression equation for QoS parameter Q will look like $\mathbf{y^j} = X\mathbf{b^j}, 1 \leq j \leq \kappa$.

For a new ticket we need to predict its QoS parameter Q by computing its membership value *wrt* the each of the fuzzy clusters. Using NLP techniques from Section III we can extract the feature vector of new ticket T_{in} as $\mathbf{x_{in}}^T = [1\ x_{in,1}\ \cdots\ x_{in,p}]$. For QoS Q, we predict it membership value $y_{in,j} = \mu_{in,j} = \mathbf{x_{in}}^T\mathbf{b^j}$, where j $(1 \leq j \leq \kappa)$ denotes the jth fuzzy cluster. Need to ensure that $\sum_{i=1}^{\kappa} \mu_{in,i} = \mu_{in,1} + \cdots + \mu_{in,j} + \cdots + \mu_{in,\kappa} = 1$. Hence we normalize the membership values as $\bar{y}_{in,j} = \frac{\mu_{in,j}}{\sum_{i=1}^{\kappa} \mu_{in,i}}$, $1 \leq j \leq \kappa$ which is finally predicted. We assign the likely QoS of an incoming ticket as a particular cluster if the predicted membership value is the maximum for that cluster.

VI. A correlation analysis of QoS parameters

In this section we discuss correlation analysis of QoS parameters including correlation between fuzzy sets and between a fuzzy set and a crisp set.

A. Correlation between fuzzy sets

We adopt the approach of correlation of fuzzy sets as proposed in [9]. Let S be a crisp set with a random set of points $\{z_1, \ldots, z_k\}$. Let A be a fuzzy set defined on S: $A = \{z_i \in S \mid 1 \leq i \leq k, \mu_A(z_i) \in [0, 1]\}$. Then we define,

$$\overline{\mu_A} = \frac{\sum_{i=1}^{k} \mu_A(z_i)}{k}, \quad s_A^2 = \frac{\sum_{i=1}^{k} (\mu_A(z_i) - \overline{\mu_A})^2}{k-1},$$

where $\overline{\mu_A}$ is the average membership grade of fuzzy set A over the set of points $\{z_1, \ldots, z_k\}$ in S. Further s_A^2 denotes the variance of the membership grades of the fuzzy set A over S and the standard deviation is s_A.

Instead of modeling tickets as fuzzy sets we model QoS parameters as fuzzy sets this time. Fix two QoS parameters Q_r and Q_s. Suppose that there are κ_r and κ_s number of clusters for Q_r and Q_s respectively. For Q_r, let the set of clusters be represented as $\Gamma_r = \{c_1^r, \ldots, c_{\kappa_r}^r\}$ where $c_j^r : 1 \leq j \leq \kappa_j$, is the center of the jth cluster determined by Fuzzy C-means algorithm as discussed before. Similarly for QoS Q_s, as the clusters are suitably represented by $\Gamma_s = \{c_1^s, \ldots, c_{\kappa_s}^s\}$. If we have a collection of tickets $\mathcal{T} = \{T_1, \ldots, T_n\}$ at our disposal then it is possible to determine the grade of membership for each of these tickets with respect to the clusters associated with Q_r or Q_s, for example, the membership of ticket T_i with respect to cluster c_j^r for QoS parameter Q_r can be computed as $\mu_{i,j}^{Q_r}$, and likewise the membership of ticket T_i with respect to cluster c_k^s for QoS parameter Q_s can be computed as $\mu_{i,k}^{Q_s}$. We now introduce fuzzy sets on each cluster for each QoS parameter defined on the tickets. That is, we define a fuzzy set representation of cluster c_p^r of QoS Q_r as

$C_p^{Q_r} = \{\mu_{1,p}^{Q_r}/T_1, \ldots, \mu_{n,p}^{Q_r}/T_n\}$ on set $\mathcal{T}$; and similarly the cluster c_q^s of QoS Q_s can be represented as a fuzzy set $C_q^{Q_s} = \{\mu_{1,q}^{Q_s}/T_1, \ldots, \mu_{n,q}^{Q_s}/T_n\}$ on $\mathcal{T}$ etc.

Then we define the correlation co-efficient between two fuzzy sets $C_p^{Q_r}$ and $C_q^{Q_s}$ (computed over the set of $|\mathcal{T}|$ tickets) as:

$$\rho^f(C_p^{Q_r}, C_q^{Q_s}) = \frac{\sum_{k=1}^{n}(\mu_{k,p}^{Q_r}(T_k) - \overline{\mu_p^{Q_r}})(\mu_{k,q}^{Q_s}(T_k) - \overline{\mu_q^{Q_s}})}{s_{C_p^{Q_r}} * s_{C_q^{Q_s}}}, \tag{5}$$

where $\overline{\mu_p^{Q_r}}$ and $\overline{\mu_q^{Q_s}}$ are the average membership grades of fuzzy sets $C_p^{Q_r}$ and $C_q^{Q_s}$. Also $s_{C_p^{Q_r}}$ and $s_{C_q^{Q_s}}$ denote the standard deviations of the membership grades of the fuzzy sets $C_p^{Q_r}$ and $C_q^{Q_s}$ respectively. By Theorem 3.3 in [9] the correlation co-efficient $\rho^f(C_p^{Q_r}, C_q^{Q_s})$ defined by Eqn 5 varies between -1 and 1.

It is possible to compute correlation co-efficient for every possible pair of clusters from two QoS parameters Q_r and Q_s, totaling up to $\kappa_r\kappa_s$ co-efficients. All these coefficients can be combined to determine the correlation between Q_r and Q_s (using ideas from [10]):

$$\rho^f_{combined}(Q_r, Q_s) = \frac{1}{\kappa_r\kappa_s} \sum_{(c_p^r, c_q^s) \in \Gamma_r \times \Gamma_s} \rho^f(C_p^{Q_r}, C_q^{Q_s}) \tag{6}$$

B. Correlation between a fuzzy set and a crisp set

We also find out the correlation between a fuzzy set and a crisp set [11]. Let $\mathcal{T}$ be the total number of tickets. As before, let a fuzzy set representation of cluster c_p^r of QoS Q_r be $C_p^{Q_r} = \{\mu_{1,p}^{Q_r}/T_1, \ldots, \mu_{n,p}^{Q_r}/T_n\}$ on set Γ_r. Let Θ be a crisp set on the set of tickets which is associated with the function $\theta : \mathcal{T} \to \mathbb{R}^{\geq 0}$ such that $\theta(T_i)$ supplies its value for ticket T_i. This function can be used to measure the length/norm of ticket T_i or the number of words appearing in ticket T_i. Then we define the correlation measure between the fuzzy set and the crisp set as

$$\rho^h(C_p^{Q_r}, \Theta) = 1 - \sum_{T_i \in \mathcal{T}} \left| \frac{\mu_{i,p}^{Q_r}(T_i)}{\sum_{T_i \in \mathcal{T}} \mu_{i,p}^{Q_r}(T_i)} - \frac{\theta(T_i)}{\sum_{T_i \in \mathcal{T}} \theta(T_i)} \right| \tag{7}$$

Now, $\min_{\sum_{T_i \in \mathcal{T}}} \left| \frac{\mu_{i,p}^{Q_r}(T_i)}{\sum_{T_i \in \mathcal{T}} \mu_{i,p}^{Q_r}(T_i)} - \frac{\theta(T_i)}{\sum_{T_i \in \mathcal{T}} \theta(T_i)} \right| = 0.$

This happens when both the subtract and subtracted are equal. This implies $\rho^h(C_p^{Q_r}, \Theta) \leq 1 - 0 = 1$.

$$\begin{aligned}
\text{Notice that } & \sum_{T_i \in \mathcal{T}} \left| \frac{\mu_{i,p}^{Q_r}(T_i)}{\sum_{T_i \in \mathcal{T}} \mu_{i,p}^{Q_r}(T_i)} - \frac{\theta(T_i)}{\sum_{T_i \in \mathcal{T}} \theta(T_i)} \right| \\
& \leq \sum_{T_i \in \mathcal{T}} \left(\left| \frac{\mu_{i,p}^{Q_r}(T_i)}{\sum_{T_i \in \mathcal{T}} \mu_{i,p}^{Q_r}(T_i)} \right| + \left| \frac{\theta(T_i)}{\sum_{T_i \in \mathcal{T}} \theta(T_i)} \right| \right) \\
& = \frac{\sum_{T_i \in \mathcal{T}} |\mu_{i,p}^{Q_r}(T_i)|}{|\sum_{T_i \in \mathcal{T}} \mu_{i,p}^{Q_r}(T_i)|} + \frac{\sum_{T_i \in \mathcal{T}} |\theta(T_i)|}{|\sum_{T_i \in \mathcal{T}} \theta(T_i)|} \\
& = 1 + 1 = 2.
\end{aligned}$$

Table I
TICKET DATA FROM DIFFERENT DOMAINS

Domain	Total tickets	No of tuples	Feature vector dimension
AMD	4505	3	145
Travel	3085	11	130
LANM	28520	40	80

We get, $\rho^h(C_p^{Q_r}, \Theta) \geq 1 - 2 = -1$. Finally, $-1 \leq \rho^h(C_p^{Q_r}, \Theta) \leq 1$.

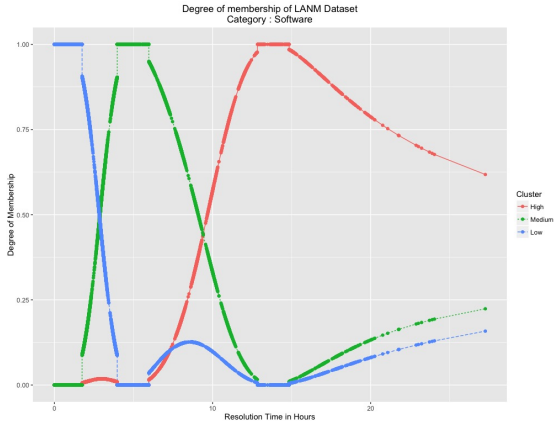


Figure 2. Fuzzy quality model for Resolution Time for LANM:'Software' data set

VII. Experimental Results

In this section we report on experiments on IT maintenance tickets originating from 3 different domains to develop QoS models. The domains are Travel request (Travel), Application Maintenance and Development (AMD) and Local Area Network Management (LANM). We identify different tuples for each data set and and then carry out further analyses for all those tuples. Table I provides the details of 3 top-most populous tuples. For all the data sets we choose only one attribute from the schema for the tuple value, the selection was made by taking the relevant practitioners' opinion into account. Also for some data set if we had more than one attribute to represent a tuple then we had to work with very few tickets for which no meaningful analysis could have been done. We further observed that similar kinds of tickets were raised across such multi-attribute tuples. So only one attribute of the schema as a tuple was a rational choice. As we have the most number of tickets for LANM domain we shall mostly publish results for this data set for lack of space. In Table II, for AMD data 'Software' tuple has the maximum number of tickets, for Travel data 'International travel request' has more tickets and for LANM data the tuple 'Software' contains maximum tickets.

A. Fuzzy clustering of QoS parameters

We run FC clustering algorithm on each tuple for all three data sets. The clustering algorithm on each tuple will be

Table II
TUPLE DETAILS FROM DIFFERENT DOMAINS

Domain	Tuple details	# of tickets
AMD	Software	2978
	Administration	1505
	Hardware	22
Travel	International Travel Request	1023
	Domestic Travel Request	537
	Accommodation	371
LANM	Software	7925
	Desktop	7148
	Laptop	2397

slow on the larger data set as a whole and tickets across the tuple may not exhibit the same QoS characteristic. We consider three QoS parameters, Response time, Resolution Time and Closure Time as discussed before. While it is possible to ascertain all the three QoS parameters for LANM ticket data we could only compute Resolution Time for other two data sets because of lack of information. First we validate the generated clusters using VB validity index. We choose the fuzzy weighing exponent m to be equal to 2 which is normally chosen in the objective function-based method [12]. Based on the value of this index the optimal number of clusters for Resolution Time, Response Time and Closure for all these tuples in LANM data set are set to 3. For these QoS parameters we consider 3 clusters, *viz.*, 'high', 'medium' and 'low'. For Travel data the number of clusters across the tuples is chosen to be 2, 'high' and 'low', while for AMD it is 3 for all the tuples. The FCM algorithm generates the cluster centers of each QoS parameters. Subsequently, these are used to produce fuzzy clustering models. One such fuzzy quality model for 'Software' tuple of LANM data is shown in Figure 2. In this figure each curve corresponds to a cluster. The cluster center is represented by the peak of the curve, which has the degree of membership equal to 1. Some observations can be made on the fuzzy quality models shown in Figure 2. Complex trouble tickets take more than 15 days on average to get resolved. Some tickets are served in the best possible manner when they are served under 3 days. On average tickets in this category are resolved in between 3 and 12 days. Similar observation can be made from each clustering model for all QoS parameters across all data sets.

B. Prediction of QoS for new tickets

For building the regression model we shall use an appropriate update algorithm to solve the NNLS problem described in Section V-A. As the TF*IDF matrix X is sparse and skinny we reduce its dimension by using principal component analysis (PCA). We carry out NNLS on the reduced matrix and obtain the regression parameters. For a new ticket we need to predict its QoS parameter Q by computing its membership value *wrt* the each of the fuzzy clusters. Using NLP techniques discussed previously we can extract the feature vector of new ticket T_{in} (using already mined keyphrases/terms) as $\mathbf{x_{in}}^T = [1\ x_{in,1}\ \cdots\ x_{in,p}]$. For QoS Q we predict its membership value $y_{in,j} = \mu_{in,j} = \mathbf{x_{in}b^j}$, where j $(1 \leq j \leq \kappa)$ denotes the jth fuzzy cluster. Need to ensure that $\sum_{i=1}^{\kappa} \mu_{in,i} = \mu_{in,1} + \cdots + \mu_{in,j} + \cdots + \mu_{in,\kappa} = 1$. Hence we normalize the membership values as $\bar{y}_{in,j} = \frac{\mu_{in,j}}{\sum_{i=1}^{\kappa} \mu_{in,i}}$, $1 \leq j \leq \kappa$, which is finally published. We assign the likely QoS of an incoming ticket to a particular cluster if the predicted membership value is the maximum for that cluster. The cluster curve can be used to determine the range of values for this predicted QoS. For example, if the resolution time for a new ticket in LANM data set is predicted to be in the 'med' cluster then we can say this ticket is likely to be resolved within 3 to 12 days by inspecting Figure 2. We validated our prediction model by dividing the data sets into training and test data sets. For test data we predict the degree of membership of each cluster for a QoS and then compare it with the actual membership value. If the maximum predicted membership value and the actual maximum membership value belong to the same cluster we say there is a match. For LANM data we get a maximum match of 51% for Resolution Time corresponding to 'Desktop' tuple, and 81% for Response Time corresponding to 'Software' tuple, these are depicted in Figure 3. It can be explained by observing that the former tuple contains less number of tickets compared to the latter, and the response time for tickets in 'Software' tuple might not vary much irrespective of systems or persons. For AMD data set we get a maximum match of 78% for 'Administration' tuple. Travel data shows a high accuracy of maximum match of 86% for 'Domestic Travel' request and 75% for 'International Travel' request. We provide confusion matrix for Resolution Time for 'Desktop' tuple of LANM data in Table III. For these particular data tickets, with low Resolution Time were predicted to be having low Resolution Time also in most of the cases. In lot of cases tickets which were resolved in high Resolution Time were predicted to be resolved in low Resolution Time. This can be due to the fact that tickets with low Resolution Time are pre-dominant in the data set. A better accuracy of matches can be obtained by refining the dimension reduction technique.

Table III
CONFUSION MATRIX FOR RESOLUTION TIME FOR 'DESKTOP' TUPLE OF LANM DATA SET

Actual Value		Predicted Value		
		High	Med	Low
	High	14	125	203
	Med	57	810	1123
	Low	237	2566	3714

C. Fuzzy Correlation of QoS parameters

We investigate how different categories (clusters) of QoS are correlated using our framework of correlation introduced in Section VI. As mentioned before, for LANM data set we have derived 3 categories for each QoS, - 'high', 'medium'

Table IV
FUZZY CORRELATION BETWEEN CLOSURE TIME AND RESOLUTION TIME FOR "SOFTWARE" DOMAIN-LANM DATA SET

	Reso-High	Reso-Medium	Reso-Low
Clos-High	0.896	0.134	-0.605
Clos-Med	-0.297	0.346	-0.145
Clos-Low	-0.261	-0.426	0.519

Table V
FUZZY CORRELATION BETWEEN CLOSURE TIME AND RESPONSE TIME FOR "SOFTWARE" DOMAIN-LANM DATA SET

	Resp-High	Resp-Medium	Resp-Low
Clos-High	0.130	0.055	-0.086
Clos-Med	-0.036	0.073	-0.060
Clos-Low	-0.045	-0.107	0.113

and 'low'. The correlation coefficients among the different fuzzy clusters in Software domain of LANM data set are depicted In Table IV. Here Resp-High, Reso-High and Clos-High denote the categories of high Resolution, Response and Closure Time respectively. The correlation coefficients signify the fact that tickets with higher Resolution Time are likely to have high Closure time. On the other hand lower Resolution Time contributes to low Closure Time. The final coefficient correlation among Resolution Time and Closure time is calculated by combining the fuzzy coefficients using Eqn 6: $\rho^{f}_{combined}(\mathcal{S},\mathcal{C}) = 0.654$, which also implies that Resolution Time and Closure Time are positively correlated.

In table V correlation among fuzzy clusters of quality parameters Closure time and Response Time for the same data set is studied. It is evident that there is no strong association among any of these fuzzy clusters. The value of combined correlation, $\rho^{f}_{combined}(\mathcal{C},\mathcal{R}) = 0.654$ implies that Closure time and Response Time for tickets are positively correlated. That is, an increase in Response Time may result in an enhanced Closure time.

For the same data set Table VI shows that there is no strong association among any categories of Resolution and Response Time. This follows causally also. Cluster wise fuzzy correlation among three quality parameters for all tuples of ticket data across the domains are also calculated for which similar kind of patterns (as of LANM data set) are observed. We then combine the correlations for a QoS for a particular domain by taking the average across the tuples. For example, the average correlation between Resolution Time and Response Time for LANM data set is 0.456 which again reflects lack of association between those quality parameters.

We also compute correlation between Norms, #words with categories of QoS using Equation 7 in Table VII. By norm

Table VI
FUZZY CORRELATION BETWEEN RESOLUTION TIME AND RESPONSE TIME FOR "SOFTWARE" DOMAIN-LANM DATA SET

	Resp-High	Resp-Medium	Resp-Low
Reso-High	0.127	0.012	-0.080
Reso-Med	0.040	0.154	-0.158
Reso-Low	-0.072	-0.149	0.171

Table VII
FUZZY-CRISP CORRELATION BETWEEN NORM AND #WORDS WITH RESOLUTION TIME FOR LANM DATSET

Domain	Crisp Set	Reso-High	Reso-Medium	Reso-Low
Travel	Norm	-0.569	-0.213	0.4
AMD	Norm	-0.967	-0.456	0.091
LANM	Norm	-0.699	-0.360	0.345
Travel	#words	-0.602	-0.353	0.310
AMD	#words	-0.959	-0.572	-0.019
LANM	#words	-0.678	-0.678	-0.678

of a ticket we mean the length (Euclidean vector) of it, and by #words, the number of keyphrases contained in a ticket summary. It is evident that norms of tickets across the data set are negatively correlated with high Resolution Time. It tells that if the norm of a tickets is small then it may take higher amount of time to resolve it. Alternately it can be said if the ticket has longer description it carries more information and hence, it is easier to resolve. Across all domains the crisp set #words also exhibit similar kind of association with high Resolution Time.

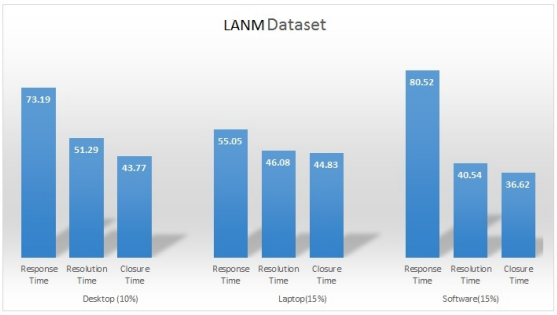


Figure 3. Matches for LANM tickets in different tuples

In future we would like to compare our technique with a simple Q0S-based clustering of tickets and classification of ticket using those cluster labels and the method of [13] by introducing suitable metrics.

VIII. Related Work

It is hard to imagine customer querying for QoS values of web services with exact parameter values since they might have no idea of difference between idealistic and non-idealistic values. The fuzzy-based categorization of QoS of web services can overcome this problem by throwing light on the performance criteria that are considered as good, moderate and poor. This also facilitates the development of fuzzy-based web services solutions. In [5], the authors use the well-known Fuzzy C-Means (FCM) for fuzzy clustering of QoS data for web services to create fuzzy quality models. These models can be used both as reference in requirements negotiations and as inference components of the fuzzy-based application development. Fuzzy analysis can also be applied to the selection of web services based on different QoS parameters. Such a selection of QoS-based web services is done in [14] by applying the fundamental principles

in the fuzzy set theory to model web service selection as Fuzzy Multiple Criteria Decision Making (FMCDM) using a synthetic weight which reflects human rating among others. Human preferences can be biased and often depend on the domain knowledge of the person concerned. So, a better way to QoS clustering is to follow a data driven approach like the proposed one. Automatic Web-service selection also involve reliable predictions for QoS based on historical service innovations. The authors in [15] aim to make highly accurate predictions for missing QoS data via building an ensemble of non-negative latent factor (NLF) models by fitting non-negative QoS data. But the data driven association among QoS parameters and the related analysis are missing, incorporation of which could result in more accurate predictions.

Regression models are also developed for different applications ranging from simulation of data to network prediction in terms of different QoS parameters like throughput, mean delay, missed deadline ratio, collision ratio etc. Dogman *et al.* proposed a QoS evaluation system for computer networks using a combination of fuzzy C-mean (FCM) and regression analysis by which they analyze and assess the QoS in a simulated network [13]. The authors therein considered network QoS parameters of multimedia applications like delay, jitter, loss ratio etc and used FCM algorithms to partition those QoS paramaters into appropriate clusters, the number of which was determined using validity index. The resulting QoS parameters were given as inputs to a regression model in order to quantify the overall QoS value. Our work is an improvement of theirs in the sense that we carry out a more rigorous and comprehensive analysis by using fuzzy regression model with non-negative least square method, which accepts the membership value of each ticket with respect to each category (cluster) of a QoS as response variables.

IX. Conclusions

Feedback of web services plays an important role in measuring the quality of services. Often the feedback depends on the performances of resolvers for a generated complaint. Trouble tickets are generated and closed after they are resolved. Thus the Resolution, Response and Closure Time of tickets influence the service performance. Other performance parameters such as Latency, Availability and Reliability can be computed in terms of Resolution time, Agreed service Time and Closure Time (which we leave for future work). Thus the determination of the performance (Resolution and Closure Time) of tickets are of paramount importance as they have high impact on the quality parameters of web services.

References

[1] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[2] D. Dubois and H. Prade, *Fuzzy Sets and Systems*. New York: Academic Press, 1988.

[3] J. Leskovec, A. Rajaraman, and J. Ullman, *Mining of Massive Datasets*, 2nd ed. Cambridge University Press, 2014.

[4] S. Roy, D. P. Muni, J. Y. T. Yan, N. Budhiraja, and F. Ceiler, "Clustering and labeling IT maintenance tickets," in *14th ICSOC'16, Banff, AB, Canada, Proceedings*, 2016, pp. 829–845.

[5] M. H. Hasan, J. Jaafar, and M. F. Hassan, "Development of web services fuzzy quality models using data clustering approach," in *Proceedings of the 1st DaEng*. Springer, 2014, pp. 631–640.

[6] H. Guldemır and A. Sengur, "Comparison of clustering algorithms for analog modulation classification," *Expert Systems with Applications*, vol. 30, no. 4, pp. 642–649, 2006.

[7] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems*, vol. 17, no. 2-3, pp. 107–145, 2001.

[8] H. Tanaka, S. Uejima, and K. Asai, "Linear regression analysis with fuzzy model," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 12, no. 6, pp. 903–907, 1982.

[9] D. Chiang and N. P. Lin, "Correlation of fuzzy sets," *Fuzzy Sets and Systems*, vol. 102, no. 2, pp. 221–226, 1999.

[10] E. Szmidt, J. Kacprzyk, and P. Bujnowski, "Pearson coefficient between intuitionistic fuzzy sets," *Notes on Intuitionistic Fuzzy Sets*, vol. 17, no. 2, pp. 25–34, 2011.

[11] B. B. Chaudhuri and A. Bhattacharya, "On correlation between two fuzzy sets," *Fuzzy Sets and Systems*, vol. 118, no. 3, pp. 447–456, 2001.

[12] N. R. Pal and J. C. Bezdek, "On cluster validity for the Fuzzy C-means model," *IEEE Trans. Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, 1995.

[13] A. Dorgman, R. Saatchi, and S. AI-Khayatt, "Quality of Service Evaluation using a Combination of Fuzzy C-Means and Regression Model," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 6, no. 1, pp. 62–69, 2012.

[14] P. Xiong and Y. Fan, "QoS-Aware Web Service Selection by a Synthetic Weight," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD'07, Proceedings, Volume 3*, 2007, pp. 632–637.

[15] X. Luo, M. Zhou, Y. Xia, Q. Zhu, A. C. Ammari, and A. Alabdulwahab, "Generating Highly Accurate Predictions for Missing QoS Data via Aggregating Nonnegative Latent Factor Models," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 27, no. 3, pp. 524–537, 2016.