In the Name of God

Journal of
Information Systems & Telecommunication

Vol. 7, No. 2, April-June 2019, Serial Number 26

Research Institute for Information and Communication Technology
Iranian Association of Information and Communication Technology
Affiliated to: Academic Center for Education, Culture and Research (ACECR)

Manager-in-Charge: Habibollah Asghari, ACECR, Iran
Editor-in-Chief: Masoud Shafiee, Amir Kabir University of Technology, Iran

Executive Manager: Shirin Gilaki
Executive Assistants: Ali Mokhtarani, Mahdokht Ghahari
Print ISSN: 2322-1437
Online ISSN: 2345-2773
Publication License: 91/13216
Editorial Office Address: No.5, Saeedi Alley, Kalej Intersection., Enghelab Ave., Tehran, Iran,
P.O.Box: 13145-799 Tel: (+9821) 88930150 Fax: (+9821) 88930157
E-mail: info@jist.ir , infojist@gmail.com
URL: www.jist.ir

Indexed by:

- SCOPUS	www. Scopus.com
- Index Copernicus International	www.indexcopernicus.com
- Islamic World Science Citation Center (ISC)	www.isc.gov.ir
- Directory of open Access Journals	www.Doaj.org
- Scientific Information Database (SID)	www.sid.ir
- Regional Information Center for Science and Technology (RICeST)	www.ricest.ac.ir
- Iranian Magazines Databases	www.magiran.com

Publisher:
Regional Information Center for Science and Technology **(RICeST)**
Islamic World Science Citation Center **(ISC)**

This Journal is published under scientific support of
Advanced Information Systems (AIS) Research Group and
Digital & Signal Processing Research Group, ICTRC

Acknowledgement

JIST Editorial-Board would like to gratefully appreciate the following distinguished referees for spending their valuable time and expertise in reviewing the manuscripts and their constructive suggestions, which had a great impact on the enhancement of this issue of the JIST Journal.

(A-Z)

- Abdolvand, Neda, Azahra University, Tehran, Iran
- Abdoos, Monire, Shahid Beheshti University, Tehran, Iran
- Ahmadizad, Arman, University of Kurdistan, Kurdistan, Iran
- Alavi, Seyed Enayatollah, Chamran University, Ahvaz, Iran
- Bashiri, Hassan, Hamedan University of Technology, Hamedan, Iran
- Baradaran,Vahid, Islamic Azad University, North Tehran Branch, Iran
- Cheraghchi, Hamideh Sadat, Shahid Beheshti University, Tehran, Iran
- Dehkharghani, Rahim, University of Bonab, East Azerbaijan, Iran
- Farsijani, Hassan, Shahid Beheshti University, Tehran, Iran
- Feshari, Majid, Kharazmi University, Tehran, Iran
- Ghasemzadeh, Mohammad, Yazd University, Yazd, Iran
- Ghaffari, Ali, Islamic Azad University, Tabriz Branch, Iran
- Ghiamy, Mohammad, Islamic Azad University, Ardabil Branch, Iran
- Hosseini, Monireh, K. N. Toosi University of Technology, Tehran, Iran
- Jazayeriy, Hamid, Babol Noshivani University of Technology, Mazandaran, Iran
- Kashef, Seyed Sadra, Urmia University, Urmia, Iran
- Mirroshandel, Seyed Abolghasem, University of Guilan, Rasht, Iran
- Mousavirad, Seyed Jalaleddin, University of Kurdistan, Kurdistan, Iran
- Mahdieh, omid, University of Zanjan, Zanjan, Iran
- Mohammadi Zanjireh, Morteza, Imam Khomeini International University, Qazvin, Iran
- Nazemi, Eslam, Shahid Beheshti University, Tehran, Iran
- Oraizi, Homayoon, University of Science and Technology, Tehran, Iran
- Pirgazi, Jamshid, Zanjan University, Zanjan, Iran
- Reshadat, Vahideh, Malek-Ashtar University of Technology, Tehran, Iran
- Rasi, Habib, Shiraz University of Technology, Shiraz, Iran
- Rafighi, Masoud, Malek-Ashtar University of Technology, Tehran, Iran
- Saeb, Armin, Stellenbosch University, Stellenbosch, South Africa
- Tavakoli, Hassan, University of Guilan, Rasht, Iran
- Taghizadeh, Alireza, Islamic Azad University, Parand Branch, Iran
- Yadollahzadeh Tabari, Meisam, Shahrood University Of Technology, Iran
- Zayani, Hadi, Shiraz University of Technology, Shiraz, Iran

Table of Contents

Wavelet-based Bayesian Algorithm for Distributed Compressed Sensing

Razieh Torkamani*
Faculty of Electrical Engineering, K.N. Toosi University of Technology, Tehran, Iran
rtorkamani@mail.kntu.ac.ir
Ramazan Ali Sadeghzadeh
Faculty of Electrical Engineering, K.N. Toosi University of Technology, Tehran, Iran
sadeghz@eetd.kntu.ac.ir

Received: 23/June/2019 Revised: 28/Oct/2019 Accepted: 20/Nov/2019

Abstract

The emerging field of compressive sensing enables the reconstruction of the signal from a small set of linear projections. Traditional compressive sensing approaches deal with a single signal; while one can jointly reconstruct multiple signals via distributed compressive sensing algorithm, which exploits both the inter- and intra-signal correlations via joint sparsity models. Since the wavelet coefficients of many signals is sparse, in this paper, the wavelet transform is used as sparsifying transform, and a new wavelet-based Bayesian distributed compressive sensing algorithm is proposed, which takes into account the inter-scale dependencies among the wavelet coefficients via hidden Markov tree model, as well as the inter-signal correlations. This paper uses Bayesian procedure to statistically model this correlation via the prior distributions. Also, in this work, a type-1 joint sparsity model is used for jointly sparse signals, in which every sparse coefficient vector is considered as the sum of a common component and an innovation component. In order to jointly reconstruct multiple sparse signals, the centralized approach is used in distributed compressive sensing, in which all the data is processed in the fusion center. Also, variational Bayes procedure is used to infer the posterior distributions of unknown variables. Simulation results demonstrate that the structure exploited within the wavelet coefficients provides superior performance in terms of average reconstruction error and structural similarity index.

Keywords: Distributed Compressive Sensing; Joint Saprsity; Signal Reconstruction; Wavelet Transform; Hidden Markov Tree Model; Variational Bayes.

1- Introduction

Compressive sensing (CS) constitutes a framework for sampling of the signals at a rate lower than the Shannon-Nyquist sampling rate [1, 2, 3]. According to the CS theory, when the signal has a sparse representation in a particular basis, one can reconstruct the original signal from a reduced number of linear projections. In order to recover the original signal from compressed measurements, CS exploits the sparsity, i.e. the intra-signal correlation of the signal. But in some applications, the signals may possess many dependencies, which is referred to as inter-signal correlation. Distributed CS (DCS), as a generalization of CS, is proposed in [4- 6], and aims to exploit the intra- and inter-signal correlations simultaneously, and jointly reconstruct a set of signals. According to the DCS algorithm, in addition to each signal being individually sparse, some of their nonzero components are common. Since the signals share a common support in DCS, they can be jointly reconstructed from dramatically fewer measurements in comparison with their independent reconstruction.

1-1- Related work

Distributed algorithms for solving multiple sparse signal reconstruction problem generally divided into two categories: centralized and decentralized [7- 11]. In a centralized approach, each node runs the CS undersampling procedure independently. Then, all of the local measurements obtained from each node are collected in a fusion center (FC) which estimates the joint-sparse signals and transmits the reconstructed sparse signals back to the respective nodes. [4, 5, 6]. In contrast, in decentralized approaches, processing the measurements is performed at each node by allowing some inter-node exchange of information. In this paper, a centralized approach is proposed for joint reconstruction of sparse signals in DCS. Also, the multiple sparse signals considered in this paper belong to the type-1 joint sparse model (JSM-1), in which all of the signals are assumed to

* Corresponding Author

have a sparse common component and an innovation component [7, 8].

Recently, many DCS algorithms have been proposed for the reconstruction of signals from CS measurements in a centralized and decentralized manner. In [4] a greedy algorithm is developed for joint signal recovery, which assumes a JSM-2 model for joint sparse signals. In [12], two DCS algorithms are proposed for the VB inference problem, which introduces a distributed VB framework for conjugate-exponential models. In the first algorithm, the global parameters at each node are optimized. In the second method, the variational optimization is redefined as a constrained minimization problem with a modified objective function. In [11] a decentralized Bayesian DCS algorithm is proposed to reconstruct multiple sparse signals. This paper uses a JSM-1 model and develops the variational approximation in the Bayesian formulation to jointly reconstruct the sparse signals. In [7] a distributed greedy algorithm based on Orthogonal Matching Pursuit (OMP) is proposed for JSM-2 signal recovery. This algorithm estimates the locations of non-zero elements of the sparse signal in an iterative manner, while considering a priori knowledge of the size of the nonzero support set, which could be unknown or hard to estimate.

Distributed recovery algorithms have been efficiently studied in many CS applications that allow joint reconstruction of multiple sparse signals. For example, in [13-15], DCS have been applied to the wireless sensor networks (WSNs), where each sensor performs its measurements independently, and then, DCS develops an algorithm for the reconstruction problem, performing most of the computations in the joint decoder, which can reduce the computational complexity and energy consumption. Also, a novel video coding framework is proposed based on DCS [16-18], in which video frames are classified as CS-frames and K-frames and encode the frames using CS. Another application of the DCS theory is image fusion methods [19-21], where two or more images of the same situation combine and constitute an image which is appropriate for practical applications. A suitable fusion of visible and infrared images can obtain a precise, reliable and proper exposition of the environmental conditions. In [22], a multi-channel SAR system based on DCS has been proposed, which exploits the coherence among multiple channels, in addition to the sparsity of each channel. The joint processing requires a reduced number of samples than the multi-channel SAR imaging system based on traditional CS. Except the above-mentioned applications, there are many other application domains for DCS including: MIMO channel [23], speech enhancement (SE) [24], multichannel electroencephalogram (EEG) [25], joint channel estimation [26], and ground moving target indication (GMTI) [27].

1-2- Contribution

As mentioned above, any DCS algorithm must rely on a dependence model that illustrates the intra- and inter-signal correlation of the sparse signals. The main drawback of the previous algorithms is that the only assumption considered for each of the signals, is the sparsity of the individual signals and they do not enumerate the interdependency structure among the sparse signal coefficients. To address this drawback, and based on the fact that the wavelet transform of many signals is sparse [28, 29], in this paper, the discrete wavelet transform (DWT) has been used as the sparsifying basis. The main contribution of this paper is to exploit the tree-structure of wavelet coefficients in the proposed reconstruction algorithm to demonstrate the dependency among the wavelet coefficients. It has been proved that exploiting signal structure in addition to sparsity for CS, results in a decrease in the number of CS measurements [28].

In this work, Bayesian method is employed for the reconstruction of the signal from underdetermined data, which results in the wavelet-based Bayesian DCS (WB-DCS) algorithm. Furthermore, a Gaussian pdf is assumed for the sparse coefficients, and variational Bayes (VB) inference is employed to derive the posterior probabilities. Experimental results show that the proposed WB-DCS algorithm provides a superior reconstruction quality than the other state-of-the-art approaches.

The reminder of this paper is organized as follows. Section 2 briefly reviews the wavelet-based BCS. The Bayesian DCS framework and the VB inference procedure are provided in Section 3. Simulation results are reported in Section 4. Finally, conclusions are discussed in Section 5.

2- Wavelet-based Bayesian Compressive Sensing

In this section, we explain the CS recovery problem via Bayesian framework and use the DWT as the sparsifying basis. Let $x \in \mathbb{R}^{N\times 1}$ denote the original signal. The DWT of the signal x can be represented as [30]

$$x = \Psi\theta \qquad (1)$$

where $\Psi \in \mathbb{R}^{N\times N}$ is the matrix containing wavelet basis vectors as its columns, and $\theta \in \mathbb{R}^{N\times 1}$ is the vector of wavelet coefficients. The wavelet coefficients can be represented in a tree structure, in which every coefficient at scale s has four children at the next scale, and the statistical relationship among the parent and children coefficients is such that if the parent coefficient is negligible, then its children are also negligible. The statistical relationship between the wavelet coefficients can be demonstrated via the hidden Markov tree (HMT) model [28, 31, 32]. Fig.1 shows the DWT with three wavelet decomposition levels and two associated wavelet trees.

It has been proved that the wavelet transform of many signals and images have sparse representation, thus, enabling us to utilize the CS theory. The classical CS data acquisition is modeled by

$$y = D\theta + n = \sum_{i=1}^{N} \theta_i z_i d_i + n \qquad (2)$$

where $y \in \mathbb{R}^{M\times 1}$ denotes the vector of CS measurements, $D \in \mathbb{R}^{M\times N}$ is the measurement matrix, $n \in \mathbb{R}^{M\times 1}$ is the measurement noise, d_i is the i'th column of D, and z_i is the support of the i'th coefficient, i.e. $z_i = 0\ (1)$ means that the i'th element of θ is zero (nonzero).

In this paper, the CS problem is formulated via a Bayesian perspective. Bayesian CS (BCS) aims to estimate the sparse coefficients vector θ from measurements y via considering a suitable pdf for hidden variables [33]. Let the measurement noise n has a zero-mean Gaussian prior distribution with precision (inverse variance) α_n. Then the likelihood function is given by

$$p(y|\theta, z) = \mathcal{N}(D\theta\,, \alpha_n^{-1} I_N) \qquad (3)$$

where $I_N \in \mathbb{R}^{M\times N}$ is an identity matrix. In this work, a zero-mean Gaussian distribution is assumed for the wavelet coefficients

$$p(\theta) = \prod_{i=1}^{N} p(\theta_i^s) \qquad p(\theta_i^s) = \mathcal{N}(0\,, \alpha_s^{-1}) \qquad (4)$$

where θ_i^s is the i'th component of sparse signal θ, which is at scale s, and α_s is the precision of the pdf, which is assumed to be common for all coefficients at scale s. In the next section, proposed JSM-1 DCS algorithm, namely WB-DCS, is introduced and present the prior distributions for other variables.

3- Distributed Compressive Sensing Model

In this section, we extend the BCS procedure explained in the previous section and present the proposed WB-DCS algorithm for joint reconstruction of multiple correlated signals. Also, the interactions of multiple signals is modeled via JSM-1 DCS model.

Suppose that the network has K nodes and can be modeled by an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{1, \dots, K\}$ is the set of nodes and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of undirected edges that characterizes the links between the nodes. In a particular graph, there is a link between two nodes if they are neighbors. Fig. 2 shows an example graph with 7 nodes.

The CS measurements for each node is given as

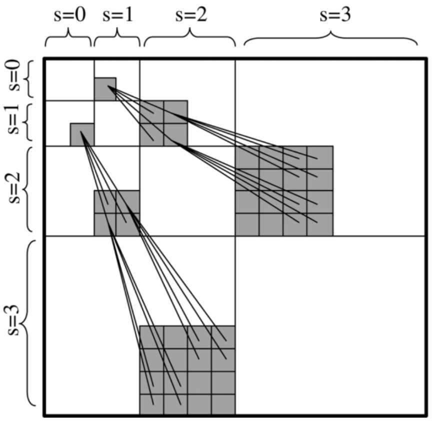


Fig. 1 The HMT structure of wavelet coefficients. [28]

$$y_k = D_k\theta_k + n_k \qquad k = 1, \dots, K \qquad (5)$$

where $y_k \in \mathbb{R}^{M_k\times 1}$, $D_k \in \mathbb{R}^{M_k\times N}$, $\theta_k \in \mathbb{R}^{N\times 1}$, $n_k \in \mathbb{R}^{M_k\times 1}$ denote the measurement vector, sensing matrix, sparse signal, and noise for k'th signal, respectively, and K is the total number of signals. According to the JSM-1 DCS model, the sparse signal θ_k can be represented as

$$\theta_k = w_c \odot z_c + w_k \odot z_k \qquad k = 1, \dots, K \qquad (6)$$

where $w_c \in \mathbb{R}^{N\times 1}$ denotes the common component of the sparse signal, $z_c \in \{0,1\}^{N\times 1}$ is the support vector of w_c, $w_k \in \mathbb{R}^{N\times 1}$ is the innovation component of the θ_k, which is specific to the k'th signal, $z_k \in \{0,1\}^{N\times 1}$ is the support vector of w_k, and $\odot$ denotes the Hadamard product.

In this paper, a zero-mean Gaussian distributions is assumed for the innovation and common components. Also, for modeling the elements of common and innovation supports, z_c and z_k, a HMT model is imposed in a statistical manner. Hence, the priors are given as

$$p(w_k) = \mathcal{N}(0\,, \Gamma_k) \qquad (7)$$

$$p(w_c) = \mathcal{N}(0\,, \Gamma_c) \qquad (8)$$

$$p\big(z_{k,s,i}\big) = Bernoulli(\pi_{k,s,i}) \qquad (9)$$

$$p\big(z_{c,s,i}\big) = Bernoulli(\pi_{c,s,i}) \qquad (10)$$

$$\pi_{k,s,i} = \begin{cases} \pi_{k,s} & s = 0,1 \\ \pi_{k,s0} & 2 \le s \le L, z_{pa(k,s,i)} = 0 \\ \pi_{k,s1} & 2 \le s \le L, z_{pa(k,s,i)} = 1 \end{cases} \qquad (11)$$

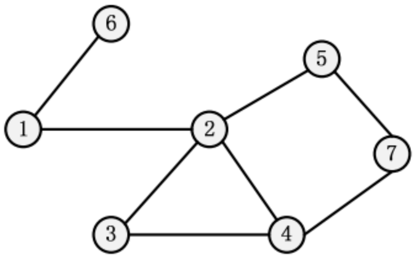


Fig. 2 A typical network structure with 7 nodes.

$$\pi_{c,s,i} = \begin{cases} \pi_{c,s} & s = 0,1 \\ \pi_{c,s0} & 2 \leq s \leq L, z_{pa(c,s,i)} = 0 \\ \pi_{c,s1} & 2 \leq s \leq L, z_{pa(c,s,i)} = 1 \end{cases} \tag{12}$$

$$p(\alpha_n) = Gamma(a_0, b_0) \tag{13}$$

$$p(\pi_{k,s}) = Beta(e_{k,s}, f_{k,s}) \qquad s = 0,1 \tag{14}$$

$$p(\pi_{k,s0}) = Beta(e_{k,s0}, f_{k,s0}) \qquad s = 2, \dots, L \tag{15}$$

$$p(\pi_{k,s1}) = Beta(e_{k,s1}, f_{k,s1}) \qquad s = 2, \dots, L \tag{16}$$

$$p(\pi_{c,s}) = Beta(e_{c,s}, f_{c,s}) \qquad s = 0,1 \tag{17}$$

$$p(\pi_{c,s0}) = Beta(e_{c,s0}, f_{c,s0}) \qquad s = 2, \dots, L \tag{18}$$

$$p(\pi_{c,s1}) = Beta(e_{c,s1}, f_{c,s1}) \qquad s = 2, \dots, L \tag{19}$$

where $\Gamma_k \in \mathbb{R}^{N \times N}$ and $\Gamma_c \in \mathbb{R}^{N \times N}$ are diagonal matrices whose elements are the precisions $\alpha_{k,s}$ and $\alpha_{c,s}$, respectively, (k, s, i) and (c, s, i) denote the index of i'th element at scale s which belongs to the k'th innovation component and the common component, respectively, $\pi_{k,s,i}$ and $\pi_{c,s,i}$ denote the mixing weights adopting Beta priors with the specified hyperparameters, $z_{pa(k,s,i)}$ and $z_{pa(c,s,i)}$ denote the support of the parent coefficient of $w_{(k,s,i)}$ and $w_{(c,s,i)}$, respectively, and L is the total number of wavelet decomposition levels.

3-1- Variational Bayes Inference

To solve the joint recovery problem and infer the posterior distributions of the proposed WB-DCS algorithm, VB inference is implemented. The fundament of VB inference is to provide an estimate of the true posterior distribution $p(.)$, say $q(.)$, by adopting a factorable distribution [34]. For simplicity, define $Y = \{y_1, \dots, y_K\}$, $W = \{w_c, w_1, \dots, w_K\}$, $Z = \{z_c, z_1, \dots, z_K\}$ and $\theta = \{\Gamma_c, \Gamma_1, \dots, \Gamma_K, \pi_c, \pi_1, \dots, \pi_K, \alpha_n\}$. Assume that the $q(W)$ and $q(Z)$ can be factorized as

$$q(W) = q(w_c)q(w_1) \dots q(w_K) \tag{20}$$

$$q(Z) = q(z_c)q(z_1) \dots q(z_K) \tag{21}$$

Then, based on the priors presented in the previous section, the estimated posterior distributions $q(w_c)$ and $q(w_k)$ at each iteration are given by

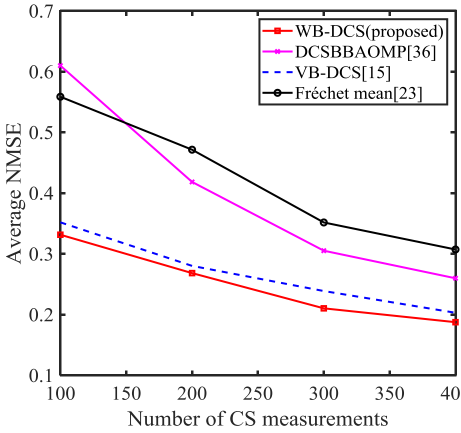
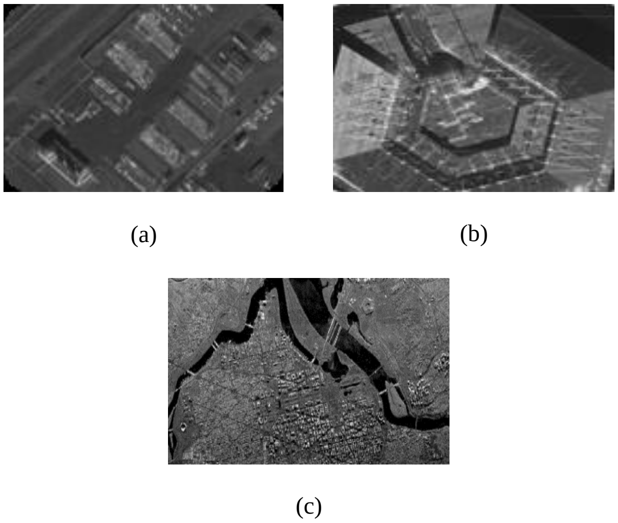


Fig. 3 Comparison of the normalized mean square error for the temperature signals.



(a)

(b)

(c)

Fig. 4 3-D SAR images downloaded from [38].

$$q(w_c) \propto \exp(\mathbb{E}_{q(w_1),\dots,q(w_K)}[\ln p(y_1, \dots, y_K, w_c, w_1, \dots, w_K$$

$$, z_c, z_1, \dots, z_K; \Gamma_c, \Gamma_1, \dots, \Gamma_K, \pi_c, \pi_1, \dots, \pi_K, \alpha_n)])$$

$$\propto \exp(\mathbb{E}_{q(w_1)}[\ln p(y_1|w_c, w_1, z_c, z_1, \alpha_n) + \cdots + \ln p(y_K|w_c, w_K, z_c, z_K, \alpha_n) + \ln p(w_c|\Gamma_c)]$$

$$= \mathcal{N}(\mu_c, \Sigma_c) \tag{22}$$

$$q(w_k) \propto \exp(\mathbb{E}_{q(w_c),q(w_j),j\neq k}[\ln p(y_1, \dots, y_K, w_c, w_1, \dots,$$

$$w_K, z_c, z_1, \dots, z_K; \Gamma_c, \Gamma_1, \dots, \Gamma_K, \pi_c, \pi_1, \dots, \pi_K, \alpha_n)])$$

$$\propto \exp(\mathbb{E}_{q(w_c)}[\ln p(y_k|w_c, w_k, z_c, z_k, \alpha_n) + \ln p(w_k|\Gamma_k)]$$

$$= \mathcal{N}(\mu_k, \Sigma_k) \tag{23}$$

where

$$\mu_c = \alpha_n \Sigma_c {Z_c}^T (\textstyle\sum_{k=1}^{K} \{D_k^T (y_k - D_k Z_k \mu_k)\}) \tag{24}$$

$$\Sigma_c = \left\{\Gamma_c^{-1} + \alpha_n {Z_c}^T (\textstyle\sum_{k=1}^{K} D_k^T D_k) Z_c\right\}^{-1} \tag{25}$$

$$\mu_k = \alpha_n \Sigma_k Z_k^T D_k^T (y_k - D_k Z_c \mu_c) \tag{26}$$

$$\Sigma_k = (\alpha_n Z_k^T D_k^T D_k Z_k + \Gamma_k^{-1})^{-1} \tag{27}$$

where $Z_{(.)} = diag(z_{(.)})$. According to the Eqs. (22) and (23), it can be authenticated that $q(w_c)$ and $q(w_k)$ are Gaussian distributions. Applying similar process for other hidden variables, where $q(w_c)$ and $q(w_k)$ are given, approximate posterior distributions are obtained as follows:

$$q(z_{k,s,i}) \propto p(z_{k,i}) \sqrt{\Sigma_{k,i}}\, exp\left(-\frac{1}{2}\frac{\mu_{k,i}^2}{\Sigma_{k,i}}\right) \tag{28}$$

$$q(z_{c,s,i}) \propto p(z_{c,i}) \sqrt{\Sigma_{c,i}}\, exp\left(-\frac{1}{2}\frac{\mu_{c,i}^2}{\Sigma_{c,i}}\right) \tag{29}$$

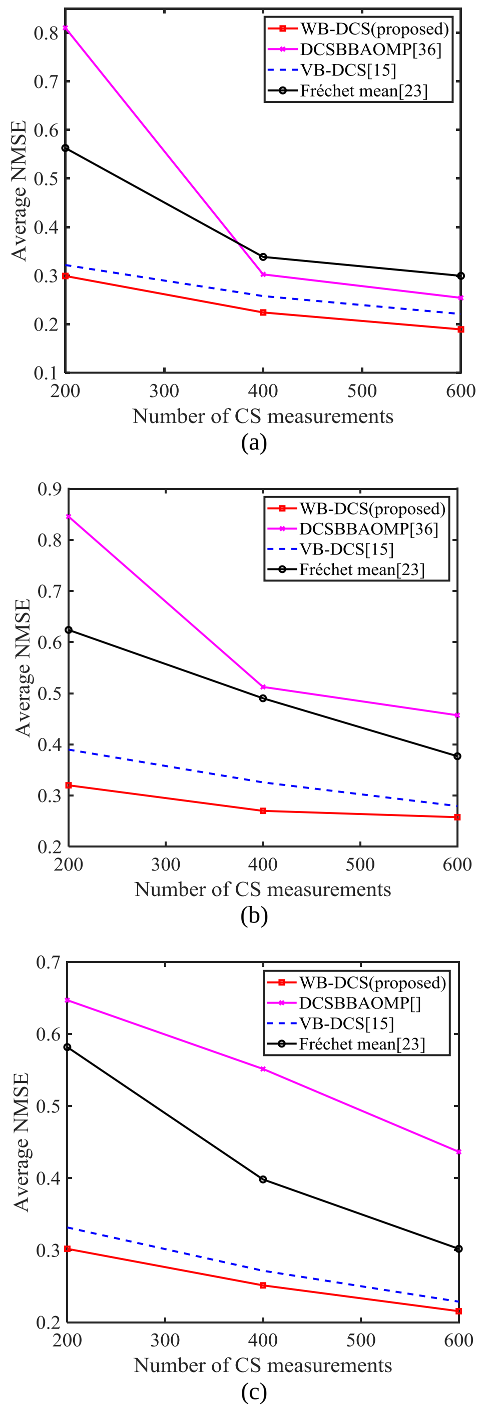
$$q(\alpha_n) = Gamma(a_0', b_0') \tag{30}$$

$$q(\pi_k) = \prod_{s=0}^{1} Beta(e_{k,s}', f_{k,s}') \times \prod_{s=2}^{L} Beta(e_{k,s0}', f_{k,s0}') Beta(e_{k,s1}', f_{k,s1}') \tag{31}$$

$$q(\pi_c) = \prod_{s=0}^{1} Beta(e_{c,s}', f_{c,s}') \times \prod_{s=2}^{L} Beta(e_{c,s0}', f_{c,s0}') Beta(e_{c,s1}', f_{c,s1}') \tag{32}$$

$$q(\alpha_{k,s}) = Gamma(c_0', d_0') \tag{33}$$

$$q(\alpha_{c,s}) = Gamma(e_0', f_0') \tag{34}$$



(a)

(b)

(c)

Fig. 5 Comparison of the normalized mean square error for the SAR images of Fig. 4.

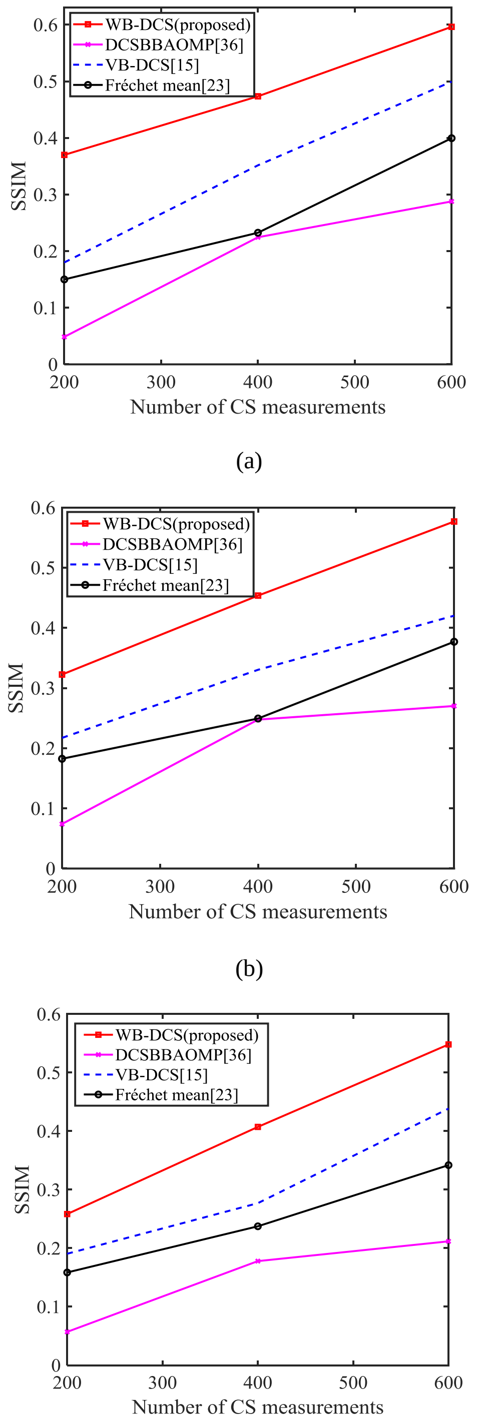
where $a_0' = a_0 + \frac{KM}{2}$ and $b_0' = b_0 + \frac{1}{2}\sum_{k=1}^{K}\|y_k - D_k\theta_k\|_2^2$. The variational optimization procedure iteratively updates until convergence occurs to stable hyperparameters. Finally, the reconstructed signal can be obtained as

$$\theta_k = \mu_c \odot z_c + \mu_k \odot z_k \tag{35}$$

4- Simulation Results

In this section, the performance of the proposed centralized WB-DCS algorithm in two settings is evaluated. First, the experiments are tested for the 1-D temperature signals. In the second scenario, the efficiency of the proposed algorithm is investigated for the 3-D SAR images. The results of the proposed algorithm and that of three recent algorithms presented for DCS signal reconstruction are compared: the centralized part of the Bayesian DCS algorithm proposed in [11], the centralized Fréchet mean approach [35], and the backtracking-based adaptive orthogonal matching pursuit for block distributed compressed sensing (DCSBBAOMP) algorithm proposed in [36]. All of the competing algorithms assume JSM-1 model for the signals and exploit both the intra- and inter-signal dependencies. The centralized algorithm used in [11] is a Bayesian DCS algorithm and estimates the jointly sparse signals based on the VB inference procedure. The Fréchet mean approach is also a centralized algorithm, but the effect of innovation components in the reconstruction of common component is ignored. The DCSBBAOMP algorithm reconstructs block-sparse signals in an iterative manner, which each iteration consists of forward selection and backward removal stages.

In the following evaluations, the DWT is used as the sparsifying transform, and the elements of all the sensing matrices D_k are i.i.d and drawn randomly from a zero-mean Gaussian distribution with variance $\frac{1}{M}$. The sparsifying domain used in [11] is the discrete cosine transform (DCT), and the sensing matrices are random partial DCT matrices. For fair comparison, the same settings is used in all simulations, i.e. the curves presented as the results of [11] are based on the sparse coefficients obtained by the DWT transform and Gaussian measurement matrices. Also, parameter setting for the DCSBBAOMP algorithm is the same as [36].



(a)

(b)

(c)

Fig. 6 Comparison of the structural similarity index for the SAR images of Fig. 4.

4-1- Experiments with 1-D Signals

In this subsection, the algorithms mentioned above are tested for the 1-D temperature signals downloaded from the Intel Berkeley Research lab [37]. A set of $K = 10$ temperature signals of length $N = 512$ is considered and in the recovery process, this signals are jointly reconstructed in an iterative manner. The comparison of competing algorithms is in terms of normalized mean square error (NMSE), $NMSE(x) = 10\log\left(\frac{\|x-\hat{x}\|_2^2}{\|x\|_2^2}\right)$, where x and $\hat{x}$ denote the original and the reconstructed signal, respectively. For each experiment setting, 100 trials are implemented and the averaged result are provided.

The average NMSE results of the reconstructed temperature images are displayed in Fig. 3. It is seen that the proposed WB-DCS algorithm, which exploits the structure of wavelet coefficients, has better performance than the other algorithms in terms of NMSE, where they only assumption is the intra- and inter-signal correlations of the signals.

4-2- Experiments with 3-D Signals

In the following set of experiments, the algorithms mentioned above are compared for 3-D SAR images. Three SAR images are selected, which are downloaded from Sandia National Laboratories in U.S. [38] and shown in Fig. 4. All simulations are for 32×32 images. For all competing algorithms, two quality assessors are used to compare the results: (1) NMSE, and (2) structural similarity (SSIM) [39], which evaluates the similarity between the original and the reconstructed image. The SSIM index is defined as $SSIM(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{(2\mu_x\mu_{\hat{x}}+C_1)(2\sigma_{x\hat{x}}+C_2)}{(\mu_x^2+\mu_{\hat{x}}^2+C_1)(\sigma_x^2+\sigma_{\hat{x}}^2+C_2)}$, where $\mu_{(.)}$ is the mean intensity, $\sigma_{(.,.)}$is the covariance, $\sigma_{(.)}^2$ is the variance, and C_1 and C_2 are some constants.

The utilization of the DCS algorithm for each SAR image, exploits structural dependencies between adjacent azimuth–range pixels and/or polarimetric channels [40].

Fig. 5 depicts the NMSE results versus the number of experiments for the SAR images of Fig. 3. Each point is based on the average of 100 trials. It can be seen that the proposed method obtains the lowest reconstruction error among all the other competing algorithms.

A SSIM comparison between the proposed WB-DCS algorithm and the competitive methods is illustrated in Fig. 6. Based on this results, it can be demonstrated that the use of a profitable model for enumerating the inter-scale, intra- and inter-signal dependencies of jointly sparse signals, simultaneously, leads to an improvement in DCS signal recovery in terms of SSIM, such that the highest SSIM is obtained by the proposed algorithm.

5- Conclusion

In this paper, a centralized wavelet-based Bayesian DCS algorithm (WB-DCS) is proposed to jointly reconstruct multiple signals. Both the inter- and intra-signal correlations are exploited by the JSM-1 DCS model. Furthermore, the DWT is used as sparsifying transform and the HMT model is employed to demonstrate the inter-scale structure associated with wavelet coefficients. Also, this correlation is statistically employed within a Bayesian prior, and the posteriors of unknown variables are estimated using the VB inference procedure. Experimental results confirmed that the proposed WB-DCS algorithm significantly outperforms the state-of-the-art DCS recovery algorithms in terms of reconstruction error and structural similarity index (SSIM). Future research includes exploiting approximate message passing algorithm for the recovery process, which can help significantly reduce the computational complexity of Bayesian inference. Also, in addition to the sparsity and the local structure of the sparse signals considered in this work, we would like to exploit the nonlocal self-similarity of images in our future work, which represents the repetitive behavior of the higher level patterns globally located in the images.

Acknowledgments

Authors wish to thank Dr. Hadi Zayyani for his beneficial discussions and comments.

References

[1] M. Fornasier and H. Rauhut, “Compressive Sensing”, 2010.

[2] E. J. Candès and M. B. Wakin, “An Introduction to Compressive Sampling”, IEEE Signal Process. Magazine, 2008, pp. 21-30.

[3] R.G. Baraniuk, “Compressive Sensing”, IEEE Signal Process. Mag., 2007, pp. 118-124.

[4] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin and R. G. Baraniuk, “Distributed Compressed Sensing of Jointly Sparse Signals”, in Proc. Asilomar Conf. Signals, Syst., Comput., 2005, pp. 1537–1541.

[5] M. Duarte, M. Wakin, D. Baron, S. Sarvotham and R. Baraniuk, “Measurement Bounds for Sparse Signal Ensembles via Graphical Models”, IEEE Trans. Inf. Theory, Vol. 59, No. 7, 2013, pp. 4280-4289.

[6] D. P. Wipf and B. D. Rao, “An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem”, IEEE Trans. Signal Process., Vol. 55, No. 7, 2007, pp. 3704–3716.

[7] D. Baron, M. Wakin, M. Duarte, S. Sarvotham and R. Baraniuk, “Distributed Compressed Sensing”, Technical Report ECE-0612, Electrical and Computer Engineering Department, Rice University, 2009.

[8] S. F. Cotter, B. D. Rao, K. Engan and K. Kreutz-Delgado, “Sparse Solutions to Linear Inverse Problems with Multiple Measurement Vectors”, IEEE Trans. Signal Process., Vol. 53, No. 7, 2005, pp. 2477–2488.

[9] T. Wimalajeewa and P. Varshney, “Cooperative Sparsity Pattern Recovery in Distributed Networks via Distributed OMP”, in Proc. ICASSP, 2013, pp. 5288–5292.

[10] G. Li, T. Wimalajeewa and P. Varshney, “Decentralized Subspace Pursuit for Joint Sparsity Pattern Recovery”, in Proc. ICASSP, 2014, pp. 3365–3369.

[11] W. Chen and I. J. Wassell, “A Decentralized Bayesian Algorithm for Distributed Compressive Sensing in Networked Sensing Systems”, IEEE Trans. Wireless Commun., Vol. 15, No. 2, 2016, pp. 1282–1292.

[12] J. Hua and Ch. Li, “Distributed Variational Bayesian Algorithms Over Sensor Networks”, IEEE Trans. Sig. Process., Vol. 64, No. 3, 2016, pp. 783-798.

[13] H. Yang, L. Huang, H. Xu, and A. Liu, “Distributed Compressed Sensing in Wireless Local Area Networks”, International Journal of Communication Systems, Vol. 27, No. 11, 2013, pp. 2723-2743.

[14] N. Youness, and Kh. Hassan, “Energy Preservation in Large-Scale Wireless Sensor Network Utilizing Distributed Compressive Sensing”, IEEE 10th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), 2014.

[15] H. Dai, Y. Zhang, and J. Liu, “Structured Variational Methods for Distributed Inference in Networked Systems: Design and Analysis”, IEEE Trans. Sig. Process., Vol. 61, No. 15, 2013, pp. 3827-3839.

[16] K. M. León-López, L. V. G. Carreño, and H. A. Fuentes, “Temporal Colored Coded Aperture Design in Compressive Spectral Video Sensing”, IEEE Trans. Image Process., Vol. 28, No. 1, 2019, pp. 253-264.

[17] C. Chen, F. Ding, and D. Zhang, “Perceptual Hash Algorithm-Based Adaptive GOP Selection Algorithm for Distributed Compressive Video Sensing”, IET Image Process., Vol. 12, No. 2, 2018, pp. 210-217.

[18] L. W. Kang, and C. S. Lu, “Distributed Compressive Video Sensing. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. 2009, pp. 1169–1172.

[19] N. Yu, T. Qiu, F. Bi, and A. Wang, “Image Features Extraction and Fusion based on Joint Sparse Representation” IEEE Journal of Selected Topics in Signal Processing, Vol. 5, No. 5, 2011, pp. 1074–1082.

[20] J. Wei, L. Wang, P. Liu, X. Chen, W. Li, and A. Y. Zomaya, “Spatiotemporal Fusion of MODIS and Landsat-7 Reflectance Images via Compressed Sensing”, IEEE Trans. Geosc. Remote Sens., Vol. 55, No. 12, 2017, pp. 7126-7139.

[21] F. Li, Sh. Hong, and L. Wang, “A New Satellite Image Fusion Method Based on Distributed Compressed Sensing”, 25th IEEE International Conference on Image Processing (ICIP), 2018.

[22] Y. Lin, B. Zhang, H. Jiang, W. Hong, and Y. Wu, “Multi-Channel SAR Imaging based on Distributed Compressive Sensing”, Sci. China Inf. Sci., Vol. 55, No. 2, 2012, pp. 245- 259. https://doi.org/10.1007/s11432-011-4452-z .

[23] W. Kong, H. Li, and W. Zhang, “Compressed Sensing-Based Sparsity Adaptive Doubly Selective Channel Estimation for Massive MIMO Systems”, Wireless Comm. and Mobile Computing, Vol. 2019, Article ID 6071672, 10 pages, 2019. https://doi.org/10.1155/2019/6071672.

[24] D. Wu, W. P. Zhu, and M. Swamy, “Compressive Sensing-Based Speech Enhancement in Non-Sparse Noisy Environments”, IET Sig. Process., Vol. 7, No. 5, 2013, pp. 450–457.

[25] D. Liu, Q. Wang, Y. Zhang, X. Liu, J. Lu, and J. Sun, “FPGA-based Real-Time Compressed Sensing of Multichannel EEG Signals for Wireless Body Area Networks” Biomedical Sig. Process. and Control, Vol. 49, 2019, pp. 221-230.

[26] A. N. Uwaechia, and N. M. Mahyuddin, “Spectrum-Efficient Distributed Compressed Sensing Based Channel Estimation for OFDM Systems Over Doubly Selective Channels”, IEEE Access, Vol. 7, 2019, pp. 35072-35088.

[27] L. Prunte, “Application of Distributed Compressed Sensing for GMTU Purposes”, IET International Conference on Radar Systems, 2012.

[28] L. He and L. Carin, “Exploiting Structure in Wavelet-Based Bayesian Compressive Sensing”, IEEE Trans. Signal Proccess., Vol. 57, No. 9, 2009, pp. 3488–3497.

[29] R. Torkamani and R. A. Sadeghzadeh, “Bayesian Compressive Sensing using Wavelet Based Markov Random Fields”, Signal Processing: Image Communication, Vol. 58, 2017, pp. 65-72.

[30] S. Mallat, A Wavelet Tour of Signal Processing, 2nd ed. New York: Academic, 1998.

[31] M. F. Duarte, M. B. Wakin and R. G. Baraniuk, “Wavelet-Domain Compressive Signal Reconstruction using a Hidden Markov Tree Model”, in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2008, pp. 5137–5140.

[32] M. S. Crouse, R. D. Nowak and R. G. Baraniuk, “Wavelet-Based Statistical Signal Processing using Hidden Markov Model”, IEEE Trans. Signal Process., Vol. 46, 1998, pp. 886–902.

[33] Sh. Ji, Y. Xue and L. Carin, “Bayesian compressive sensing”, IEEE Trans. Signal Process., Vol. 56, No. 6, 2008, pp. 2346-2356.

[34] M. Beal, “Variational Algorithms for Approximate Bayesian Inference,” Ph.D. thesis, Univ. College of London, London, U.K., May 2003.

[35] W. Chen, M. Rodrigues and I. Wassell, “A Fréchet Mean Approach for Compressive Sensing Date Acquisition and Reconstruction in Wireless Sensor Networks”, IEEE Trans. Wireless Comm., Vol. 11, No. 10, 2012, pp. 3598 –3606.

[36] X. Chen, Y. Zhang, and R. Qi, “Block Sparse Signals Recovery Algorithm for Distributed Compressed Sensing Reconstruction”, Journal of Inf. Process. Systems, Vol. 15, No. 2, 2019, pp. 410-421.

[37] P. Bodik, W. Hong, C. Guestrin, S. Madden, M. Paskin and R. Thibaux. (2004) Intel lab data. [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html.

[38] http://www.sandia.gov/radar/sar-data.html.

[39] Z. Wong, A.C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity”, IEEE Trans. Image Process., Vol. 13, No. 4, 2004, pp. 600-612.

[40] E. Aguilera, M. Nannini and A Reigber, “Multisignal Compressed Sensing for Polarimetric SAR Tomography”, IEEE Geoscience and Remote Sensing Letters, Vol. 9, No. 5, 2012, pp.871-875.

Razieh Torkamani received the B.S. and M.S. degrees from the Iran University of Science and Technology (IUST), Tehran, Iran, in 2010 and 2012, respectively. She is currently working toward the Ph.D. degree in the faculty of Electrical Engineering and Computer Science, in the K.N. Toosi University of Technology. Her research interests include compressive sensing and recovery algorithms, statistical signal processing, sparse signal representations, and graphical models.

Professor R.A. Sadeghzadeh is a full professor of Communications Engineering at the faculty of Electrical Engineering of the K.N. Toosi University of Technology in Tehran, Iran. He received his B.Sc. in 1984 in telecommunication Engineering from the K.N. Toosi, University of Technology and M.Sc. in Digital Communications Engineering from the University of Bradford and UMIST (University of Manchester Institute of Science and Technology), UK as a joint program in 1987. He received his Ph.D. in electromagnetic and antenna from the University of Bradford, UK in 1990. Professor Sadeghzadeh worked as a Post-Doctoral Research assistant in the field of propagation, electromagnetic, antenna , Bio-Medical, and Wireless communications from 1990 till 1997 at the University of Bradford, UK. From 1984 to 1985 he was with Telecommunication Company of Iran (TCI) working on Networking. Since 1997 He is with the K.N. Toosi University of Technology working with Telecommunications Department at the faculty of Electrical Engineering. He has published more than 200 referable papers in international journals and conferences. Professor Sadeghzadeh current interests are numerical techniques in electromagnetic, antenna, propagation, radio networks, wireless communications, nano-antennas and radar systems.

Reliable resource allocation and fault tolerance in mobile cloud computing

Zahra Najafabadi Samani
Department of Computer Architecture, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran
najafabadizahra@gmail.com
Mohammad Reza Khayyambashi
Department of Computer Architecture, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran
M.R.Khayyambashi@eng.ui.ac.ir

Received: 08/Jul/2019 Revised: 11/Nov/2019 Accepted: 24/Dec/2019

Abstract

By switching the computational load from mobile devices to the cloud, Mobile Cloud Computing (MCC) allows mobile devices to offer a wider range of functionalities. There are several issues in using mobile devices as resource providers, including unstable wireless connections, limited energy capacity, and frequent location changes. Fault tolerance and reliable resource allocation are among the challenges encountered by mobile service providers in MCC. In this paper, a new reliable resource allocation and fault tolerance mechanism is proposed in order to apply a fully distributed resource allocation algorithm without exploiting any central component. The objective is to improve the reliability of mobile resources. The proposed approach involves two steps: (1) Predicting device status by gathering contextual information and applying TOPSIS to prevent faults caused by volatility of mobile devices, and (2) Adapting replication and checkpointing methods to fault tolerance. A context-aware reliable offloading middleware is developed to collect contextual information and manage the offloading process. To evaluate the proposed method, several experiments are run in a real environment. The results indicate improvements in success rates, completion time, and energy consumption for tasks with high computational loads.

Keywords: Mobile Cloud Computing; Fault tolerance; Reliability; Replication; Checkpointing.

1- Introduction

As a result of the recent developments in mobile technologies, mobile devices (e.g., smartphone and tablet PC) have become an integral part of human life as highly effective and convenient means of communication. However, mobile devices face an array of challenges in terms of both resources (e.g., battery life, storage, and bandwidth) and communications (e.g., mobility and security). To overcome this limitation, offloadding computations from mobile devices to cloud is proposed [1, 2]. A three-tier architecture is defined for Mobile Cloud Computing (MCC) including remote cloud servers, local servers known as cloudlets, and adjacent mobile devices. Offloading to remote servers can be costly and introduces latency. Besides, they are not always available and cloudlets limit mobility of mobile devices [1, 3-6]. To address this issue, in this paper, the third tier of the MCC (consist of neighboring mobile devices) are consider where both service requester and service providers are mobile devices.

However, there are several issues in mobile devices as resource providers such as unstable wireless connection, limited energy capacity and frequent changes of location. Thus, fault and fault tolerance are among the main challenges faced by mobile resource providers, which should not be ignored. Faults in the offloading process are primarily caused by energy constraints, mobility, and availability associated with mobile devices. Many previous works try to select appropriate resource providers among available mobile devices to allocate tasks by gathering contextual information [6-8]. However, they mostly fail to address fault prevention and tolerance. Others consider only prevention of fault [3, 13, 18-21] or fault tolerance [16, 14]. Moreover, limited energy has not been investigated as a factor of fault prevention [11-13] or only replication techniques are applied in fault tolerance [14-17]. It is well noted that replication techniques are very costly in high computational load. The objective here is to maximize the success rate of the offloading process so that some Quality of Service (QoS) constraints are satisfied. In this paper, a fully-distributed reliable resource allocation algorithm is used wherein, energy, mobility and availability of mobile devices are considered as fault factors. This approach promotes system robustness by predicting the states of mobile devices to avert the faults caused by mobile volatility. Then depending on the size of the task, checkpointing or replication methods are used as the fault tolerance methods. In order to provide dynamic and accurate resource allocation and to predict the states of mobile devices, contextual information needs to be gathered from devices, applications, and the environment to be applied in decision-making. A context-aware offloading middleware is developed to collect contextual information and manage the offloading process. To evaluate this new proposed method, our experiments are run in a real environment.

The rest of the paper is organized as follows: Section 1 reviews the related works. In Section 2, a reliable system is formulated and solved using a two-stage approach. Section 3 pertains to ranking and classification of mobile devices using TOPSIS. In Section 4, the architecture of the context-aware reliable offloading middleware is described. Section 5 presents the evaluation results of the proposed approach. Finally, the paper is concluded in Section 6.

2- Related work

There are many studies that investigate the task allocation problem by collecting contextual information in MCC [6-8]. However, those do not consider fault and fault-tolerant methods. Shi et al. [7] propose Serendipity which enables offloading through gathered information profile and releasing two versions of task allocation algorithm, energy aware Serendipity and time optimizing. [6] considers the local tier of mobile cloud where both service requester and service providers are mobile. The paper proposes a resource allocation algorithm with multi-objective optimization to minimize completion time and energy consumption of all participating mobile devices. Ref. [8] considers all three layers of MCC and proposes a context-aware offloading decision algorithm to derive an optimal offloading decision under the context of the mobile device and cloud resources.

Several studies consider fault in MCC. However, some of them consider only prevention of fault and none of fault tolerance method (e.g. replication or checkpointing) is adopted after failure occurrence [3, 12, 13, 18-21].

In [18, 19], a monitoring approach based on Markov chain are proposed for analyzing and predicting states of resources. They propose a monitoring time interval rate in order to monitor the correct state information of mobile resources. In these papers, the manner of applying monitoring information in fault is missed. In the context of mobile cloud, Ref. [20] considers mobile devices as resources and improves mobile resource efficiency through energy-aware management by selecting appropriate mobile devices. In this scheme, firstly devices are divided into four groups in terms of efficiency and mobility based on a threshold value. The devices in each group are then ranked. To prevent faults, no tasks are assigned to unreliable mobile devices. In [13, 21], a dynamic grouping scheme is presented to manage mobile devices in MCC. In [13] availability and mobility, and in [21] mobility and efficiency are considered as fault factors. Then, cut-off points are adopted by entropy. Next, according to the cut-off points, mobile devices are arranged into several groups. In these works, mobile devices' energy is not considered as fault factor in grouping. In all these papers [13, 18-21], a central component is applied to monitor and manage mobile devices that cause a single failure and limit the mobility of mobile devices.

Ref. [12] proposes a reliable resource allocation method according to availability and mobility of mobile devices in MCC. Initially, subtasks are assigned to mobile devices with minimal mobility and then resources with the highest availability are selected. Despite being a limitation, energy is not considered as a fault factor. In [6], a context-aware offloading scheme for mobile peer-to-peer environments is proposed in which both servers and clients are mobile. Their scheme chooses adjacent reliable mobile devices to assign subtasks in order to prevent fault. This dissertation just supports fault occurrence by mobility and ending energy of mobile.

On the other hand, in [14-17] fault tolerance is achieved using only replication which is generally not very efficient for task with high computational loads, imposes high costs, and occupies many resources. In addition, in [16, 14], to prevent fault occurrence, the system does not select adjacent reliable mobile devices to assign subtasks. In Hyrax [16], a Hadoop-based platform is proposed that supports cloud computing on smartphones. Replication is used for fault tolerance; subsequent to failure, failed subtasks are re-executed without user intervention. Although Hyrax provides high scalability, the system exhibits poor performance with CPU-intensive tasks. The central server in Hyrax causes a bottleneck in the system and limits the mobility of mobile devices. Ref.[14] presents the implementation of a platform for providing fault tolerance in MCC. Their approach improves reliability by the use of a dynamic and adaptive replication which uses the minimum number of replicas. In this model, a new replica is placed on the most reliable node until the required reliability level is reached.

Ref. [15] considers fault tolerance and quality of service in social mobile cloud computing environment by using Content Addressable Network. In this paper the cloud server selects the best resource from the adjacent mobile devices in terms of quality of public and mobile device services and, replication is used for fault tolerance. Nevertheless, energy and mobility are not considered as a fault factor in the paper. In [17], a framework is proposed to support fault tolerance which integrates the k-out-of-n reliability mechanism into mobile cloud formed only by mobile devices. The framework provides services for applications that aim to reliably store and process data in the mobile cloud such that the energy consumption for retrieving and processing the data is minimized. An

unrealistic assumption in this work is that the nodes have equal energy consumption for processing. It is possible that subtasks are allocated to nodes with high-energy consumption, so these nodes will fail in the future.
[9] applied the fault tolerance technique to handle the faulty VMs in the MCC using the human disease resistance mechanism which identifies the faulty virtual machines and reschedules the tasks to the identified suitable virtual machines. In this paper, they do not consider fail in mobile devices as resources.

In [11, 10], a dynamic classification scheme for reliable management of mobile devices as resources in MCC is used. In [11] mobile devices are grouped according to availability and mobility by adopting entropy, and in [10] mobile devices are classified into groups based on their processing capacity, availability, and communication condition by adopting tree learning. Then, checkpointing and replication are applied to groups with high and low reliability, respectively. However, in these works, administrative tasks are handled by a proxy which is in conflict with nature of mobile networks and mobility of mobile devices. Moreover, task assignment has not been determined. In addition, in [11] energy is not considered as a fault factor, and [10] considers only the remaining battery of the devices and study of mobility model and trajectory is overlooked.

3- Proposed Reliable System Model

Reliability is one of the most important aspects in mobile cloud computing due to mobility and resource constraints of mobile devices. Faults and failures should be managed and controlled in an active manner to minimize the effects of failures on the system. There are two complementary approaches to establish reliability: (1) Fault Prevention and (2) Fault Tolerance [23].
In the proposed model, each task has n independent subtasks $\{T= t_i \ |1\leq i\leq n\}$. And the mobile cloud environment is formed by h mobile device $\{S=s_k| 1\leq k\leq h\}$(where s_h is a client mobile device that does task offloading). In this model, at any given time, only one offloading is executed, applications are segmented before, and offloadable subtasks are independent and executed in parallel.

3.1- Fault Prevention

The goal of fault prevention is to prevent system failure by ensuring that all possible causes of unreliability are removed [23]. One of the challenges in mobile cloud is selecting reliable resources from adjacent mobile devices to assign tasks to and prevent system failure. The energy, mobility and availability factors of mobile devices result in more frequent system faults. Due to dynamic changes which disrupt mobile device applications, these features increase device load and decrease system performance. To the best of our knowledge, these factors have not been jointly addressed. In this paper, we seek to prevent system failure by selecting reliable devices according to three important factors which cause system faults: energy, mobility and availability.

3.1.1- Energy

Battery energy is among the main constraints in mobile devices. As battery life is inherently limited, it may run out at any moment, causing premature job termination. Hence, mobile devices must preserve battery power during and after processing a job. In mobile cloud environments, subtasks should be distributed in a manner that network lifetime is maximized and fault caused by battery energy limitations are avoided. This proposed energy model is inspired by [6]. In This model energy consumption and remaining energy levels of all nodes involved in offloading are considered. Variable notations are shown in Table 1.

Table 1. Definitions of notations.

n	Number of subtasks
h	Number of mobile devices
R_j	jth resource provider
t_j	Time needed to execute subtasks on Rj
e_j	Energy consumption per second running each subtask on Rj
et_j	Energy consumption on Rj to transfer one unit of data
$E0_j$	Initial energy level of Rj
Vin	Input size of subtask
$Vout$	Output size of subtask
b_j	Number of subtasks on Rj
E_j	Energy consumption on Rj

The energy consumptions of both the server and the client are indicated through Equ. (1). In the client side ($j = h$), energy consumption includes energy consumption of running local subtasks, transferring offloaded subtasks to nearby mobile devices, running the resource allocation algorithm. The energy consumption in server sides consists of resource provider Rj ($1\leq j<h$) consists of energy consumed for running the assigned subtasks and transmitting the results.

$$predicted\ E_j = \begin{cases} b_j * (t_j * e_j) + \sum_{j=1}^{h-1} b_j * (et_j * vin) + algE, & j = h \\ bj * [(t_j * e_j) + (et_j * vout)], & 1 \leq j < h \end{cases} \quad (1)$$

Then, Equ. (2) is applied to avoid allocating subtasks to nodes with low remaining energy. Here, the network's lifetime increases and the risk of energy depletion during processing is prevented.

$$E0_j - predicted\ E_j \geq \alpha \quad for\ all\ j= 1, \dots, h-1 \qquad (2)$$

Next, Equ. (3). allows the subtasks to be allocated to the nodes according to their energy level:

$$Energy_j = \frac{predicted\ E_j}{E0_j - predicted\ E_j} \quad (3)$$

In this model, in addition to taking care of nodes with low energy levels, the subtasks are fairly distributed among other nodes. We aim to maximize the lifetimes of nodes having the low remaining energy and high energy consumption rates. Here, network energy is reduced and no task is assigned to nodes with low remaining energy, which prevents failures caused by battery drain and increases minimal remaining energy of nodes. In [17], only remaining energy and transferred energy among nodes is considered while energy requirements are assumed to be equal for all nodes, which is not a realistic assumption.

3.1.2- Mobility Factor

Mobile devices as resource providers can join and leave the Mobile Cloud environment in an unpredictable manner. This interrupts the operation and may cause a system failure. Given their low reliability, it is difficult to consider mobile devices as resources. In this study, prediction colocation between client and server mobile devices is used to avoid the failure by mobility during offloading. In fact, a colocation time between two users is the time that two users visit each other and stay together. Many studies use a temporal mean for predicting colocation time [13, 21, and 22]. In this technique, time is split in slots, and a temporally varying mean of the encounters with other user is kept. Under this assumption, the estimated colocation time is accurate since user paths are well defined. However, it may present issues when applied to new environments. To overcome this issue, in this paper, the spatial-temporal mean is adopted to predict colocation time between server and client that follows the same principle, but taking into consideration the activity and place when the encounter between users take place. The anticipated colocation and movements of users are based on intermittent user behaviors to see a place and move among locations. Users periodically see a specific location and regularly stay there, allowing their routes and stationary time intervals to be estimated [3, 13, 22, and 24]. Here, a sequence of places *p* with time stamps *ten* and *tex* is stored for each user. The tuple $<p_i, ten_i, tex_i>$ indicates that user *u* enters place p_i (*p* is the place covered by one cellular tower or AP) at time ten_i and exits at time tex_i. A trajectory information group G_k is generated which contains *k* tuples. Each time the user moves from a location to another, the trajectory information is updated. This sequence is expressed by Equ. (4).

$$G_k = \{<p,\ ten, tex>|<p_{i-k},\ ten_{i-k}, tex_{i-k}>, \ldots, <p_{i-1},\ ten_{i-1}, tex_{i-1}>, <p_i,\ ten_i, tex_i>\} \quad (4)$$

In the proposed scheme, first, client device movements are predicted. Markov chain is perhaps the most widely used model for human mobility due to its simplicity and efficiency. Many recent works adopt the approach for location prediction [25-28]. A Markov chain is a sequence of random variables X_t, which represent the states of a system, provided that the current state (X_t) depends only on the previous state (X_{t-1}) [29]. This could be expressed through Equ. (5).

$$\mathrm{P}(X_t = x | X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) = \mathrm{P}(X_t = x | X_{t-1} = x_{t-1}) \quad (5)$$

The chain could also be depicted using a directed graph, whose edges are labeled with the probability of going from one state to another, from time *t - 1* to time *t*, Fig. (1).

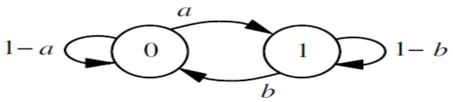


Fig 1.Markov chain model

In this study, the probability of going from location *i* to location *j* is calculated according to [28]. Here, locations are considered as states of the Markov chain. As mentioned earlier, in a Markov chain, the states are independent; in other words, the location subsequent to *i* is not dependent on the history of visited locations. Accordingly, at processing time, the probability of a transition from *i* to *j* equals the number transitions from *i* to j at processing time is divided by the total number of outgoing transitions from *i* (i.e. the definition of probability). Respective probabilities are calculated on the client mobile device (during processing) by applying Equ. (6).

$$\mathrm{P}\big(l_{t+1} = p_j | l_t = p_i\big) = \frac{\mathrm{n}(l_{t+1} = p_j | l_t = p_i)}{\sum_{d=1, d\neq i}^{k} n(l_{t+1} = p_d | l_t = p_i)} \quad (6)$$

The location with maximum probability is selected as the next location of the client during the process. The neighbor nodes are then informed of the client's next location. Adjacent nodes that are willing to cooperate, calculate their own colocation probabilities with the client node. Equ. (7) is applied to calculate the colocation between the service provider and the client after it is determined that the client node is going to stay in its current location during the process.

$$\mathrm{P}(l_{t+1} = p_i | l_t = p_i \,, tex_i - \mathrm{tc} \geq \mathrm{tre}) = \frac{\mathrm{n}(l_{t+1} = p_i | l_t = p_i \,, tex_i - \mathrm{tc} > \mathrm{tre})}{\sum_{d=1,d\neq i}^{k} n(l_{t+1} = p_d | l_t = p_i\,)} \qquad (7)$$

In this equation, according to the characteristic of Markov chain and the definition of probability, the number of times the service provider visits the current location at processing time for *tre* is divided by total number of times it moves from *i* to any other location at processing time. Here, *tre* is the time required for processing. The variable *tc* denotes current time. The non-equation $tex_i - tc > tre$ indicates that service providers need to remain in the vicinity of the client as long as the task continues to execute.

Once it is determined that the client node intends to move to another location during the process, Equ. (8) Calculates the colocation between the service provider and the client. To find the colocation probability, the number of times the service provider has moved from *i* to *j* (along with the client during the process) is divided by the number of its movements from *i* to any other location. The non-equation $(tex_i - \mathrm{tc}) + (tex_j - ten_j) \geq \mathrm{tre}$ means that the colocation time must exceed the time necessary for processing.

$$\mathrm{P}\big(l_{t+1} = p_j \big| l_t = p_i \,, (tex_i - \mathrm{tc}) + (tex_j - ten_j) \geq \mathrm{tre}\big) = \frac{\mathrm{n}\big(l_{t+1} = p_j \big| l_t = p_i \,, (tex_i - \mathrm{tc}) + \big(tex_j - ten_j\big)}{\sum_{d=1,d\neq i}^{k} n(l_{t+1} = p_d | l_t = p_i\,)} \qquad (8)$$

Finally, service providers send their corresponding colocation probabilities to the client node. In this manner, the best resources from the adjacent mobile devices are automatically chosen. The best resources have the most chance of staying in colocation.

3.1.3- Availability

Availability can have several meaning according to the system requirements. In the context of mobile devices used as resources, it refers to the probability that a mobile device performs and replies correctly while acting as a resource. High availability systems aim to minimize downtime and repair costs. In mobile applications, availability is a crucial requirements and an achievable objective. In the context of MCC or mobile grid, mobile devices are often classified according to availability [11-13]. In [11], the number of faults caused by a mobile device is used for availability information. Accordingly, in this paper the success rate is applied to obtain availability. The higher values in the past predict greater availability and success in the future. Each service provider calculates its availability using Equ. (9). and the resulting values are transmitted to the client mobile device.

$$Availability = \frac{number\ of\ successful\ tasks}{\text{total number of tasks}} \qquad (9)$$

The client then selects the devices with the highest success rate (i.e. availability) as the resources. Thus, tasks are assigned to devices that are more reliable and devices with high failure rates are avoided.

3.1.4- Estimating context information

There are several techniques to predict runtime, energy consumption, and output size of a subtask on a mobile device [32, 33]. We use a dual profiling approach inspired by [33], which consists of a peer-centric and a task-centric profile. The former is a history-based profile maintaining the runtimes and energy consumptions of the last *n* runs of a subtask on a specific peer. Here, the average profile data from the last ten runs is calculated as an estimate. Since the mobile cloud is a dynamic environment and there may be new to which the task has never been assigned, a task centric profile is used for new devices. In this approach, a device is selected as the base; then, by comparing the processing power of the base device with that of the new device, subtask energy consumption and runtime on the new device are estimated.

3.2- Fault Tolerance

A fault-tolerant system should be able to manage defects in hardware or software components as well as other unexpected downtimes. Despite fault prevention approaches, failure is inevitable; therefore, it is necessary to take appropriate measures after system fault [23]. There are numerous methods for achieving fault tolerance in distributed systems, among which replication and checkpointing are most popular.

3.2.1- Active Replication

Replication techniques replicate similar tasks, which can be run in a simultaneous manner on several devices. When one of the replicas fails another performs its task [30]. In this paper, active replication is used to promote resources reliability against faults caused by the volatility of mobile devices for tasks with low computational load. The reasons for this choice are: (1) active replication is a non-centralized technique which suits the fully distributed design of the proposed system; (2) Failures are fully hidden from the clients, since requests are still processed even if one replica fails; (3) short response time even in case of failures because each replica works independently which is important for mobile users; and (4) simplicity which is important for mobile devices with limited resources.

In the proposed method once a service provider fails for any reason, another replica is responsible for the client node requests. When the client node receives the first response from a replica, others halt processing in order to reduce energy consumption. Although replication techniques have many advantages with high fault tolerance, they are costly for tasks with high computational loads.

3.2.2- Independent Checkpointing

Once a failure occurs, it is vital to restore the process to its correct state. So, in this paper, independent checkpointing is applied for tasks with high computational load. This is better for independent processes and reduces the computational overhead [31]. Independent checkpointing mainly involves restoring the system from its present erroneous state back into a previously correct state. This requires the state of the system to be occasionally recorded so that, when faults occur, the state can be restored. The state of the system is stored on the client mobile devices at regular intervals. When a subtask fails, that subtask along with the previously correct state, is assigned to another reliable service provider without user intervention.

3.3- Reliable task allocation in the Mobile Cloud

In the mobile cloud, when a client mobile device wants to run a compact application, it first asks for an offloading service. Next, it applies service discovery to identify adjacent mobile devices by transmitting a broadcast message containing its location during the service. As the service provider receives the offloading request message, it calculates the energy consumption ratio in relation to residual energy after performing the subtask through Equ. (3). Then, it calculates its colocation probability with the client and availability using Equs. (7-8) and Equ. (9), respectively. Finally, it sends this information along with its energy consumption rate and current energy in response to the client. Algorithm 1 outlines the execution algorithm of the subtasks on the service provider.

Algorithm 1: Execution Subtasks Algorithm

```
Numet=0;//number of all execution tasks
Numset=0;//number of all successfully executed tasks
Collect location information;
Receive request from client;
If mobile device is willing to cooperate then
   Collect energy information;
   Calculate context information();
   Send context-inf to client;
   Receive subtasks;
   If subtask is large then
     While(Receive request for checkpoint)
        Send checkpoint;
     If chechpoint $\neq 0$ then
        Execute subtask from checkpoint;
        Send result;
   Else
     Execute subtask;
     Send result;
   Receive ack from client;
   If ack==1 then
       Numet++;
       Numset++;
     Else
       Numet++;
 Function Calculate context information()
   Calculate colocation probability with client(p-location);
   Calculate consumable battery(cb);
   Calculate ratio of consumable battery to remaining battery(E);
   Calculate availability;
   Context-inf set(p-location, E, current E,cb, availability);
 End
```

The client discovers its adjacent mobile devices and obtains their context information through messages sent by adjacent mobile devices in response to the client. First, the service providers' remaining energy levels are checked. If the levels are lower than the threshold from Equ. (1), the client does not assign any subtasks to these service providers. Next, the client predicts the providers' behaviors by using their context information. They are ranked accordingly by applying Equ. (10-18). The service providers are then partitioned into two groups of high and low reliability. Finally, in order to take advantage of fault tolerance techniques, according to the computational load of the task, either replication or checkpointing is adopted. For tasks with low computational load, active replication is employed and the subtasks in several groups with the highest reliability are replicated in a simultaneous manner.

Subtasks are allocated to groups according to the rank of devices through Equ. (13). More subtasks are assigned to devices with higher ranking. Every service provider executes the assigned subtasks and replies the results to the client. Once a subtask execution is finished on one of the replicas and the result is transmitted to the client, the subtask is no longer executed on other replicas. This contributes to saving resources and reducing traffic on the network. Replication in several devices for tasks with low computational load is not costly and occupies few resources; however, in such cases, restarting a subtask and taking checkpoints are of high overhead and cost. The decision-making algorithm on the client side is presented in Algorithm 2.

Algorithm 2: Dynamic Reliable Context-aware Decision making Algorithm

```
Function main()
  Collect location information; //call monitoring
  Request offloading();
  Decision offloading();
  Receive result;
  While  not receive result of all subtask do
  Request offloading;
  Receive result;
  Merge all results;
End
Function Request offloading()
  Calculate maximum job execution time ;
  Calculate next location in maximum job execution time ;
  Send broadcast with next location;
  While (true) do
    Receive reply from vicinal mobile devices;
    Context ← Set context(energy, location,  availability, id);
```

```
   For i=1 to all vicinal device do
     Rank ◄— TOPSIS(context);
     If Rank ≥ 0.5 then
       Group1 ◄— mobile device;
     Else
       Group2 ◄—mobile device;
   In Group1
      Sort in ascending order of Rank;
   In Group2
      Sort in ascending order of Rank;
End
Function Decision offloading()
    If subtasks is small then
      Call Replication;
    Else if subtasks is large then
      Call check pointing;
End
```

Replication is very costly for tasks with high computational load and occupies many resources. Consequently, in this paper, checkpointing is applied for tasks with high computational load. In checkpointing, first, subtasks are assigned to the group with the highest reliability according to device rankings. Next, at regular intervals, system state is stored on the client mobile device. Given the dynamic nature of the mobile cloud environment, after a device fails, client mobile device discovers its adjacent mobile devices again and receives their context information. Then once more, the client ranks and classifies adjacent mobile devices where the failed subtask with previously correct state is assigned to another reliable service provider without user intervention. Due to the high computational load of the subtask, the system replaces an erroneous state with an error-free state. Thus, the new service provider does not need to start from the beginning to process the failed subtask. This proposed approach saves time and cost of resources if a fault occurs in the final moments of processing.

Once the client receives a result from a service provider, it replies with an acknowledgment message to the sender, who then increments its number of successful tasks. Contrarily, if a subtask fails, it is re-executed on other adjacent mobile devices without user intervention. Ultimately, the client collects and merges all the results.

4- Ranking and grouping mobile devices with TOPSIS

In this study, TOPSIS is applied to rank and group mobile devices [34]. There are two reasons for this choice: (1) the concept of TOPSIS is rational, and (2) its algorithm is simple and light which is suitable for mobile devices with limited resources. In addition, TOPSIS can take objective weights into consideration in the comparison process. The technique includes a number of options and attributes for decision making. The options must be ranked according to these attributes. This solution procedure takes the following seven steps:

- **Step1.** Preparing the decision-making matrix with p rows and three columns where r_{pq} are the elements of this matrix, Equ. (10). The rows represent mobile devices in the vicinity of the client device while the columns represent the decision making attributes. Three attributes, namely energy consumption to the remaining energy, colocation probability between the service provider and the client and availability, are considered for ranking mobile devices.

$$D = \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_q \end{array} \\ \left[\begin{array}{cccc} r_{11} & r_{12} & \cdots & r_{1q} \\ r_{21} & r_{22} & \cdots & r_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pq} \end{array}\right] \end{array} \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_p \end{array} \qquad (10)$$

- **Step2.** In this step, the decision-making matrix (D) is normalized by using Equ. (11).

$$n_{ij} = \frac{r_{ij}}{\sqrt{\sum r_{ij}^2}} \qquad i = 1,2,\dots,p \ , \qquad j = 1,2,\dots,q \qquad (11)$$

- **Step3.** Each attribute is assigned a weight based on the concept of Shannon entropy. Entropy is a major concept in the information theory and represents the amount of unreliability in a discrete probability distribution (random variable) [35]. Lower weights are assigned to attributes with identical values. This is because such attributes do not contribute to distinguishing options. They are, thus, less prominent. For each attribute, entropy is defined through Equ. (12).

$$E_j = -\frac{1}{\ln r} \sum_{i=1}^{p} n_{ij} \, ln \, n_{ij} \qquad (12)$$

The degree of divergence dj (the contrast intensity of each attribute) and the weight for each attribute is indicated through Equ. (14) and (13), respectively.

$$d_j = 1 - E_J \qquad (13)$$

$$W_j = \frac{d_j}{\sum_{k=1}^{q} d_k} \qquad (14)$$

- **Step4.** This step involves the calculation of the weighted normalized decision matrix using Equ. (15).

$$v_{ij} = n_{ij} * w_j \quad i = 1,2,\dots,p \ , \quad j = 1,2,\dots,q \qquad (15)$$

- **Step5.** In this step to rank alternatives, the solutions are compared with the positive ideal solution (A^+) and the negative ideal solution (A⁻), Equ. (16) (J and J' are the index sets of the benefit and cost attributes, respectively)

$$A^+ = \{(\min v_{ij} | j \in J), (max v_{ij} | j \in J') \,|\, i = 1,2,\dots,p\} = \{v_1^+, v_2^+, \dots, v_q^+\}$$
$$A^- = \{(\text{man}\, v_{ij} | j \in J), (min v_{ij} | j \in J') \,|\, i = 1,2,\dots,p\} = \{v_1^-, v_2^-, \dots, v_q^-\} \quad (16)$$

- **Step6.** The distances of each solution from the ideal solution (d_i^+) and negative ideal solution (d_i^-) is calculated through Equ. (17).

$$d_i^+ = \sum_{j=1}^{3} (v_{ij} - v_j^+)^{\frac{1}{2}} \quad i = 1,2,\cdots,p$$
$$d_i^- = \sum_{j=1}^{3} (v_{ij} - v_j^-)^{\frac{1}{2}} \quad i = 1,2,\cdots,p \quad (17)$$

- **Step7.** This step involves calculating ranks i.e. relative closeness of a solution to the ideal solution. The devices are ranked according to Equ. (18).

$$CL_i = \frac{d_i^-}{d_i^- + d_i^+} \quad i = 1,2,\cdots,p \quad (18)$$

Once mobile devices are ranked according to the aforementioned criteria, each device is placed in a group of high or low reliability based on its rank. Given the dynamic nature of mobile devices, their context information is continuously updated. Thus, devices are dynamically ranked and grouped. The number of subtasks assigned to each mobile device is determined according to its rank and reliability, Equ. (19).

$$nT_{R_i} = n \times \frac{CL_i}{\sum_{j=1}^{m} CL_i} \quad i = 1,\cdots,m \quad (19)$$

5- Context-aware reliable offloading middleware

In this section, a middleware is designed and implemented to employ reliable offloading and apply fault tolerance methods in the mobile cloud. This middleware collects contextual information, manages offloading steps, and adopts fault tolerance methods.

5.1- Architecture

This middleware consists of a client-side, which requests the offloading service, and a server-side, which provides services. A mobile device runs both parts, Fig. (2).

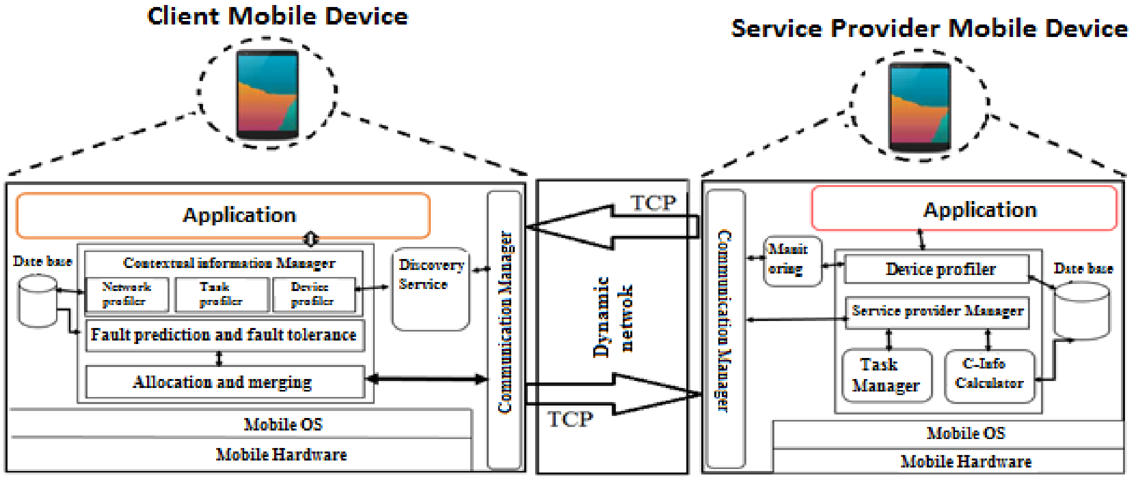


Fig 2. Context -aware reliable offloading middleware

5.1.1- Client-Side Middleware

The client-side middleware includes a discovery service, a contextual information manager, fault manager, Allocation and Merging, and a communications manager.

• Discovery Service: Adjacent mobile devices can be discovered using either pull or push techniques [18]. The former has a relatively small monitoring overhead, because the server's resource information is only requested when it is needed. Consequently, given resource constraints in mobile devices, and to reduce monitoring overhead, we propose a monitoring scheme which relies on the pull model. In this technique, when a client mobile device decides to offload a task to adjacent mobile devices, it sends a broadcast message, to which willing devices respond by transmitting their contextual information. This allows the client to discover its adjacent mobile devices and obtain their contextual information.

- Contextual information manager: This component gathers contextual information pertaining to tasks, resource providers, and the network through their aggregators i.e. task profiler, device profiler and network profiler, respectively.

a) The device profile contains energy consumption as well as remaining battery energy, trajectory information, total number of tasks assigned to the device, and the number of successfully performed by the device.

b) Task profile contains information about runtime and input/output size of each subtask. In this paper, it is assumed that application developers present offloadable parts of the task in the task profile.

- Fault manager: Initially, this component detects adjacent unreliable mobile devices; next, adjacent mobile devices are ranked according to their reliability and partitioned into groups of high or low reliability. Finally, in order to take advantage of fault tolerance techniques due to computational load of the task, either replication or checkpointing approach is applied.
- Allocation and Merging: this component assigns subtasks to adjacent mobile devices according to the fault manager component. In addition, this component merges the results. If a subtask fails or this component does not receive any result, the failure is reported to the fault manager component.
- Communications Manager: The Communications manager on the client provides network communication among client and service provider devices while monitoring this communication. In case the connection is disconnected, this component notifies the client to take appropriate fault tolerance measures depending on the condition.

5.1.2- Server-Side Middleware

This part of middleware includes a device profiler, a task manager, a context information calculator, a service provider manager, and a communications manager.

- Device profiler: Contextual information from the devices is collected and stored in a database on the device.
- Context information calculator: The three relevant criteria (i.e. colocation probability, energy ratio, and availability) are calculated based on the information received from the client and sent to the service provider manager component.
- Task manager: The component handles the request, executes the offloaded code, and relays the results to the service provider manager component.
- Service provider manager: It is responsible for coordinating the components on the service provider and following up processing on the server. Finally, this component collects the results.
- Communications Manager: The component handles communication between client and server on the server side, receives data and control data from the client, and sends context information and result to client.

6- Evaluation

In this section, the performance of the proposed system is evaluated by conducting real experiments on multiple mobile devices. A middleware is implemented into a library on Android operating system, which can be added in Android application for development. A face detection application is applied as the case study; it analyzes an assortment of photos as subtasks to identify the comprising faces. The application is implemented for the Android platform by applying android. media [36]. The middleware consists of approximately 6500 lines of Java code, which is used in following extensive experiments. In addition to the networking components, collecting and receiving checkpoints, collecting contextual information and tasks execution on this versions are run in separate threads. To evaluate the performance of the proposed method, a testbed of mobile devices is set up as show in Table 2. The mobile devices, which is creating a mobile cloud, are connected to an ad-hoc network using Wi-Fi with a mean bandwidth of 1.86 MBps. For each device, energy levels required for executing various tasks, the location of the device at any time, and availability are stored on its database. To measure the energy consumption of smartphones, we take advantage of PowerTour [37]. Furthermore, real-time device coordinates are obtained via GPS.

To demonstrate the performance of our proposed method, the proposed algorithm is compared with four other algorithms with different numbers of subtasks: random allocation, reliable allocation, replication, checkpointing algorithms, and dynamic grouping in [11]. In random allocation, mobile devices are randomly selected as resources without any measure of fault tolerance. Reliable allocation just aims to choose reliable mobile devices as resources and to assign tasks with no fault tolerance. The difference in the checkpointing algorithm lies in the application of checkpointing for fault tolerance. Another option is to use replication for fault tolerance, like [16, 14]. In these experiments, four criteria are applied to evaluate the performance of the proposed method:

Completion time: the amount of time to complete offloading plus the time of task allocation algorithm.

Success rate: indicative of successful offloading.

Consumed Energy: the total energy consumption of all devices involved in offloading. It is the sum of the values calculated for each node through Equ. (1).

Percentage of task failure: the percentage of subtasks that fail after being offloaded to adjacent mobile devices.

The impact of computational load (required processing time) of subtasks in this set of experiments is explored, where two sets of subtasks with different computational loads are of concern:

Case 1: Set of subtasks with low computational load.

Case 2: Set of subtasks with high computational load.

To evaluate the proposed algorithm, the two cases are examined in two scenarios, where two failure models are of concern [17]:

Fail-fast: a node fails at the first time-slot and cannot complete any task.

Fail-slow: a node may fail at any time; thus being able to complete some of its assigned tasks before the failure

Table 2. Features of mobile devices in testbed.

ID	***Mobile Device***	***CPU***	***Memory***	***OS***	***Battery capacity (Joule)***
A	Samsung Galaxy core 18260	Dual-Core 1.2GHZ Cortex-A5	1GB	Android OS, V4.1.2	24624 Joule
B	Samsung Galaxy Grand2	Quand-Core 1.2GHZ cortex-A7	1.5GB	Android OS, V4.4.2	35568 Joule
C	Samsung Galaxy Note 800	Quand-Core 1.4GHZ cortex-A7	2GB	Android OS, V4.1.2	100000 Joule
D	LG L Fino	Quand-Core 1.2GHZ cortex-A7	1GB	Android OS, V4.4.2	25992 Joule
E	Huawei Ascend G730	Quand-Core 1.3GHZ cortex-A7	1GB	Android OS, V4.4.2	31464 Joule

6.1- First Scenario

Here, the proposed method is evaluated and compared by considering fail-fast in both cases.

First Sample: The foregoing algorithms are reviewed and compared by considering fail-fast in case 1. In this situation, replication is applied in the proposed algorithm because the subtasks have low computational load.

Fig. (3.a) depicts the corresponding success rates. As evident, in all of the algorithms, the rate begins to suffer as the number of subtasks grows. This is because greater network traffic leads to higher failure rates. Likewise, an increase in the number of subtasks reduces the colocation probability between service provider and client node, thus increasing failure due to mobility. The random allocation algorithm has the lowest success rate while the proposed algorithm has the highest success rate with an average of 95.5%, which is 70%, 27%, 21%, and 14% higher than random, reliable offloading, checkpointing, and dynamic grouping, respectively. The reason is that the proposed algorithm selects reliable mobile devices with the highest rank as resources followed by replicating subtasks to several devices. However, the reliable allocation and checkpoint algorithms merely select reliable mobile devices without replicating subtasks in several mobile devices.

Completion times of the algorithms for the first sample can be seen in Fig. (3.b). The completion time in random allocation algorithm is higher than that of the other algorithms because it has a high failure rate forcing the failed subtasks to be reassigned. Contrarily, the proposed algorithm exhibits the lowest completion time, which is 19%, 6.5%, and 0.6% lower than random, checkpointing, and dynamic grouping, respectively. This is attributed to two reasons: (1) maximum success rate minimize the need to re-assign failed subtasks and (2) the response from the fastest replica is regarded as the ultimate result. The client may receive their results earlier than when the subtasks are not replicated. If the size of the task grows, checkpointing experiences a boost in performance because the overhead tends to dwindle. Applying checkpointing for big task reduces completion time. While for small task, overhead for getting checkpoint is large compared to the task size.

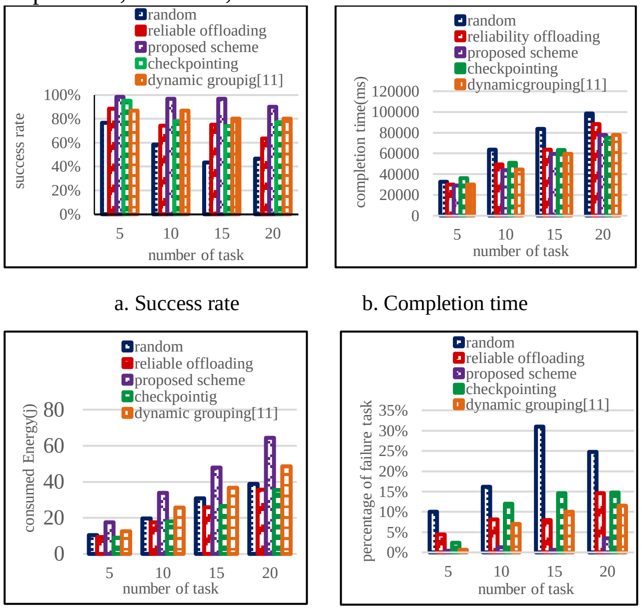
Total energy consumptions in the first sample are shown in Fig. (3.c). The proposed algorithm has the highest energy consumption, since it replicates tasks in several mobile devices. This increasing energy consumption is not significant because subtasks are small. Contrarily, reliable allocation and checkpointing have the lowest total energy consumption as they select mobile devices with low energy consumption rates and do not perform replication. Compared with the proposed algorithm, dynamic grouping uses less energy (about 32%) since it takes advantage of checkpointing for reliable groups.

The percentages of task failures in first sample are illustrated in Fig. (3.d), where the percentage of failure task in this proposed algorithm is lower than others. on average, percentage of failure task in this proposed algorithm is 11% lower than reliable offloading and 6% lower dynamic grouping, because tasks are replicated in several mobile devices.

Second Sample: Here, the aforementioned algorithms are reviewed and compared by taking fail-fast in case2. In this situation, the proposed algorithm applies checkpointing for fault tolerance because subtasks have high computational loads.

The success rate in the second sample is shown in Fig. (4.a), where replication achieves first ranks (12% higher than the proposed method) given that it replicates tasks in several mobile devices. However, it is not efficient in this sample because subtasks have high computational loads.

The completion time in second sample is expressed in Fig. (4.b), where the proposed algorithm and has the lowest completion time because failed subtasks resume execution on new service provider devices from the latest recorded checkpoint. As it is illustrated, on average, the completion time in the proposed algorithm by 11%, 16%, 13%, and 27% is lower than reliable offloading, replication, and dynamic grouping, respectively. Thus, proposed algorithm saves completion time. This improvement is not very impressive, however, because of fail-fast.



a. Success rate

b. Completion time

c. Consumed Energy

d. Percentage of task failure

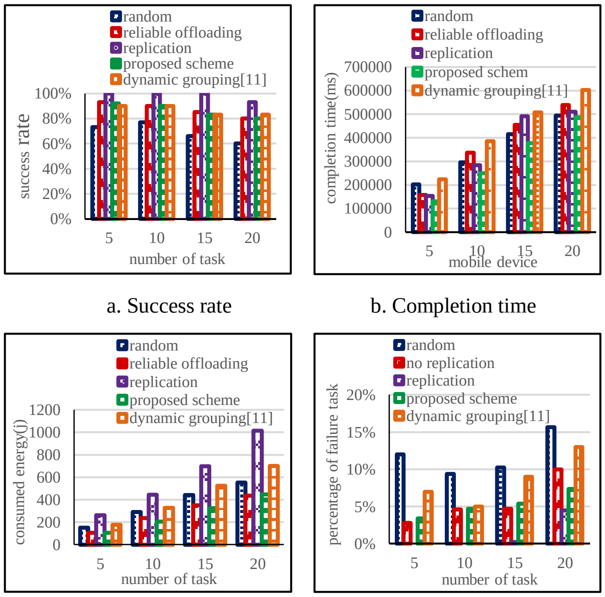
Fig 3. Impact of different numbers of subtasks on the proposed algorithm and previous methods according to fail-fast and subtasks of case1

The success rate in the second sample is shown in Fig. (4.a), where replication achieves first ranks (12% higher than the proposed method) given that it replicates tasks in several mobile devices. However, it is not efficient in this sample because subtasks have high computational loads.

The completion time in second sample is expressed in Fig. (4.b), where the proposed algorithm and has the lowest completion time because failed subtasks resume execution on new service provider devices from the latest recorded checkpoint. As it is illustrated, on average, the completion time in the proposed algorithm by 11%, 16%, 13%, and 27% is lower than reliable offloading, replication, and dynamic grouping, respectively. Thus, proposed algorithm saves completion time. This improvement is not very impressive, however, because of fail-fast.

Fig. (4.c) displays the overall energy consumption of all devices for each algorithm with the second sample. Replication has the highest consumption of energy since tasks are replicated on several mobile devices. This increased energy consumption is impressive because subtasks are large. The proposed algorithm, on the other hand, exhibits the lowest energy consumption, which is 25%, 4.5%, 50%, 37% greater than random, reliable offloading, replication, and dynamic grouping, respectively. The reason is that the new service provider does not need to execute subtask from the beginning, which leads to a reduction of energy requirements.

Finally, Fig. (4.d) shows task failure percentages with the second sample. Failure percentage of the proposed algorithm is lower compared to dynamic grouping and random allocation. This can be attributed to the selection of reliable mobile devices.



a. Success rate

b. Completion time

c. Consumed Energ

d. Percentage of task failure

Fig 4. Impact of different numbers of subtasks on the proposed algorithm and previous methods according to fail-fast and subtasks of case 2

6.2- Second Scenario

In this scenario, the proposed algorithm is evaluated and compared by considering fail-slow in both cases.

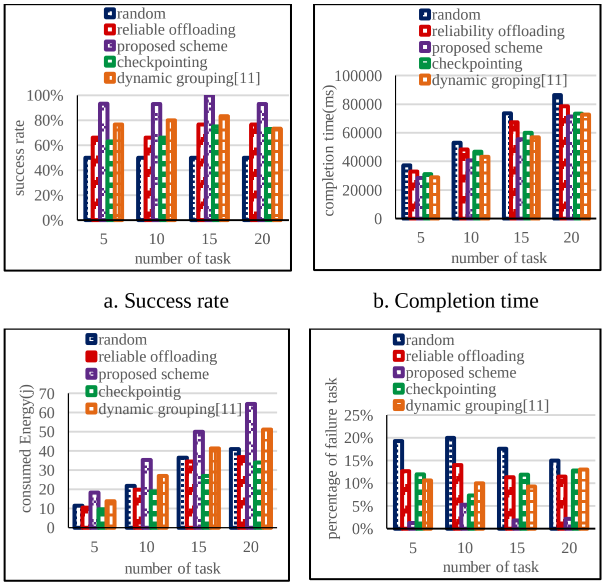
First Sample: The mentioned algorithms are reviewed and compared by considering fail-slow in case1. In this situation, after the proposed algorithm selects reliable mobile devices it applies replication for fault tolerance because subtasks have low computational load.

The success rate with the first sample is shown in Fig. (5.a). The success rate of the proposed algorithm is higher than that of the other algorithms because reliable mobile devices with the highest rank are selected as a resources and subtasks are replicated to several devices. Hence, as it is illustrated, the proposed algorithm with average of 89%, 32%, 36%, and 21% surpasses random, reliable offloading, checkpointing, and dynamic grouping, respectively.

Completion times with the first sample are shown in Fig. (5.b). As evident, the proposed algorithm has the lowest completion. It is expected that checkpointing have the lowest completion time because new service providers do not need to execute failed subtask from the beginning. However, this is not true in this case as, for small tasks, the overhead of getting checkpoints is larger than that of restarting.

The total energy consumption of all devices for each algorithm with the first sample is shown in Fig. (5.c). Once again, the proposed algorithm has the highest consumption since tasks are replicated to several devices, while dynamic grouping consumes lower energy, which exceeds the proposed method by 25%.

By comparing the algorithms using the first sample of the two scenarios, it can be deduced that success rate of the proposed algorithm is higher than other algorithms. Moreover, the proposed algorithm has the highest energy consumption. This increasing energy consumption is not significant because subtasks are small.



a. Success rate

b. Completion time

c. Consumed Energy

d. Percentage of task failure

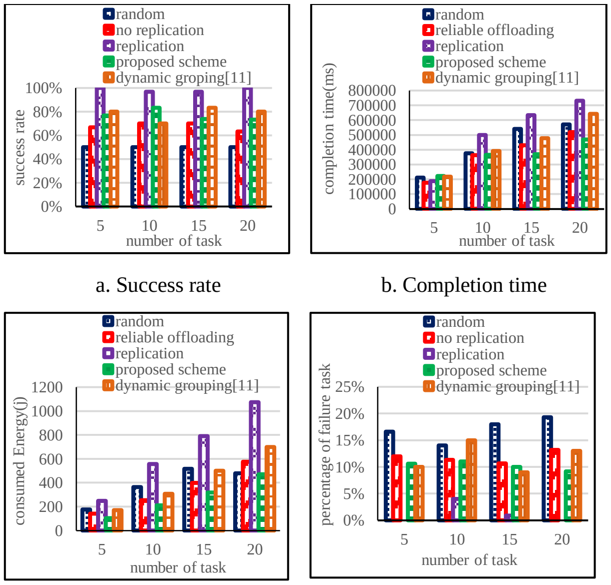
Fig 5. Impact of different numbers of subtasks on the proposed algorithm and previous methods according to fail-slow and subtasks of case1

Second Sample: Here, the mentioned algorithms are reviewed and compared by considering fail-slow in case2. Under these circumstances, after the proposed algorithm selects reliable mobile devices as resources to assign tasks, it applies checkpointing for fault tolerance because subtasks have high computational loads.

Success rates with the second sample are shown in Fig. (6.a). The success rate of replication is higher than that of the other algorithms (28% higher than the proposed method).

The completion time in the second sample is shown in Fig. (6.b). The completion time of the proposed algorithm is lower than that of the other algorithms (16%, 4%, 30%, and 17% lower than random, reliable offloading, replication, and dynamic grouping, respectively). The improvement in completion time is significant because of fail-slow and tasks with high computational loads. However, when the number of subtasks decreases, checkpointing overhead increases in relation to task size. Therefore, completion time increases**.**

The total energy consumption of the algorithms can be seen in Fig. (6.c). Replication uses the largest amount of energy (59% more than the proposed algorithm). The increase in energy consumption is significant because subtasks are large. The total energy consumption in the proposed algorithm is lower than that of the other algorithms (28%, 19%, and 34% lower than random, reliable offloading, and dynamic grouping, respectively).



a. Success rate

b. Completion time

c. Consumed Energy

d. Percentage of failure task

Fig 6. Impact of different numbers of subtasks on the proposed algorithm and previous methods according to fail-slow and subtasks of case2

After comparing the algorithms in the second sample of the first and second scenarios, it can be deduced that in the second sample, the total energy consumption and completion time of the proposed algorithm are lower than other algorithms because, in the proposed algorithm, new service providers do not need to execute failed subtask from the beginning. This improvement in second scenario is better than the first because of fail-slow.

These results indicate that replication is very costly for tasks with high computational load because, in replication, energy consumption is high. Although, success rate in checkpointing is lower than that of replication algorithm, checkpointing is more appropriate for tasks with high computational loads. Consequently, in this paper we tried to make a trade of between energy and success rate. As it is illustrated, our algorithm saves impressive amount of energy in tasks with high computational load, while success rate of proposed algorithm is lower than Replication. In contrast, for tasks with low computational load, the success rate of the proposed algorithm is higher than other algorithms, while it has the highest energy consumption. However, this increasing energy consumption is not significant because subtasks are small.

7- Conclusion

In this paper, a new approach was proposed for fault tolerance where a fully distributed resource allocation algorithm was applied without using any central component with the objective to improve the reliability of mobile resources. Mobile devices in this algorithm are adopted as resources to which tasks are assigned. In the context of mobile devices, energy constraints, mobility, and availability are considered as fault factors used to predict device states and prevent faults caused by volatility of mobile devices. The algorithm applied replication or checkpointing for fault tolerance according to the task size. Here, a context-aware reliable offloading middleware was developed to collect contextual information, manage reliable offloading processes, and fault tolerance. To evaluate the proposed method, several experiments were run in a real environment. The results showed higher success rate as well as significant improvements in completion time and energy consumption for tasks with high computational loads.

In future studies, secure offloading by assigning tasks to trusty devices would be of concern to overcome malicious users. Moreover, extending the proposed method for scenarios in which multiple offloading requests are submitted simultaneously is regarded as another future work.

References

[1] N. Fernando, S. W. Loke, and W. Rahayu, “Mobile cloud computing: A survey,” *Future Generation Computer Systems*, vol. 29, no. 1, pp. 84-106, 2013.

[2] M.Othman, S. A. Madani, and S. U. Khan, “A survey of mobile cloud computing application models,” IEEE Communications Surveys & Tutorials, vol. 16, no. 11, pp. 393-413, 2014.

[3] G. F. Huerta Cánepa, “A context-aware application offloading scheme for a mobile peer to peer environment,” Ph.D. dissertation, Department of Information and Communication Engineering, KAIST, South Korea, 2012.

[4] M. Conti, et al. “Looking ahead in pervasive computing: Challenges and opportunities in the era of cyber–physical convergence,” Pervasive and Mobile Computing, vol. 8, no. 1, pp. 2-21, 2012.

[5] V. Cardellini, V. De NitoPersoné, V. Di Valerio, F. Facchinei, V. Grassi, F. Lo Presti and V. Piccialli, “A game-theoretic approach to computation offloading in mobile cloud computing,” Technical Report, 2013..

[6] S. G. Falavarjani, M. Nematbakhsh, and B. S. Ghahfarokhi, "Context-aware multi-objective resource allocation in mobile cloud," Computers & Electrical Engineering,vol. 44, pp. 218-240, 2015.

[7] C.Shi, et al. “Serendipity: enabling remote computing among intermittently connected mobile devices,” In Proceedings of the thirteenth ACM international symposium on Mobile Ad Hoc Networking and Computing. ACM, 2012, pp. 145-154.

[8] B. Zhou, A. V. Dastjerdi, R. N. Calheiros, S. N. Srirama, and R. Buyya, “A context sensitive offloading scheme for mobile cloud computing service,” In Proceedings of IEEE of 8th International Conference on In Cloud Computing (CLOUD), 2015, pp.869-876.

[9] D. N. Raju, and V. Saritha. "Architecture for fault tolerance in mobile cloud computing using disease resistance approach." International Journal of Communication Networks and Information Security, vol. 8, no. 2, 2016.

[10] B. Zhou and R. Buyya, "A Group based Fault Tolerant Mechanism for Heterogeneous Mobile Clouds," In Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, Nov 2017.

[11] J. Park, H. Yu, H. Kim, and E. Lee, “Dynamic group- based fault tolerance technique for reliable resource management in mobile cloud computing,” Concurrency and Computation: Practice and Experience, John Wiley & Sons, vol. 26, no. 17, Jan. 2014.

[12] J. Park, H. Yu, E. Lee, “Resource allocation techniques based on availability and movement reliability for mobile cloud computing,” Distributed Computing and Internet Technology, Springer Berlin Heidelberg, 2012,pp. 263–264.

[13] J. S. Park and E. Y. Lee, “Entropy-based grouping techniques for resource management in mobile cloud computing,” Ubiquitous Information Technologies and Applications, Springer Netherlands, 2013, pp. 773-780.

[14] P. Stahl, et al, "Dynamic Fault-Tolerance and Mobility Provisioning for Services on Mobile Cloud Platforms." In Proceedings of the 2017 5th International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), IEEE, 2017, pp. 131-138.

[15] S. Choi, K. Chung, and H. Yu, “Fault tolerance and QoS scheduling using CAN in mobile social cloud computing”, Cluster Computing, vol.17, no.3, pp. 911-926, 2014.

[16] E. E. Marinelli, Hyrax: cloud computing on mobile devices using MapReduce. No. CMU-CS-09-164. Carnegie-mellon univ Pittsburgh PA school of computer science, 2009.

[17] C-A. Chen, et al., “Energy-efficient fault-tolerant data storage and processing in mobile cloud,” IEEE Transactions on cloud computing, vol. 3, no. 1, pp. 28-41, 2015.

[18] J. Park, H. Yu, K. Chung, and E. Lee, “Markov Chain based Monitoring Service for Fault Tolerance in Mobile Cloud Computing,” In Proceedings of IEEE Workshops of International Conference on Advanced Information Networking and Applications, 2011, pp.520-525.

[19] P. Patel and V. Prakash, “FTAB: Fault Tolerance Approach by Using HMM with BAUM-WELCH Algorithm in MCC,” In Proceedings of Tenth international conference on Wireless and Optical Communication Network (WOCN), 2013, pp.1-4.

[20] L. Ling, W. Zhulin and Y. Xiuhua, “Mobile Resource Reliability-Based Task Allocation for Mobile Cloud”, In Proceedings of the 2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC), IEEE, 2015, pp.1746-1750.

[21] J.Park, et al, “Two- phase grouping- based resource management for big data processing in mobile cloud computing,” International Journal of Communication Systems, vol. 27, no. 6, pp. 839-851, 2014.

[22] L. McNamara, C. Mascolo, L. Capra, “Media sharing based on colocation prediction in urban transport,” In Proceedings of the 14th ACM international conference on mobile computing and networking, ACM, 2008. pp. 58–69.

[23] P. A, Lee, and T. Anderson, “Dependable computing and fault tolerant systems,” Fault Tolerance: Principles and Practice, Springer Verlag, NewYork, vol. 3, 1990.

[24] Ch. Song, et al. “Limits of predictability in human mobility,” Science 327.5968 , 2010, pp.1018-1021.

[25] J. Nicholson and B. D. Noble. “Breadcrumbs: forecasting mobile connectivity,” In Proceedings of the 14th ACM international

conference on Mobile computing and networking. ACM, 2008, pp. 46-57.

[26] Huang, Wei, et al. “Predicting human mobility with activity changes,” International Journal of Geographical Information Science, vol. 29, no. 9, pp. 1569-1587, 2015.

[27] S. Ch. Shah, “Energy efficient and robust allocation of interdependent tasks on mobile ad hoc computational grid,” Concurrency and Computation: Practice and Experience, vol. 27, no. 5, pp. 1226-1254, 2015.

[28] M.Sepahkar and M. R. Khayyambashi, “A novel collaborative approach for location prediction in mobile networks,” Wireless Networks, pp. 1-12, DOI 10.1007/s11276-016-1304-1, 2016.

[29] J. A. Gubner, Probability and random processes for electrical and computer engineers, Cambridge University Press, 2006.

[30] M. Wiesmann, et al. “Understanding replication in databases and distributed systems”, In Proceedings of the 20th International Conference on IEEE, Distributed Computing Systems, 2000, pp. 464-474.

[31] R. Tuli and P. Kumar, “Analysis of recent checkpointing techniques for mobile computing systems,” International Journal of Computer Science & Engineering Survey, vol. 2, no. 3, 2011.

[32] E. Cuervo, et al. “MAUI: making smartphones last longer with code offload,” In: Proceedings of the 8th international conference on mobile systems, applications, and services, MobiSys’10, 2010, pp. 49–62.

[33] M.D. Kristensen, “Scavenger: transparent development of efficient cyber foraging applications,” In: Proceedings of the IEEE international conference on pervasive computing and communications (PerCom); 2010. p. 217–26.

[34] C. L. Hwang, and K. Yoon. Multiple attribute decision making: methods and applications a state-of-the-art survey, Springer Science & Business Media, Vol. 186, 2012.

[35] C. E. Shannon, “A mathematical theory of communication,” ACM SIGMOBILE Mobile Computing and Communications Review, vol. 5, no. 1, pp. 3-55, 2001.

[36] FaceDetector, http://developer.android.com/reference/android/media/FaceDetector.html Avalaible Online @ September 2016.

[37] L. Zhang, et al. “Accurate online power estimation and automatic battery behavior based power model generation for smartphones,” In Proceedings of the eighth IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis. ACM, 2010, pp. 105-

Zahra Najafabadi Samani received her B.Sc. degree in Computer Hardware Engineering from Azad University, Najafabad Branch, Iran in 2010, and M.Sc. degree in Computer Architecture from University of Isfahan, Iran, in 2016. Her research interests include Distributed Systems and High Performance Computing, Cloud and Mobile Cloud Computing, Optimization and multi-criteria decision analysis method, Future Internet Network Architectures.

Mohammad Reza Khayyambashi received his B.Sc. degree in Computer Hardware Engineering from Tehran University, Tehran, Iran in 1987. He received his M.Sc. in Computer Architecture from Sharif University of Technology (SUT), Tehran, Iran in 1990. He got his Ph.D. in Computer Engineering, Distributed Systems from University of Newcastle upon Tyne, Newcastle upon Tyne, England in 2006, he was research assistant during his Ph.D. course at University of Newcastle upon Tyne, Newcastle upon Tyne, England. He is now working as an Associate Professor at the Department of Computer Architecture, Faculty of Computing Engineering, University of Isfahan, Isfahan, Iran. His research interests include Distributed Systems, Computer Networking, Mobile Cloud, Cloud Computing, Software Define Network (SDN), Mobile and Social Networks.

Standard Deviation Characterization of a Small Size Reverberation Chamber by Using Full-wave Simulation and E-Field Probe

Ehsan. Poodineh
Department of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran,
e.poodineh@email.kntu.ac.ir
Farhad. Ghorbani
Department of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran,
farhadghorbani@email.kntu.ac.ir
Reza. Asadi
Department of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran,
R_asadi@email.kntu.ac.ir
Hadi. Aliakbarian*
Department of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran,
aliakbarian@ieee.org

Received: 30/Oct/2019 Revised: 30/Nov/2019 Accepted: 26/Dec/2019

Abstract

Reverberation Chamber (RC) is a new type of measurement equipment used in electromagnetic compatibility and antenna tests and capable to produce an almost uniform electric filed inside a Working Volume (WV). In this paper, the field uniformity of an actual small size RC is studied. At first, the mode density of the chamber which should be larger than unity is investigated. In the next step, the Standard Deviation (SD) of a small size RC, as a field uniformity criterion is investigated in an existing RC. A highly detailed three-dimensional model of chamber including its stirrers, antenna, WV, and its door create to verify the field uniformity of a RTS60 reverberation chamber. The removal of reverberation chamber's stirrers shows that they have a direct effect on the uniformity of field. As the stirrer moves during the test, the effect of three different position of stirrer on the field uniformity is investigated. The transmission antenna, as an important component of these rooms, is simulated and investigated separately. The reflection coefficient of that antenna fit the working frequency band of the chamber. In a real scenario, the SD of the chamber is measured by using an electric field probe. A comparison between the simulation and measurement also is done in order to confirm the uniformity of the electric fields.

Keywords: Reverberation chamber; stirrer; Electromagnetic Compatibility tests; Antenna gain and efficiency; standard Deviation.

1- Introduction

As technology spreads in every aspect of human life, we need electrical devices to work side-by-side without interfering with each other operation. Electromagnetic Compatibility (EMC) standards and consequently tests have been defined to ensure the proper performance of electronic equipments in the vicinity of each other and in different electromagnetic conditions [1]. To test and study the effects of electromagnetic wave's interference on electrical device, some standard tools, which provide uniform field distribution within specific environments, has been designed and implemented. The importance of field uniformity is to simulate the actual interference condition.

Today, various test rooms, including acoustic anechoic chambers [2] , RF shielded Room [3] , radio-frequency anechoic chambers [4] , semi-anechoic chambers [5] and GTEM CELL [6, 7] , are designed and manufactured to perform various tests, including antenna testing, radar cross-section testing, and EMC testing. Reverberation chambers, which is the most recent invention in this regard has been more successful in implementing the real situation compared to other rooms [8] . In recent years, the use of RC in the measurement of electromagnetic compatibility has become more prominent. Measurement of radiated power and antenna efficiency are two commonly capability of this type of rooms. Furthermore, enclosure shielding enables multipath environments simulation, and biological effects studies [2].

These rooms have a relatively large WV that capable measuring the characteristic of large device. Beside of that

* Corresponding Author

the RC is economically more cost effective than other test rooms. The field uniformity in the WV is the main feature of this chamber. Furthermore this room have been accepted by many international standards, including the US military industry standard ML-STD-246E, which was issued in 1999 [9] , and the international standard IEC61000-4-21, which was issued in 2003 [10].

Many studies have done to investigate the characteristics of RCs. Reference [11] provides a general method for determination of the number of independent stirrer positions. The Total Scattering Cross Section (TSCS), which is a proper parameter to describe the performance of the stirrer in the reflection chamber, is presented in Reference [12]. In terms of RC simulation, the structures of RC are electrically very large (in the order of *100×100* wavelengths). So the simulation is made very time consuming and difficult specially because mode-stirrer are moving during the operation of RCs. Reference [8] explores the full-wave Finite-Difference Time-Domain (FDTD) method for simulating large Reverberation chambers with realistic stirrers and antennas. In Reference [13] the performance of the mechanical stirrer in the reflection chamber is numerically investigated.

As we were to use our RC, it was also important to us to investigate the important characteristics of the room. Moreover, this help us to repeat it when designing a new RC or making any change in the current one. In our paper for the first time, the SD characteristics of an RC has been compared and studied practically in measurements and in full wave simulation. The first important parameter of a RC is its field uniformity and SD over the working frequency band. Having the fact that no full wave characterization of SD has been reported before. For this purpose, our RTS60 Bluetest reverberation chamber in Wireless Terminal Measurement laboratory (WiTeM) of K.N.Toosi University of Technology, which is to the authors' knowledge the only small size RC in Iran, has been selected. The RC is modeled and simulated in CST Microwave Studio with maximum details. To validate the simulation results the practical measurement performed based on requirement of standards. The field obtained by both methods, is used to analyze the field uniformity in term of SD parameter. In addition, the effect of stirrer motions and WV rotation on field uniformity are investigated. The measured and simulated results show that the field all over the Working Volume (WV) is distributed uniformly for different position of stirrers.

This paper is organized as follows. In section II, different parts of a RC is briefly introduced. Section III introduces the SD criterion and its requirement. The RTS60 dimensions and components are shown in section IV. The CST modeling is described in section V. The practical test procedure and the results discussed in section IV.

2- Reverberation Chamber Structure

Reverberation chambers have many advantages including a large working volume, better field uniformity, lower construction costs, etc. The structure of RC generally made of high quality factor shielded rooms with one or more stirrers to change the distribution of the electromagnetic field. Two important features that distinguish the RC from the other rooms are isotropic field and random polarization. The isotropic field causes the energy is distributed uniformly in all direction and the random polarization provides the randomly polarization of the radiated waves. These features make the RC test environment more similar to realistic environment. The statistical field uniformity criteria and operating frequency are two key specifications for the performance of the reverberation chambers.

The RC generally consists of the following sections: shielding cavity, stirrers, antennas and working volume. All the component of the RC must be isolated from the unwanted external field, Thus they are put inside a shielded cavity. The material of shielding cavity walls should be from a very good conductive material such as galvanized steel sheets. This feature provides a high quality factor (Q). The shape of reverberation chambers is generally rectangular, and the size of the chambers depends on the lowest frequency. As the chamber size became larger the operation frequencies, became smaller [14]. . The positions of different components inside the shielding cavity of RTS60 RC are shown in Fig.1.

Stirrers are one of the key components in a reverberation chamber which enable use creating a uniform field distribution by manipulating electromagnetic fields. The inner walls and the stirrers often reflect the electromagnetic waves transmitted by the antenna. The movement of the stirrer inside the room maximizes the field strength in the WV. The main purpose of these stirrers is to create a variable electromagnetic field with amplitude that is statistically uniform. According to the IEC standard, the stirrer radius should not be less than $\lambda/4$ (λ wavelength minimum operating frequency). Generally the stirrer radius is chosen between $\lambda/2-\lambda/3$ and also the largest stirrer dimension should not be less than $3/4$ smallest dimension of the chamber [10]. One, two, or three Stirrer blades are usually used in RCs.

In a rectangular cavity, the number of resonance modes is calculated by equation (1)[15]

$$N=\frac{8\pi}{3}(a\times b\times d)\frac{f^3}{c^3}-(a+b+d)\frac{f}{c}+0.5 \tag{1}$$

where N is the number of modes, a,b,d are the dimension of cavity, f is the frequency and c is the speed of light. The mode density is defined as $D_{S(f)}=dN/df$ for each cavity resonator. For rectangular cavity, this parameter can be calculated from equation (2) based on equation (1)[15].

$$D_{S(f)} = 8 \times \pi(a \times b \times d)\frac{f^2}{c^3} - \frac{a+b+d}{c} \qquad (2)$$

In order to have field uniformity in a RC this parameter must be higher than unity. The modes stimulated inside this chamber could be estimated by the first-order modes stimulated in an ideal rectangular cavity. These modes are combined by the mechanical stirrers inside the chamber to change the boundary conditions. Each mode has a different resonant frequency compared with the ideal initial chamber modes. When the mechanical stirrers start moving, the resonant frequencies of these modes varied. Combination of these modes results in a complex Gaussian field distribution throughout the chamber[15].

Regarding the working frequency, any type of antenna which works in this band can be selected as transmitter antenna. The number and the locations of the antennas inside the RC are determined by the type and the pattern of antenna.

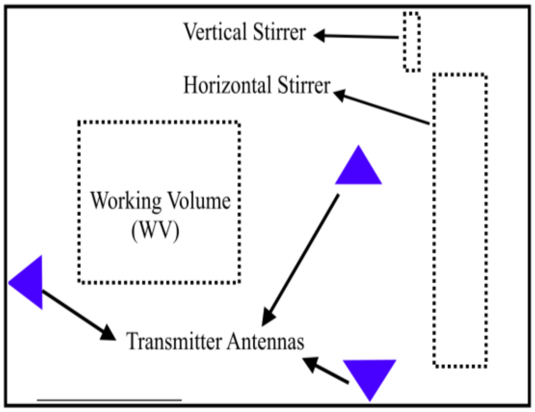


Fig 1 Main components of a RC

The field inside the reverberation chamber is statistically uniform only in the working volume of the chamber (WV). According to the IEC 61000-4-21standard, the working volume is a rectangular area that should be positioned at the opposite side of the room where stirrers are. Also the minimum distance of the WV from the walls, stirrers, antennas or anything else should be $\lambda/4$ [10].

In this work, a RTS60 RC, made by Bluetest AB is used as the test platform. The structure of the RTS60 is shown in Fig. 2. This chamber has dimensions of 1.75m×1.88m ×1.75m. This chamber has 100 dB shielding that is capable for measuring receiver sensitivity without disturbing external wireless devices such as cell phones. By using these dimensions, the mode density of RTS60 can be calculated by equation (2) .In Fig. 3, the mode density of RTS60 in frequency band of 600-6000 MHz is shown. As illustrated in this figure, the mode density is higher than 1 within the working frequency of the chamber.

According to the IEC standard, the lowest operating frequency of a room can be obtained by multiplying the lowest resonant frequency by a factor of 5 or 6 [16]. For our reverberation chamber with the mentioned dimensions, the lowest resonant frequency is approximated to be 100 MHz. Therefore, the lowest operating frequency of RTS60 is approximately 600 MHz. The upper working frequency of this structure is determined 6 GHz by company.

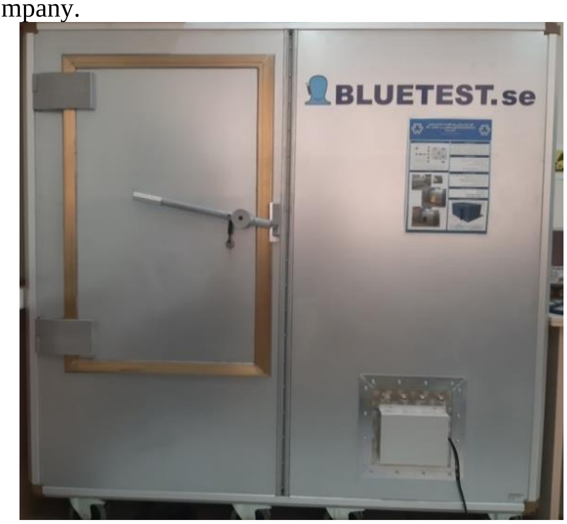


Fig. 2 RTS60 chamber in WiTeM lab, KN Toosi University of Technology

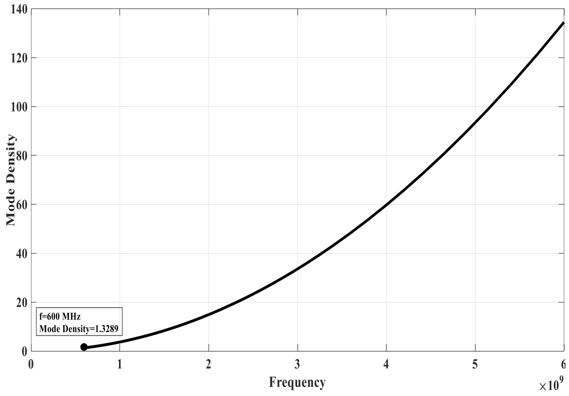


Fig. 3 Mode density of RTS60 for different frequency

3- Field Uniformity Parameter

The field uniformity parameter or Standard Deviation (SD) is one of the most important parameters that must be determined in order to show that statistically the electric field is uniform in the reverberation chambers. According to IEC61000-4-21, the field inside the reverberation chamber is considered uniform if its SD is lower than the threshold that is determined for each frequency. The standard explains that the acceptable SD is below 4 dB from 80 MHz to 100 MHz frequency range. The standard value decrease linearly to 3 dB from 100 MHz to 400MHz, and in frequency higher than 400 MHz is below 3 dB [10]. For a give chamber, SD or σ can be obtained by equation (3)[16] :

$$\sigma = \sqrt{\frac{\sum_{m=1}^{M}\sum_{n=1}^{N}(|E_{m.n}| - |E_{m\times n}|)^2}{(M\times N)-1}} \qquad (3)$$

where *n* varies among three *(N=3)* coordinate axes, usually three *x,y,z* axes, and *m* denotes the number of electric field probes selected in the WV that in our case *M=8*. The IEC recommends to place at least eight probes in the WV. By using the data obtained from probes, the maximum electric field $(E_{m.n})$ and, the average of the maximum electric field $(E_{m\times n})$ calculated. The value of σ in dB can be obtained as follows [16]:

$$\sigma = 20\,log\left(\frac{\sigma + |E_{m\times n}|}{|E_{m\times n}|}\right) \qquad (dB) \qquad (4)$$

In a reverberation chamber, it is not necessary to obtain measurement parameters in all frequencies. According to the IEC 61000-4-21standard, it is enough to set twenty frequency points logarithmically within the frequency range *fs* to *3fs*, fifteen frequency points within the frequency range *3fs* to *6fs*, and finally ten frequencies points within the frequency band *6fs* to *10fs*.

4- Full Wave Simulation of RC

The excitation antennas are Bow-Tie Blade Monopole antenna that fixed at the three chamber walls. The antennas are installed perpendicularly in order to produce wave in all directions uniformly. As shown in Fig.4, the Bow-Tie Blade Monopole is simulated in CST Microwave Studio using its real dimension. The main parameters of the antenna are summarized in Table 1.

The reflection coefficient of the antenna is presented in Fig. 5 showing that it works within the chambers frequency band. The antenna's S_{11} is below -10 dB from 600 MHz up to 6 GHz. The monopole like radiation pattern of the antenna in 3 GHz is also presented in Fig. 6

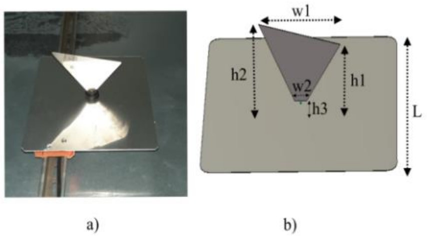


Fig. 4 Bow-Tie Blade Monopole Antenna used in the RTS60
a) fabricated model b) simulated model in CST Microwave Studio

Table 1 Main parameter of the antenna

Parameter	Value
L	*20 cm*
h1	*4.9 cm*
h2	*6.4 cm*
h3	*0.4 cm*
w1	*11 cm*
w2	*1.5 cm*

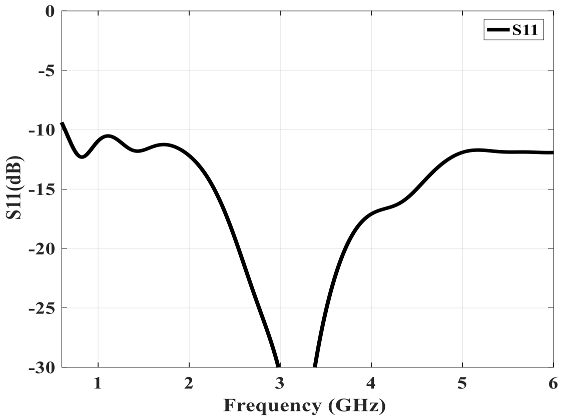


Fig. 5 Simulated reflection coefficient of the Bow-Tie Blade Monopole antenna

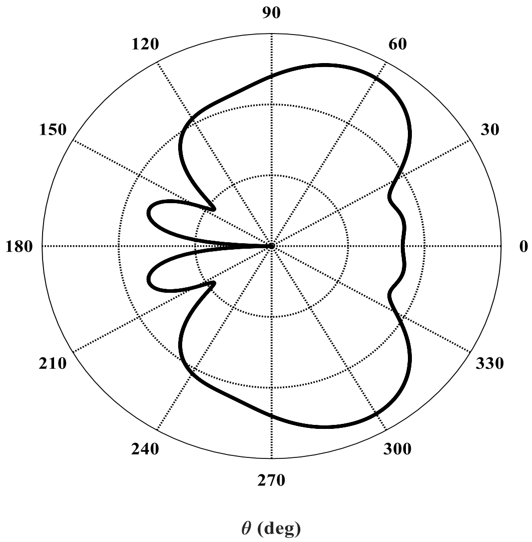


Fig. 6 Simulated radiation pattern of Bow-Tie Blade Monopole Antenna in elevation plane (f= 3 GHz)

The considered RTS60 room has two stirrer blades, one of them moves horizontally and the other one moves vertically. In Fig. 7 the location of these stirrers inside the chambers is shown.

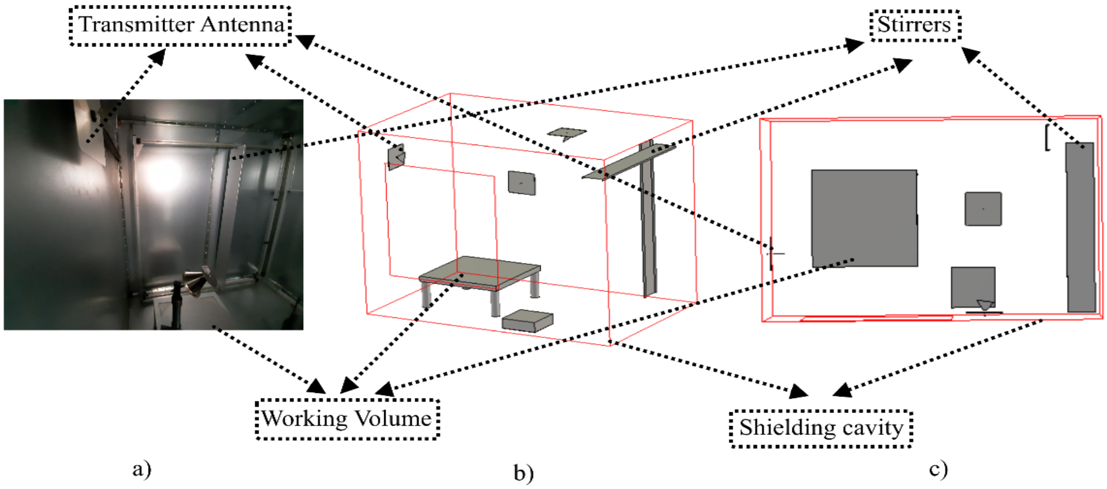


Fig. 7 RTS60 component a) inside component of actual RTS60 b) 3D view of modeling in CST c) upper view of modeling in CST

The location where the table with the dimension of *0.604m × 0.604m* and height of *30 cm* is placed, defines the working volume of the RTS60. This table is able to turns continuously during the tests so that the distribution of electromagnetic waves, affecting all parts of the device under test, becomes statistically uniform.

In order to illustrate the distribution of electric field inside the RTS60, the simulated electrical field intensity inside the chamber is shown in Fig. 8. In this simulation, all of the chamber antennas have been excited simultaneously and equally.

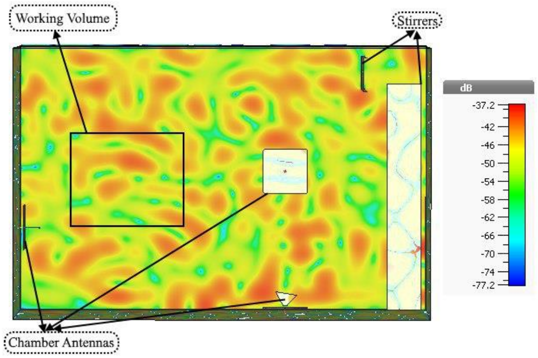


Fig. 8 The electrical field intensity inside the RTS60, simulated in CST (1.5 GHz)

In order to measure the electric field around the table according to IEC 61000-4-21, eight electric field probes at two different heights and at four corners of the table are placed, in accordance with Fig. 9 .The results demonstrate the field uniformity at each point of the table.

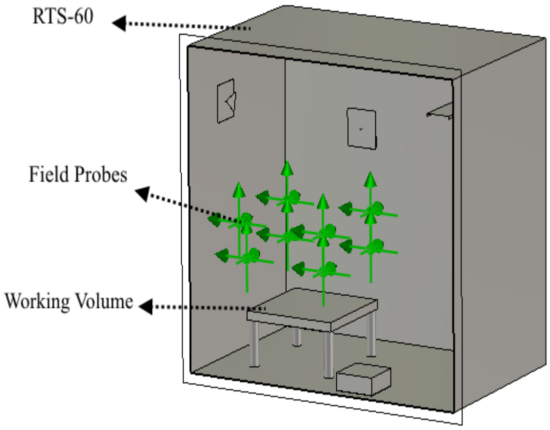


Fig. 9 Location of probes in WV modeled in CST

5- Test Procedure and Results

To measure the SD parameter of this reverberation chamber, the method, introduced in [16], is used. The measurement is performed at the frequency list, which is logarithmically opted based on the method mentioned in section III. For each frequency point, all three antennas are stimulated inside the room. Field probes at each direction individually measure the electric fields produced by each antenna. In the first step, the maximum electric fields of each antenna are obtained separately for each probe location and each particular direction, called Em.n. In the next step, the average of all Em.n for each frequency is calculated and named Em×n in (3) . By placing each of these parameters in (3), the SD parameter is obtained for a given frequency and a specific direction. This process is repeated for each direction and all frequency points. The stirrer is the main component of the Reverberation chamber, which determines the statistical uniformity of the field. The shape, dimensions and the number of stirrers are important parameters that affect the uniformity of the field [17]. To investigate the stirrer effects on the field uniformity, they are initially removed from another simulation and the field uniformity parameter, SD, is obtained again. The results, shown in Fig. 10, demonstrates that by removing the stirrers, the field uniformity inside the RC is deteriorated. As can be seen, the SD goes above 3dB in many frequency points.

Then, the stirrers are added to the simulation. As the field uniformity must be maintained during the movement of stirrers, the SD parameter is checked at different stirrer locations. Ideally, infinitely different locations should be considered for the stirrers to ensure uniformity of field, which is extremely time-consuming. However, due to the electrically large volume of the simulation, we have only obtained the SD parameter in three different locations of stirrers in this work. Full wave Simulation of each position of the stirrers takes four days by using a computer with 64GB of RAMs and core i7 2.4GHz CPU. In the Fig. 11, these three locations of stirrers are shown.

In the Figs. 12-14, the SD parameter of the chamber is given, respectively for three all positions, P1, P2, P3. As shown in these diagrams, the SD value in the frequency range of 600 to 6000 MHz is below the threshold of the IEC 61000-4-21standard, which is almost 3 dB. This verifies that the field uniformity in the working volume of the chamber becomes acceptable with the presence of the stirrers in every position.

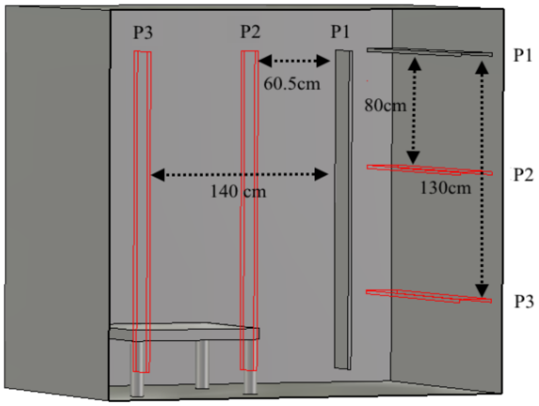


Fig. 10 Three different position of stirrers modeled in CST

In order to ensure the performance of the RTS60 chamber, as well as to verify the simulation results, a set of measurements have performed by putting a field probe inside the chamber. As shown in Fig. 15 the measurements that are performed on the absolute value of SD parameter in this frequency range confirming that the RST60's SD is below the 3dB level set by the standard. The measurement setup consists of a signal generator and a Narda NBM-550 probe is shown in Fig. 16. The probe has a directional sensor that can measure electric field from 3 MHz to 18 GHz with a precision of 1µV. Since the working frequency of the signal generator is limited to 3 GHz, the measurement is performed up to this frequency. The result diagram also makes it clear that the simulated and measured results are agreement. The difference between the practical and the simulation results shown in Fig. 15 Is due to the small details of the low impact those are not included in the chamber modeling.

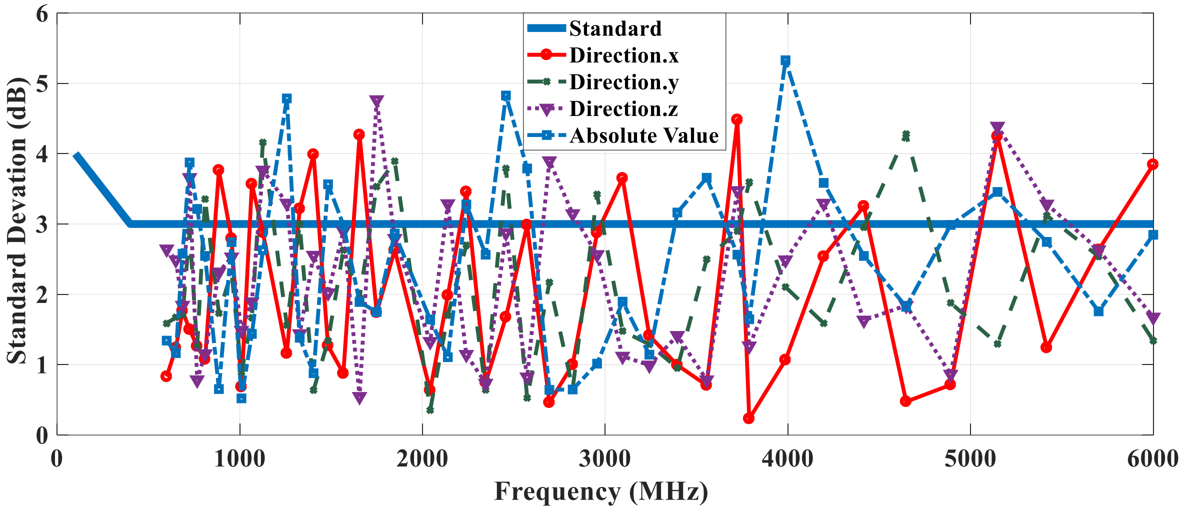


Fig. 11 Four different calculated SD based on simulation results for RC without stirrers

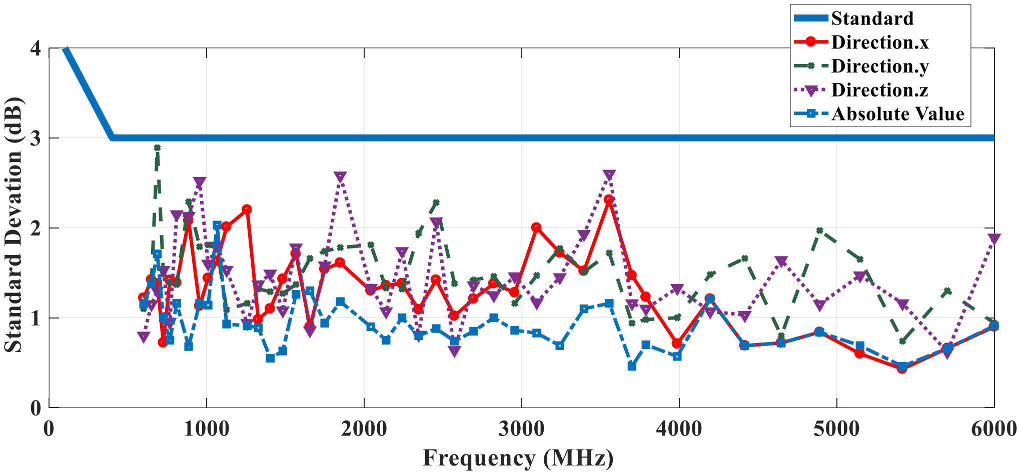


Fig. 12 Four different calculated SDs based on simulation results for the stirrer location *P1*

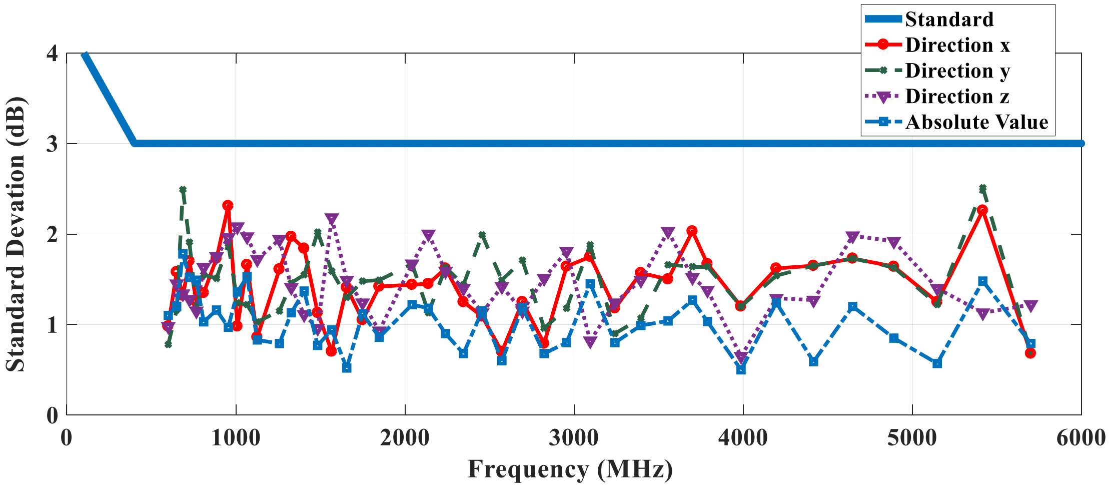


Fig. 13 Four different calculated SDs based on simulation results for the stirrer location *P2*

* Corresponding Author

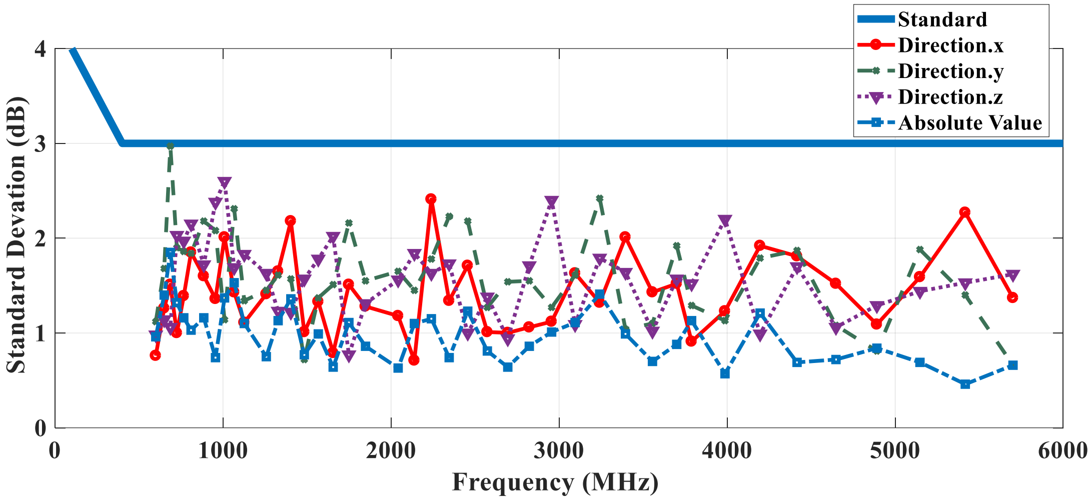


Fig. 14 Four different calculated SDs based on simulation results for the stirrer location *P3*

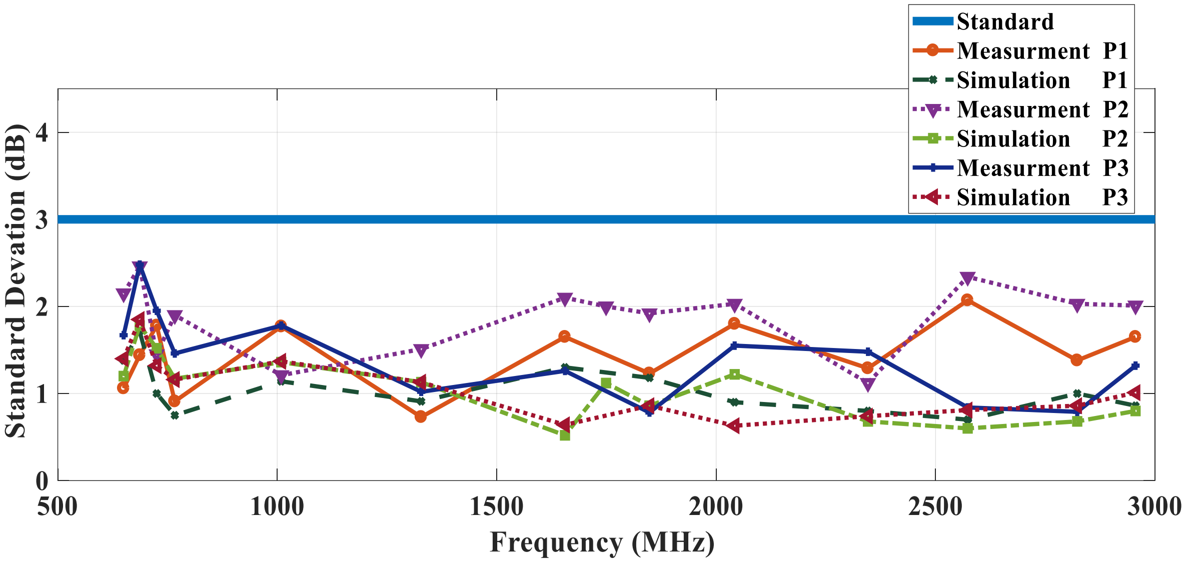


Fig. 15 Calculated absolute value of SD based on measured and simulated results for all three locations, *P1*, *P2* and *P3*

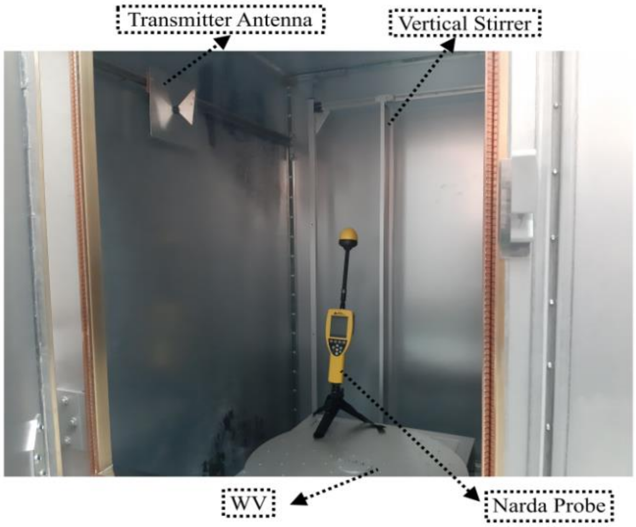


Fig. 16 Electrical field measurement by probe NBM-550, placed inside RTS60

6- Conclusion

In this paper, the field uniformity of a real reverberation chamber (RC), RTS60, has been investigated. As the main feature of real reverberations, the Standard Deviation (SD) of the chamber has been theoretically discussed, simulated and then practically verified. It is also confirmed the stirrers has a crucial role in making the electric field uniform. The simulated and measured results validates that the field distribution within the working volume of the RTS60 is better than the level set by the IEC 61000-4-21 standards.

References

[1].Leach, R. and M.B. Alexander, Electronic systems failures and anomalies attributed to electromagnetic interference. , is presented at the space programs and Technologies conference,alabam1995.

[2].Stephenson, J.H., D. Weitsman, and N.S. Zawodny, Effects of flow recirculation on unmanned aircraft system (UAS) acoustic measurements in closed anechoic chambers. The Journal of the Acoustical Society of America, 2019. 145(3): p. 1153-1155.

[3].Smith, K.M., A homogeneous RF-shielded magnet for low-field magnetic resonance studies," MASTER OF SCIENCE, Department of Physics and Astronomy, in Winnipeg, Manitoba. 2019, The University of Manitob: canada.

[4].J. W. Hansen, E.M.S., J. D. Cleveland, S. M. Asif, B. Brooks, B. D. Braaten, A Systematic Review of In-vitro and in Vivo Radio Frequency Exposure Methods. IEEE reviews in biomedical engineering, 2019.

[5].Vinci, J.J., Sparse Aperture Measurement in a Non-Ideal Semi-Anechoic Chamber., University of Dayton United States,2019.

[6].Dubey, S.K. and V. Ojha, Numerical Analysis and Measurement of Electric field Strength inside GTEM Cell at GSM Frequencies. Defence Science Journal, 2019. 69(5): p. 423.

[7].A.Sahraei, H.A., On the Design and Fabrication of a Large GTEM Cell and its Challenges. to be published in IEEE EMC Magazine, 2020.

[8].A. K. Fall, P.B., C. Lemoine, M. Zhadobov, and R. Sauleau, Design and experimental validation of a mode-stirred reverberation chamber at millimeter waves. EEE Transactions on Electromagnetic Compatibility, 2014. vol. 57: p. pp. 12-21.

[9].States, U., A document that establishes uniform engineering and technical requirements for military-unique osubstantially modified commercial processes, procedures, practices, and methods. 1999. p. pp. 707-710.

[10].Compatibility, E., Part 4-21: Testing and measurement techniques—reverberation chamber test methods. IEC Standard, 2003: p. 61000-4.

[11].Pfennig, S., A general method for determining the independent stirrer positions in reverberation chambers: Adjusting the Correlation Threshold. IEEE Transactions on Electromagnetic Compatibility, 2016. vol. 58: p. pp. 1252-1258.

[12].Q. Xu, Y.H., L. Xing, Z. Tian, C. Song, and M. Stanley, The limit of the total scattering cross section of electrically large stirrers in a reverberation chamber. IEEE Transactions on Electromagnetic Compatibility, 2016. vol. 58: p. pp. 623-626.

[13].L. Bastianelli, V.M.P., and F. Moglie, Stirrer efficiency as a function of its axis orientation. IEEE Transactions on Electromagnetic Compatibility, 2015. vol. 57: p. pp. 1732-1735.

[14].Bruns, C.a.R.V., A closer look at reverberation chambers-3-D simulation and experimental verification. IEEE Transactions on Electromagnetic Compatibility. IEEE Transactions on Electromagnetic Compatibility, 2005. vol. 47: p. p. 612-626.

[15].Hill, D.A., Electromagnetic fields in cavities: deterministic and statistical theories vol, J.W. Sons, Editor. 2009.

[16].B. Urul, I.B.B., T. Göksu, and S. Helhel, CST simulation of reverberation chamber for improved field uniformity, in International Conference on Electrical and Electronics Engineering (ELECO). 2017. p. pp. 1070-1074.

[17].Yamanaka, K.H.a.Y., FDTD analysis on the effect of stirrers in a reverberation chamber,. International Symposium on Electromagnetic Compatibility, 1999: p. pp. 260-263.

Ehsan Poodineh received the B.S. degree in electrical and communication engineering from Shiraz University of Technology (SUTech), Shiraz, Iran in 2013, and M.S. degree in electrical and communication engineering from KN Toosi University of Technology, Tehran, Iran, in 2020. He has been working as a researcher at the WiTeM Laboratory since 2020. His research interests include Electromagnetic Compability (EMC), Reverberation Chambers and Small Antennas

Hadi Aliakbarian is an Assistant Professor in K.N.Toosi University of Technology in Iran since 2013. He received the B.S and M.S degrees in Electrical and Telecommunication Engineering from the University of Tehran in 2002 and 2005, and the Ph.D degree in Electrical Engineering from the Katholieke Universiteit Leuven (KU Leuven) in 2013. He worked for the microwave laboratory and the Center of Excellence on Applied Electromagnetics at the University of Tehran as an Associated Researcher from 2005 to 2007. He currently leads Wireless Terminal Measurements lab (WiTeM) in KN Toosi University of Technology. His research interests include different aspects of antennas, propagation, electromagnetic compatibility and electromagnetics in health and agriculture.

SGF (Semantic Graphs Fusion): A Knowledge-based Representation of Textual Resources for Text Mining Applications

Morteza Jaderyan
Department of Computer Engineering, Bu Ali Sina Uinversity, Hamedan, Iran
m.jaderyan92@basu.ac.ir
Hassan Khotanlou*
Department of Computer Engineering, Bu Ali Sina Uinversity, Hamedan, Iran
khotanlou@basu.ac.ir

Received: 05/May/2019 Revised: 29/Sep/2019 Accepted: 29/Nov/2019

Abstract

The proper representation of textual documents has been the greatest challenge in text mining applications. In this paper, a knowledge-based representation model for text analysis applications is introduced. The proposed functionalities of the system are achieved by integrating structured knowledge in the core components of the system. The semantic, lexical, syntactical and structural features are identified by the pre-processing module. The enrichment module is introduced to identify contextually similar concepts and concept maps for improving the representation. The information content of documents and the enriched contents are then fused (merged) into the graphical structure of a semantic network to form a unified and comprehensive representation of documents. The 20Newsgroup and Reuters-21578 datasets are used for evaluation. The evaluation results suggest that the proposed method exhibits a high level of accuracy, recall and precision. The results also indicate that even when a small portion of the information content is available, the proposed method performs well in standard text mining applications.

Keywords: Semantic document representation; Ontology; Knowledge base (KB); Semantic network; Information fusion.

1- Introduction

The text mining techniques are heavily dependent on the generated representation of text documents; their performance is highly affected by it. In most text mining applications and techniques, a specific model is utilized to represent the information content. Most text mining systems employ simple representation models such as “Bag-of-Words” model to represent the contents. These models combined with an “exact term matching” method are used for retrieving the most relevant information to user preferences. However, such representation models suffer from serious drawbacks and limitations which are documented in [1, 2, 3]. Some of the most serious drawbacks and limitations of these systems are: (1) the inherent ambiguity of natural language, (2) synonymy and (3) polysemy. One solution is the integration of ontology and KBs and using knowledge-based representation models [4, 5]. These methods employ the structured knowledge of ontologies and knowledge bases (KBs) to overcome the ambiguity, to represent the content, to model the semantics and to develop text mining applications.

One of the most important aspects of semantic document representation is to introduce a mechanism for representing the contents, the semantics and also efficiently utilizing them in the intended applications. In this regard, filtering the relevant features and ignoring the irrelevant ones will be the main challenge. Introducing a semantic and knowledge-based document representation model using the graphical structure of semantic networks is the main idea of this paper. The semantic network representation generally consists of a number of interconnected nodes. The connecting links represent the semantic relations between the concepts.

The main contributions of this paper are: integrating the structured knowledge of ontology and KBs in every component of the proposed representation model, using semantic network for representing the contents of documents and the user preferences, introducing a knowledge-based approach for content enrichment and merging the document semantic networks with enriched concept maps to create a comprehensive representation of contents. Therefore, the idea presented in this paper can be summarized as follows. The semantic, lexical, structural and syntactical features of the documents are identified and extracted and the concepts are weighted. The content enrichment module identifies and extracts the concepts and semantic structures. In the next step, the semantic relations are established between concepts using the structured knowledge of ontology and KBs. Then, the identified concepts, semantic structures and the semantic relations between them are represented in a graphical structure of semantic networks. In the end, the concepts and the identified semantic structures are merged with semantic

* Corresponding Author

networks. It can be used in a variety of applications such as information retrieval, indexing, recommender system and information filtering and management.

The rest of the paper is structured as follows: In the second section, the related works are studied. In the third section, the structure of ontology, Wikipedia and WordNet is examined. In the fourth section, the proposed method of knowledge-based document representation is introduced. In the fifth section, the evaluation results are presented. In the sixth section, we will have the discussion and in the seventh section, the conclusion is presented.

2- Related works

In most text mining applications, the semantic and comprehensive representation of documents is the factor that guarantees the optimal performance and effectiveness of the implemented system. The information models determine how the documents should be represented. In this regard, text mining techniques can be classified into three categories: (1) techniques that employ information models, (2) techniques that employ intelligent learning models and (3) techniques that exploit the structured knowledge of ontology and KBs to represent the content. The probabilistic and vector space models (VSMs) are among the most popular and widely used information models for document representation [6]. The language models [7] and the Bayesian network models [8] are considered to be probabilistic models. They use the probability and statistics principles to generate information models. Whereas, the vector space models [9] utilize a vector form for representing the content of documents. In [10], the authors address the problem of text classification by considering Sentence-Vector Space Model (S-VSM) and Unigram representation models. A neural network based representation is then used to capture the semantic information.

Most VSMs-based information models are based on Salton et al.'s researches [9]. In recent years, there has been several studies [1, 4, 11, 12, 13], exploring the idea of employing the structured knowledge of ontology and KBs for constructing semantic information models. These models exploit the structured knowledge of ontologies and KBs to represent the content of documents. In these studies, using the structured knowledge, for creating information models, has shown promising results in this field. The Natural Language Processing (NLP) -based [14], rule-based [15], ontology-based [16] and fuzzy-based [17] are among the knowledge-based content representation models.

In [11, 13], using ontologies and KBs in semantic information indexing and retrieval is studied. In these studies, semantic networks are used for representing the information content. In [12], a personalized method for document search and retrieval is introduced. In this paper, the documents are represented by mapping the concepts to a graph-like structure. The relations between concepts are established using a web-based ontology called ODP [18]. In [19], a method for document indexing in engineering domain is introduced. A domain ontology is employed to represent the content of documents in the form of semantic networks. The Wikipedia is also used for representing content in text mining applications. In [20], the authors introduce a Wikipedia semantic matching approach for text document classification. In order to model the text semantics, documents are represented as concept vectors in Wikipedia semantic space. In [21], the authors introduce a two-level representation model (2RM) for representing text data. At the syntactic level, a document is represented as a term vector (tf-idf) and the Wikipedia concepts, related to the identified terms in syntactic level, are used to represent the document in semantic level, and a multi-layer classification framework (MLCLA) is then used to generate the output.

In [22], a graph-based feature extraction is used to extract meaningful features. The documents are represented as graphs and a weighted graph mining algorithm are applied to extract frequent sub-graphs. The sub-graphs are then further processed to produce feature vectors for classification.

Machine learning techniques can be also used for document representation. The authors in [23] propose the bag-of-concepts method as a document representation method. The proposed method creates concepts through clustering word vectors generated from word2vec. It uses the frequencies of these concept clusters to represent document vectors. Discourse analysis is a collection of Natural Language Processing tasks that are designed to identify linguistic structures and contextual information from textual resources. The extracted linguistic structures are identified at different levels, so that they can be utilized to implement NLP applications such as text analysis, Question Answering and text summarization. One of the most important discourse analysis systems is discourse parser system that is used to represent the structure of a document by a tree-based structure. The key similarities and differences between our approach and the concept of discourse analysis are:

- Both approaches build a tree-based representation of textual resources which are used to capture the semantics and the relations between linguistic elements.
- The resulting representation from a discourse parser is based on a tree-like structure. However, the proposed representation model exploits the graphical structure of semantic networks.
- A discourse analysis system is designed to create a formal representation of linguistic context. On the other hand, the proposed approach exploits the structured knowledge of ontology and KBs to compute and model the conceptual relations between extracted features.
- The discourse analysis uses rhetorical relations such as contrast, explanation and cause to define the semantics.

However, the proposed representation model exploits the ontology-based relations to represent the textual resources.
In recent years, there is a growing interest in integrating model-based, learning-based and Knowledge-based approaches for document representation [24]. In [25], a novel framework for incorporating knowledge bases (KBs) into the neural network is introduced. In this method, a raw text is conceptualized and represented by a set of concepts using a knowledge base. The neural network is then used to transform the conceptualized text into a vector, in which both the semantics and the content information are encoded.
In most text mining applications, the ontology and KBs are used either to compute the similarity of documents or to represent the content of documents. However, in this paper, the structured knowledge of ontology and KBs are integrated with core components of the proposed framework (pre-processing, enrichment and representation). Also, most similar approaches rely solely on the content of a document to identify most similar documents to user preferences. In this paper, the ontology and KBs are used to infer contextually similar concepts and semantic structures.

3- The Structure of Ontology and KBs

One of the most important features of the proposed method is the integration of ontologies and knowledge bases in every component of the method. Therefore, it is necessary to examine their features and information structures.

3-1- OntoWordNet Top-Level Ontology

Each concept in the ontology is organized as synonym set so that the contextually similar concepts can be identified. This facilitates content enrichment [26]. The classes are organized in the form of a sequence. This sequence defines synonym concepts that bear similar meaning in different contexts. OntoWordNet defines three important semantic relations: Superclass, Subclass, Synonymy and Part_of relations.

3-2- WordNet

WordNet[27] models a semantically enhanced lexicon for English language and consists of synsets. The synset organizes a set of synonym concepts. Every synset consists of several senses (different meanings of a concept).

3-3- Wikipedia

Wikipedia data are available for academic use through D.I.S.C.O project [28]. The Wikipedia consists of two sets of data [28, 29]: first-order word vector: which contains words that occur together in Wikipedia and second-order word vector: which contains words that occur in similar contexts.

4-The proposed Method

The proposed method integrates the structured knowledge of KBs into the document representation model. Incorporating the extracted semantics and informational structures into the representation model is the main idea of this paper. Such a representation model brings three important benefits: It exploits extracted information content and the enriched semantics to create a comprehensive representation model, It can be used as a multi-purpose information model in a variety of text mining applications and facilitates the process of matching documents to user preferences. Figure 1, shows the overview of the proposed method. It consists of three modules: the semantic document processing, the content enrichment and the semantic representation module.

4-1- The Semantic Document Processing Module

It performs a number of pre-processing operations to extract four types of features (semantic, morphological, syntactical and structural). This is the first step toward constructing a multi-level representation of documents. The following pre-processing operations are performed: stop-word removal, bi-gram and Uni-gram processing, Part of Speech (POS) tagging [30], lemmatization [31], named-entity recognition [32, 33] and shallow parsing of sentences [34, 35]. Each operation is designed to extract specific type of features. The features are then weighted using the CF-IDF [36]. Let $D = \{d_1, d_2, \dots\dots, d_n\}$ be the set of documents and $d_i = \{t_i^1,\ t_i^2,\ t, \dots\dots, t_i^n\}$ be the document vector d_i, after the weighting method is applied, $w_i = \{w_i^1,\ w_i^2, w_i^3, \dots\dots, w_i^n\}$ is the set of weights assigned to each member of d_i.

4-2- The Content Enrichment Process

The enrichment module enables system to infer useful knowledge and extract informative features from a set of concepts. This component exploits the structured knowledge of ontology and Wikipedia to discover additional informative features that might have been left out. This module can also be used to find concepts that can improve the information content of a given document.

4-2-1- Enrichment by OntoWordNet Ontology:

The notion of "concept map" graph is employed for enriching the content of documents. Each concept is mapped to a class in the OntoWordNet ontology. The concept and the corresponding class are then converted to the concept map. The concept maps are used to annotate the corresponding concepts in a semantic network. At first, for each concept, the corresponding classes of the ontology are extracted. Considering the structure of the ontology, a concept map consists of a concept and a set of corresponding classes. The links between the concept and the ontology classes are the “equivalent” property and the “subclass” relation. The concept maps are represented by a sub-ontology using OWL/XML schema. Such a representation would allow us to merge the generated concept maps with document semantic networks (see section 4.4). An example of a concept map for the concept “news story” is illustrated in Figure 2. The concept maps help the system discover commonalities between document semantic networks and user preferences. Also, the semantic structure of concept maps is vital in constructing a multi-level representation of documents. The superclass and the equivalent concepts are then weighted and appended to the document vector.

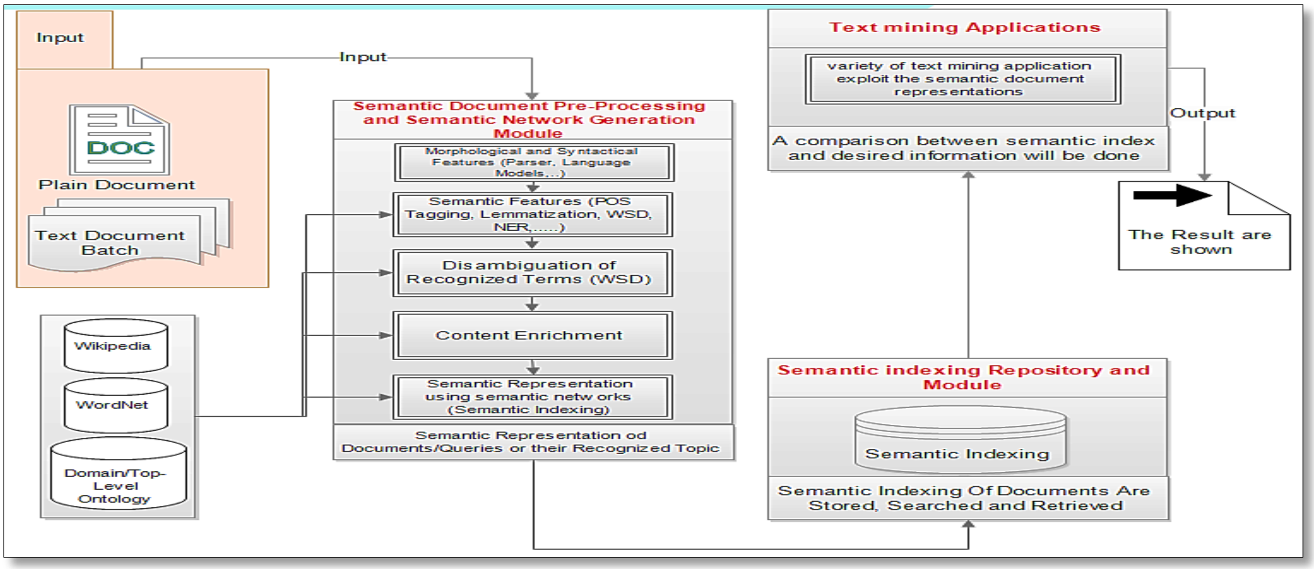

Input
Input
DOC
Plain Document
Text Document Batch
Wikipedia
WordNet
Domain/Top-Level Ontology
Semantic Document Pre-Processing and Semantic Network Generation Module
Morphological and Syntactical Features (Parser, Language Models,..)
Semantic Features (POS Tagging, Lemmatization, WSD, NER,.....)
Disambiguation of Recognized Terms (WSD)
Content Enrichment
Semantic Representation using semantic networks (Semantic Indexing)
Semantic Representation od Documents/Queries or their Recognized Topic
Text mining Applications
variety of text mining application exploit the semantic document representations
A comparison between semantic index and desired information will be done
Output
The Result are shown
Semantic indexing Repository and Module
Semantic Indexing
Semantic Indexing Of Documents Are Stored, Searched and Retrieved

Figure 1- Overview of the proposed knowledge-based representation method

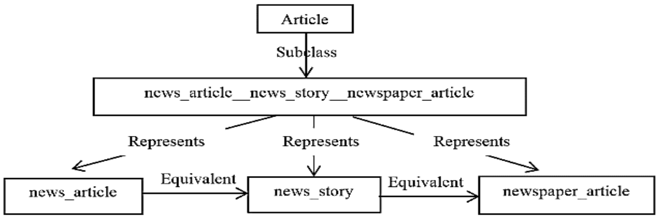


Figure 2- A representation of a generated Conceptual Map

4-2-2- Enrichment Using Wikipedia KB:

In this paper, the Wikipedia second-order word vector is used to enrich the document content. For each concept in the concept vector, co-occurring and contextually similar concepts are retrieved and appended to document vector.

The enriched concepts (contextually similar, co-occurring, superclass and equivalent concepts) are weighted and appended to the document vector. Since the new concepts are inferred indirectly from ontology and KBs, their assigned weight will be lower. The weighting equation is estimated using a subset of evaluation data.

$$Weight_{Related\ Conceot(ontology\ and\ Wiki)} = Weight_{Initial\ Concept} * 0.8 \quad (1)$$

Let $ec_i = \{t_i^{e1},\ t_i^{e2},\ t_i^{e3}, \dots \dots, t_i^{en}\}$ be the set of enriched words/concepts for document d_i, then after appending the enriched words/concepts to the original document vector, $d_i = \{t_i^1,\ t_i^2,\ t_i^3, \dots \dots, t_i^n, t_i^{e1},\ t_i^{e2},\ t_i^{e3}, \dots \dots, t_i^{en}\}$ is the extended document vector.

4-3- The Word Sense Disambiguation of Concepts

Semantic network representation of documents depends on accurate modelling of semantics and relations between features. In this regard, all features need to be cleared of ambiguity. In order to handle the word ambiguity issue in text documents, a method of word sense Discrimination, inspired by [37], is introduced. The underlying assumption of this method is that similar senses occur in similar contexts. In other words, by comparing the collective contextual features of a concept with the information content of each possible sense, we can induce its true contextual meaning. This method relies on the structured knowledge of Wikipedia and WordNet. To this end, the following procedures are performed:

(1) A ±7 context window around the desired concepts in the message is created. Also, the first-order word vector for each member of the context window is retrieved and appended to context window. The window and the appending vectors create a “context vector” for each concept, (2) all possible senses of the concept, their usage example in a sentence and their brief definition for each sense is extracted from WordNet. This will form a “sense vector” for each sense. The first-order word vector for each member of a sense vector is also retrieved and appended to the corresponding sense vector. Finding the similarity between each sense and the context vector determines the contextual similarity between them and (3) a combination of cosine [38] and Jaro-

Winkler [38] measures are used to calculate the similarity score as follows.

$$Sim(Sense_{Vctor}, context_{Vector}) = \frac{1}{2}(Cosine_{Sim}(Sense_{Vctor}, context_{Vector}) + Jaro\_winkler_{Sim}(Sense_{Vctor}, context_{Vector})) \quad (2)$$

The sense vector with highest similarity score is selected as the correct sense vector and the corresponding sense is used to annotate the concept. The output is a set of weighted concepts that are annotated by their true contextual meaning.

4-4- Semantic Document Representation Module

Various models have been proposed to represent the information content of textual resources. Such models include machine learning-based models such as Word embedding model, vector space models, and models based on ontology and structured knowledge bases [11, 4, 16]. One of the most important issues with vector space models and Word embedding models is their inability to model meaning in the information content of documents. On the other hand, the proposed method, which is based on the structured knowledge of the ontology, enables the system to identify semantic relations between words, extract the latent semantics of documents, and ultimately, map the extracted information structures and inferred background knowledge into a semantic network, without losing the semantics in the process. Therefore, the proposed model is a far better choice than vector space and machine learning models. Moreover, the ontology and structured knowledge bases provide useful information such as semantic relations between concepts, vectors representing co-occurring and contextual similarity relations between words/concepts; which makes them the perfect choice for content modelling.

In this paper, semantic networks are used for document representation. The underlying assumption about the representation model is that the information content would be better represented by a percentage of concepts rather than all the concepts. The CF-IDF weighting method is used to determine what percentage of concepts are optimal. Ontology-defined relations are then used to link the concepts in the graphical structure of semantic network.

The enriched concepts and semantic structures play an important role in creating a fully-connected semantic networks. The semantic network generation process consists of two phases: selecting the top n% concepts and generating the semantic networks by linking the concepts.

4-4-1- The Semantic Network Generation Process:

The first step toward creating a semantic network representation of documents is to establish relations between concepts. To this end, the top-n% of concepts are projected onto OntoWordNet ontology. A number of separated concept clusters are then formed. The main reason for this phenomenon is that concepts, which can link the separated cluster, are not identified or they are left out. In summary, the semantic network generation process is carried out as follows: (1) the extracted features and enriched content are weighted using the CF-IDF method and the top-and% of concepts are selected, (2) the proposed semantic network generation algorithm links the concepts together one by one using ontology-defined semantic relations and (3) the liaison concepts connect the separated concept clusters. Figure 3 illustrates how semantic networklinks the concepts and how the liaison concepts link the separated concept clusters. Also, Figure 4 illustrates the proposed semantic network generation algorithm. As shown in Figure 3, after projecting the concepts onto the

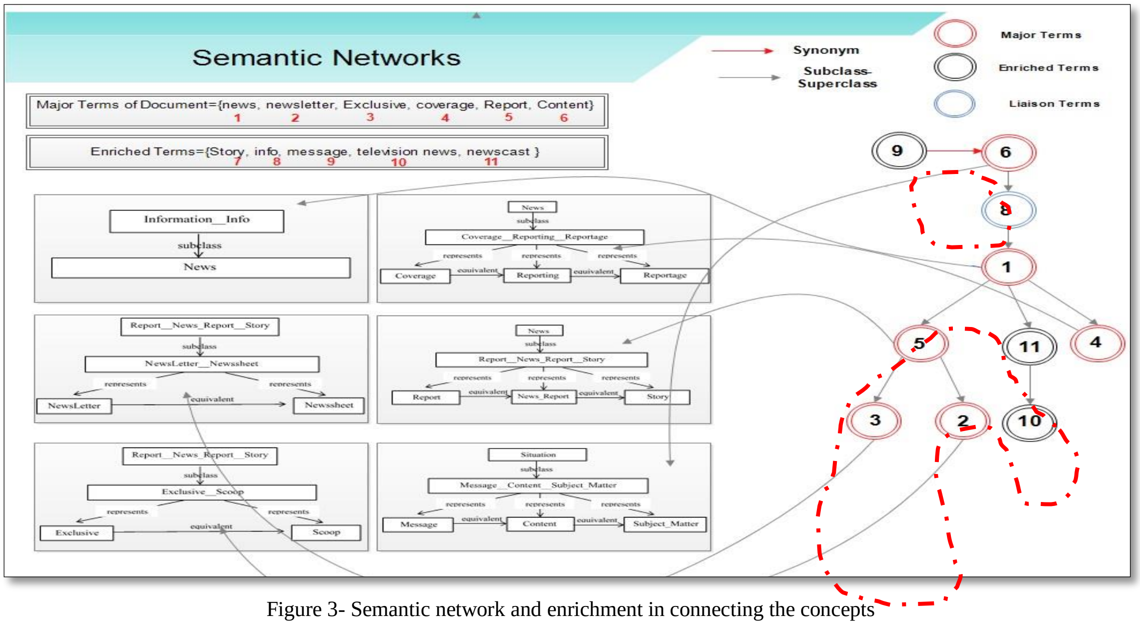


Figure 3- Semantic network and enrichment in connecting the concepts

OntoWordNet ontology, two separated concept clusters are formed. By analyzing the OntoWordNet ontology, it can be understood that the concept “info__information” is the “liaison” concept for connecting the two separated clusters. After enriching, the concept “info__information” is appended to the document semantic network and the connection between the two separated clusters is established. Also, the concepts “television_news” and “newscast” act as the “liaison” concepts for connecting the constructed semantic network with concepts in the deeper hierarchical structure of the ontology.

The semantic networks will be represented as a sub-ontology using OWL/XML schema. Such a representation not only makes the generated semantic networks machine-readable, it also enables the system to merge semantic networks with concept maps. This will create a unified and comprehensive representation of documents.

4-4-2- Merging the Enriched Information with Document Semantic Network:

Incorporating the enriched information and semantic features into the semantic networks will help the system create a fully-connected and comprehensive representation model. Therefore, the assumption is that, merging information from different knowledge sources will result better precision for the system. The following principles are used as a guideline to merge the semantic network with concept maps:

1) The document semantic network is selected as the “stable ontology”. The stable ontology is the more preferred ontology. In case of merging classes, the name of the class in the stable ontology will become the merged class name. Also, if there is a conflict when classes are merged, the stable ontology will be preferred.

2) The class names in both ontologies are scanned to find lexically identical, or linguistically similar the class names. Several factors can be considered to determine the level of similarity between class names, namely synonymy of classes, common sub-strings and common prefix/suffix.

3) In order to merge two classes: If the name of the classes is identical, either the classes will be merged or one of classes will be removed. If the name of the class is linguistically similar, a link between tween two classes will be created. The label of this link will be “similar to”. The class in the “stable ontology” will be linked to other ontology’s class.

4) If a class sub-graph in “stable ontology” is similar to a class sub-graph in the other ontology, they are merged.

5) Automatic updates will be done and the steps *2*, *3* and *4* will be repeated until the ontologies are fully merged.

Figure 5 illustrates the process of merging semantic networks with concept maps.

```
Input: documents D={D_1, D_2,…D_n}, concepts in each document D'={t_1, t_2,…, t_n}
• Loop: for each concept in D
    • Loop: until D' is empty
        • Condition: if semantic network is empty
            • Append the first concept to semantic network.
            • Delete the first Concept from D'
        • End of Condition
        • Min_Node= the minimum of nodes between concepts in hierarchical structure of ontology and KB
        • Loop: for each $t_i$ that already exists in the semantic network
            • Loop: for each $t_j$ in the D'
                • Condition: if the distance between $t_i$ and $t_j$ is less than Min_Node
                    • Source= $t_i$
                    • Destination= $t_j$
                    • Min_Node= the minimum distance between $t_i$ and $t_j$
                • End of Condition
            • End of Loop
        • End of Loop
        • Add “Destination” to semantic network
        • Remove the “Destination” from D'
        • Condition: if Min_Node is equal to 1
            • Connect $t_i$ and $t_j$ via superclass/subclass relation
        • Condition: if Min_Node is greater than 1
            • For each edge between $t_i$ and $t_j$
                • Add the endpoint concept of the respective edge to semantic network
        •   End of Condition
        •   End of Condition
    • End of loop
• End of Loop
Output: the generated semantic network for the D'
```

Figure 4- The semantic network generation Algorithm

4-5- Employing the Semantic Graph Representations for Text document Ranking and Classification

In the final step, in order to demonstrate the effectiveness of the proposed representation model, the documents semantic networks are utilized for ranking and classifying documents according to user preferences. To this end, a hybrid semantic scoring function is introduced. The proposed function estimates the similarity between a document semantic network and user preferences based on four criteria: common information content, common semantic relations, the shortest path between concepts in the hierarchical structure of the ontology and lexical commonalities between concepts. The most similar documents to user preferences are ranked, classified and displayed to the user. It should be noted that the hybrid scoring function is an "ad-hoc" approach. It is designed for document ranking and classification tasks. The following methods are tailored to find similarity based on lexical, semantic and structural features of documents. The following depicts how semantic networks are formalized:

$SN(d_i) = [(d_i), \cup\ rel_i]$.	a generated semantic network for document d_i
$rel_i = \{(t_j, rel, t_k) \mid t, t_k \in (d_i)\}$	a semantic relation between a subject t_j and an object t_k in document d_i
UP	A user profile

4-5-1- Measuring the Commonalities in Information Content and Semantic Features

This method measures the commonality based on the notion of information content (IC) of the Least Common Subsumer (LCS) [27] in WordNet. IC is a measure of the specificity of a concept, and the LCS of concepts A and B is the most specific concept that is an ancestor of both A and B. This method considers the information content (IC) of the LCS concept as the most significant factor in computing the semantic similarity. High IC commonality between two concepts indicates that two concepts are semantically similar. This method is called “normalized Jiang and Conrath measure” [39].

$$IC_{Score}(A,B) = j\&c(A,B) = 1 - \left(\frac{\left[IC_{nrm}(A) + IC_{nrm}(B) - 2 * IC_{nrm}\left(LCS(A,B)\right)\right]}{2}\right) \quad (3)$$

This method computes the semantic similarity between all the possible pair of concepts in the document semantic networks and user preferences. It generates a number between [0, 1] indicating the similarity score.

4-5-2- Measuring the Common Semantic Relations

To this end, two measures are introduced: Explicit_relation: measures the amount of shared information content between the relations in two semantic networks and Implicit_relation: measures how much a document semantic network resembles the user preferences.

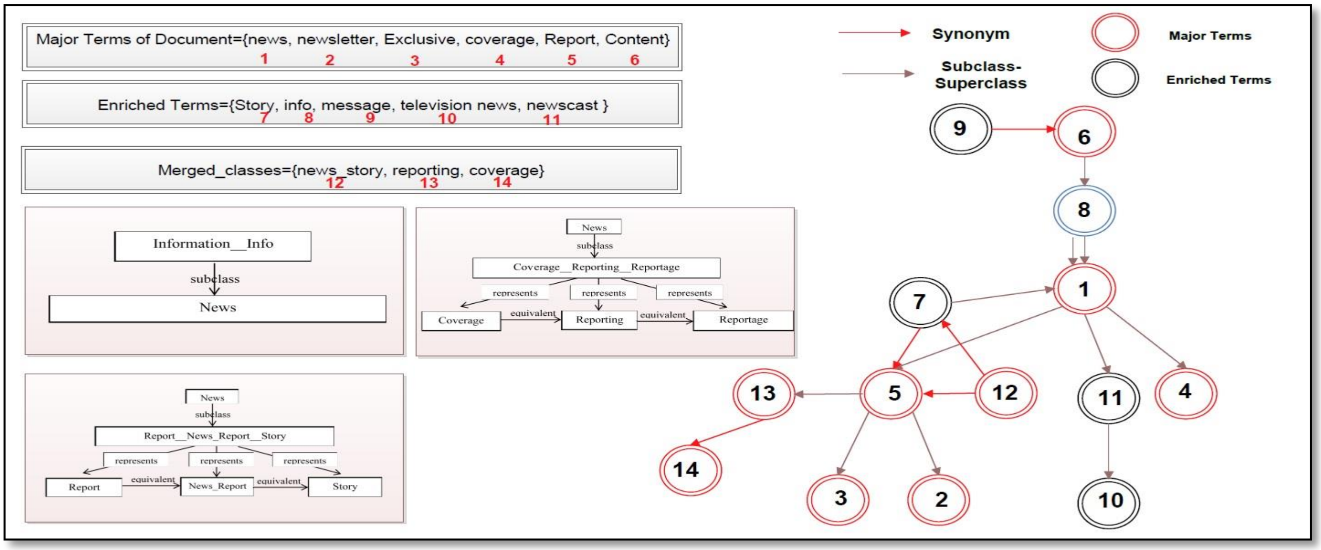


Figure 5- Merging concept maps with the document semantic networks

$$Score_{Explicit}\left(\cup \left(t_j, rel, t_k\right) \in SN(d_i)\right) = \frac{\sum_{all\ the\ triplets} Score\_term_{\text{explicit}(t_j,rel,t_k)}}{number\ of\ triplets\ in\ the\ semantic\ net.}$$

$$Score\_Term_{\text{explicit}(t_j,rel,t_k)} = \begin{cases} \delta_{exp}, & t_j, t_k\ are\ in\ user\ profile \\ 1-\delta_{exp} & o.w. \end{cases} \quad (4)$$

$$Score_{implicit}\left(\cup \left(t_j, rel, t_k\right) \in SN(d_i)\right) = \frac{\sum_{all\ the\ triplets} Score\_relation_{\text{implicit}(t_j,rel,t_k)}}{number\ of\ triplets\ in\ the\ semantic\ net.}$$

$$Score\_relation_{implicit}\left(\left(t_j, rel, t_k\right)\right) = \begin{cases} \delta_{imp}, & (t_j, rel, t_k)\ is\ in\ user\ profile \\ 1-\delta_{imp} & o.w. \end{cases} \quad (5)$$

Where, δ_{exp} and δ_{imp} are thresholds between [0, 1]. Also, A and B denote the subject and object of relations. These methods generate a number between [0, 1] indicating the similarity score.

4-5-3- Calculating the Path between Concepts in the Hierarchical Structure of WordNet

This method calculates the shortest path between concepts in the hierarchical structure of WordNet. This measure is called Wu and Palmer [27].

$$Path_{Score}(A,B) = WordNet\_Path\_based(A,B) = max\left[\frac{2 * Depth(LCS(A,B))}{Length(A,B) + 2 * Depth(LCS(A,B))}\right] \quad (6)$$

Where, $Depth$(A) calculates the depth of concept A and $ength(A,B)$ calculates the shortest path between A and B. This method computes the semantic similarity between all the possible pair of concepts in the document semantic networks and user preferences. It generates a number between [0, 1] indicating the similarity score.

4-5-4 Measuring lexical Commonalities between Concepts

For this purpose, Jaro-Winkler measure is used.

$$Lexical\ (A,B) = d_j + lp\left(1 - d_j\right) \quad (7)$$

Where, d_j is the Jaro similarity score for Concepts A and B. Also, l_p is the length of the common prefix between two concepts and acts as a control parameter. The following equation is used to compute the Jaro similarity score.

$$d_j = \begin{cases} 0 & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|A|} + \frac{m}{|B|} + \frac{m-t}{m}\right) & otherwise \end{cases} \tag{8}$$

In this equation, m is the number of matched characters and t is half the number of characters displacements between two concepts. This method computes the semantic similarity between all the possible pair of concepts in the document semantic networks and user preferences. It generates a number between [0, 1] indicating the similarity score. The overall similarity is calculated by a linear and a weighted combination of the each computed score as follows:

$$\begin{aligned} &Sim_{Score}(SN(d_i), UP) \\ &= \left(k_1 * Score_{Explicit}\left(\bigcup(t_j, rel, t_k) \in SN(d_i)_{\forall t_k \in d_i}^{\forall t_j \in d_i}\right)\right) \\ &+ \left(k_2 * Score_{implicit}\left(\bigcup(t_j, rel, t_k) \in SN(d_i)_{\forall t_k \in d_i}^{\forall t_j \in d_i}\right)\right) \\ &+ k_3. \frac{\sum_{\forall \boldsymbol{B} \in UP}^{\forall \boldsymbol{A} \in d_i} IC(A,B).\, cf\_idf_{score}(A)}{\sum_{\forall \boldsymbol{A} \in d_i} cf\_idf_{score}(A)} \\ &+ k_4. \frac{\sum_{\forall \boldsymbol{B} \in UP}^{\forall \boldsymbol{A} \in d_i} Path(A,B).\, cf\_idf_{score}(A)}{\sum_{\forall \boldsymbol{A} \in d_i} cf\_idf_{score}(A)} \\ &+ k_5. \frac{\sum_{\forall \boldsymbol{B} \in UP}^{\forall \boldsymbol{A} \in d_i} Lexical(A,B).\, cf\_idf_{score}(A)}{\sum_{\forall \boldsymbol{A} \in d_i} cf\_idf_{score}(A)} \end{aligned} \tag{9}$$

Where, k_1, k_2, $k_3, k\ _4$and k_5 are the weighting parameters between [0, 1] and their sum is equal to 1. These parameters are estimated using a subset of the training data.

5-Evaluation

The proposed method is developed for text mining application. The *20Newsgroup* [40] and *Reuters-21578* [41] datasets are used to evaluate the performance of the proposed method. In the case of *20Newsgroup* dataset, the evaluation data are classified in 20 different newsgroups. However, some newsgroups are contextually related and can be further categorized into five broad categories or "Topics", namely computer, politics, science, religion and recreation. 4000 randomly selected documents are used to evaluate the proposed method (800 documents from each topic). Also, in the case of *the Reuters-21578 dataset*, 4000 documents from five categories of dataset, namely "earn", "acq", "interest", "trade" and "crude", are randomly selected (800 documents from each category).

5-1- Evaluating the Performance of the Proposed Method in Classifying and Ranking Documents

For each dataset, five tests are designed to evaluate the performance of the proposed method in classifying and ranking documents according to user preferences. We assume that the user preferences are exactly the same as the contents of the documents in one of the topics. In order to create a semantic representation of user preferences, two queries are created using the document in the respective topic. So, the documents in each topic are analyzed to identify the most frequent and informative concepts/words. Then, a list of candidate concepts/words is formed and presented to the experts to select the concepts/words that can describe the underlying topic the best and the queries for each topic are formed. In the next step, a semantic network representation of each topic is created. In other words, the queries are converted to semantic networks. The created semantic networks represent the user preferences in each test. The semantic networks are then used for classifying documents in their respective topics. In other words, the semantic networks are used for comparing the information content of each topic to the information content of documents. To this end, the hybrid semantic scoring function (introduced in section 4.5) is employed. In each test, the semantic similarity between each document and the semantic representation of the respective topic is measured. Then, each document is classified into the topic that most closely resembles it. The results of five tests are used to evaluate the system performance. To this end, the Mean Average Precision (MAP) score is used. The MAP value is the arithmetic mean of the average precision values for the individual information needs [30]. For a set of given queries q_i, the MAP value is calculated as follows:

$$MAP_i = \frac{1}{m}\sum_{k=1}^{m} Precision\ (R_k) \tag{10}$$

where m is the number of retrieved documents, R_k is the set of ranked results from top until the k-th document.

At first, the validity of the assumption made in section 4.4 is examined. To this end, different percentages of concepts are used to create the document semantic networks. The semantic similarity between the document semantic networks and the queries is then calculated. The performance of the proposed method is then evaluated using the average MAP score of the 10 queries. Also, the overall effect of the enrichment module on the accuracy and precision of the proposed method is evaluated. To this end, the performance of the proposed method with the enrichment module is evaluated against the performance of the proposed method without enrichment module. The results of these experiments are illustrated in Figures 6, 7, 8 and 9. As can be seen in Figures 6 and 7, the information content of the documents in "20newsgroup" dataset are better represented by Top-50% of the concepts. Also, when the semantic networks of documents in the "*Reuters-21578*" dataset are constructed by top-60% of the concepts, the system performs better. Also, the results in Figures 8 and 9 indicate that the enrichment process has a positive impact on the overall performance, even when a small portion of the information is available.

Also, the effect of merging concept maps with the semantic networks on the overall precision of the proposed method is evaluated. The results are depicted in Figure 10, 11. As evident from the results, in both datasets, when a small percentage of concepts are used to generate semantic network, the effect of merging concept maps with the semantic networks is minimal. However, when the amount of available information content grows, the positive effect of merging increases. Therefore, the assumption, that merging information from different knowledge sources yields better precision also holds true. This would also imply that when more information about context is present, the information fusion would result in a higher performance and precision.

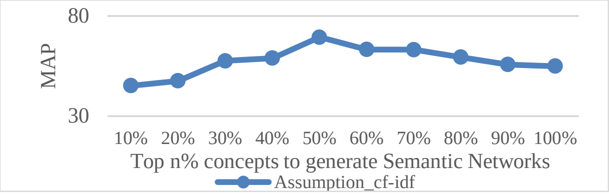


Figure 6- Evaluating the validity of assumption on 20newsgroup

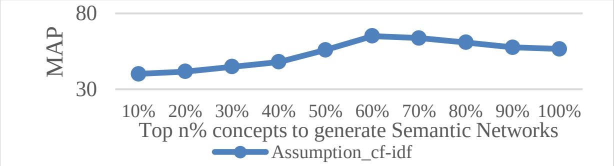


Figure 7- Evaluating the validity of assumption on Reuters-21578

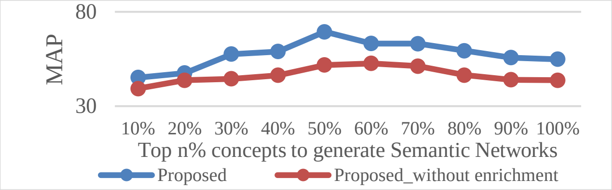


Figure 8- Evaluation of Enrichment Process on 20newsgroup

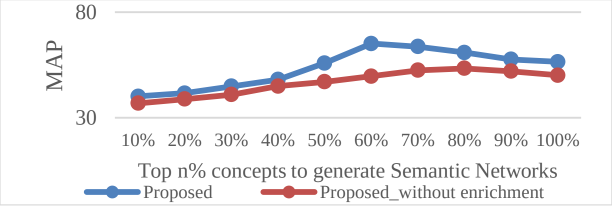


Figure 9- Evaluation of Enrichment Process on Reuters-21578

In the next step, the performance of the proposed method is compared with similar approaches. The first similarity is called the Vector Space Model (VSM)-based (Lucene) scoring function [42]. Also, the performance of the proposed method is compared with MCS-mcs document ranking and retrieval method proposed in [20].

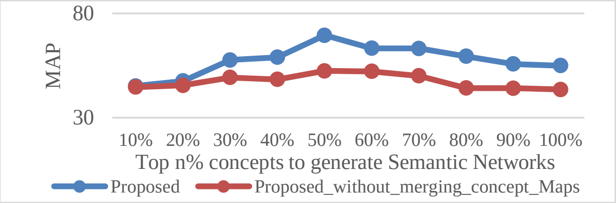


Figure 10- Evaluation the effect of merging concept maps with the semantic networks on the overall performance on 20newsgroup dataset

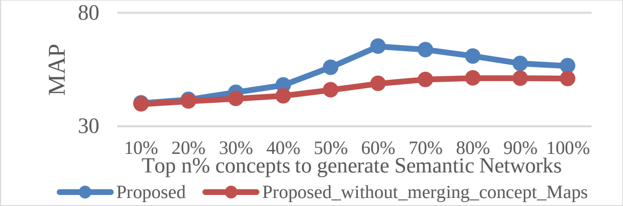


Figure 11- Evaluation the effect of merging concept maps with the semantic networks on the overall performance on the Reuters-21578

The parameters of the VSM-Based model and the MCS-mcs method have been tuned up to achieve the best possible results. The evaluation is carried out by calculating the average MAP score of designed queries for each topic. The results are illustrated in Figures 12 and 13.

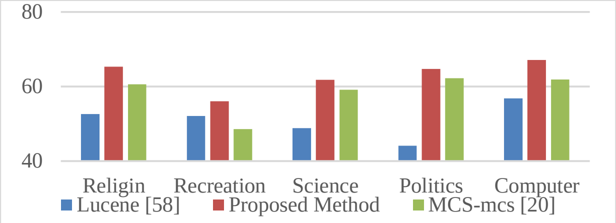


Figure 12- the comparison with similar approaches on 20Newsgroup

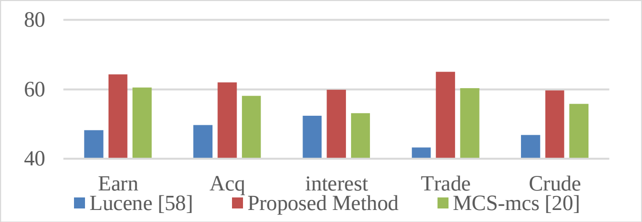


Figure 13- the comparison with similar approaches on Reuters-21578

The results suggest the proposed method outperforms all similar methods and exhibits better performance compared with MCS-mcs method. The results also suggest that the proposed method is effective in correctly classifying and ranking documents according to user preferences.

5-2- Evaluating the Performance of the Proposed Method in Identifying the Most Relevant Documents

The 20 Newsgroup and Reuters-21578 datasets are used. The Accuracy, Precision, Recall and F-measure are used for benchmarking. The evaluation data is identical to the previous experiments. In this stage, a single test for each topic is designed and the performance is evaluated. We assume that in each test, the user preferences are exactly the same as the contents of the documents in one of the topics. At first, the procedure described in section 5.1 is used for creating the semantic network representation of each topic (user queries). Then, the document semantic networks are constructed. For each test, 800 documents are labelled as relevant and 3200 documents are labelled irrelevant. Then, the similarity between document semantic networks and each query is calculated. The system will classify each document in the most similar topic. According to the results, the document is either labelled with TP (True Positive), TN (True Negative), FP (False Positive) or FN (False Negative) labels. Finally, the performance is evaluated using the following measures. The evaluation results are illustrated in Tables 2 and 3.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table 2- The evaluation results on 20 Newsgroup dataset

Test	Topics	Accuracy	Precision	Recall	F-measure
Test#1	Computer	*99.025%*	*97.62%*	*97.5%*	*97.56%*
Test#2	Religion	*98.55%*	*96.26%*	*96.5%*	*96.38%*
Test#3	Politics	*97.675%*	*93.27%*	*95.25%*	*94.25%*
Test#4	Recreation	*96.575%*	*89.7%*	*93.625%*	*91.62%*
Test#5	Science	*96.875%*	*90.52%*	*94.25%*	*92.35%*
Mean Performance		*97.74%*	*93.474%*	*95.425%*	*94.432%*

Table 3- The evaluation results on Reuters-21578 dataset

Test	Topics	Accuracy	Precision	Recall	F-measure
Test #1	Earn	*97.025%*	*91.47%*	*93.875%*	*92.66%*
Test #2	Acq	*96.875%*	*90.42%*	*94.375%*	*92.36%*
Test #3	Interest	*96.3%*	*89.27%*	*92.625%*	*90.92%*
Test #4	Trade	*97.55%*	*93.44%*	*94.375%*	*93.917%*
Test #5	Crude	*96.025%*	*89.13%*	*91.25%*	*90.18%*
Mean Performance		*96.755%*	*90.746%*	*93.3%*	*92.0074%*

As shown in Tables 2 and 3, the proposed method performs well in identifying the most relevant documents. However, the precision values in some of the topics are lower than the mean precision. After careful study of the documents in these topics, we have concluded that high levels of distinction between the topics and some of the documents and also a high degree of overlap between these documents and other topics are the reasons. The mean performance of the proposed method compared with similar text analysis methods are illustrated in Tables 4 and 5.

Table 4-The evaluation results of MCS-mcs and Lucene on 20Newsgroup

Methods	Mean Accuracy	Mean Precision	Mean Recall	Mean F-measure
MCS-mcs	*95.98%*	*89.916%*	*90%*	*89.958%*
Lucene (VSM-based)	*93.935%*	*86.238%*	*82.875%*	*84.5212%*
Proposed Method	*97.74%*	*93.47%*	*95.43%*	*94.43%*

The illustrated results indicate that the proposed method outperforms other similar text analysis approaches. The multi-level representation brings several advantages that help the system outperform other methods. First, semantic, syntactical/structural and lexical features are incorporated into the semantic networks. Second, available information about the context and semantics from different knowledge sources are merged into the semantic networks.

Table 5-The evaluation results of MCS-mcs and Lucene on Reuters-21578

Methods	Mean Accuracy	Mean Precision	Mean Recall	Mean F-measure
MCS-mcs	*95.72%*	*87.364%*	*91.9%*	*89.5774%*
Lucene (VSM-based)	*94.795%*	*86.26%*	*88%*	*87.118%*
Proposed Method	*96.76%*	*90.75%*	*93.3%*	*92.00%*

into the semantic networks. Second, available information about the context and semantics from different knowledge sources are merged into the semantic networks.

Next, we have decided to implement a number of learning methods for topic and text classification and evaluate these methods on Reuters and 20Newsgroup data. The evaluation results are compared with the proposed method. Three well-known machine learning algorithms for classification purposes are considered: Extreme gradient boosting [43], Random Forest [44], and Recurrent Neural Network (RNN)-Long Short Term Memory (LSTM Network) [45].

Extreme gradient Boosting (XGB): This learning model is very similar to the gradient boosting framework. It uses a linear model solver and a decision tree learning algorithm to learn the underlying data model and predict their labels. These models are decision tree-based ensemble models, and are used to reduce bias and variance of learning models.

Random Forest: Random Forest models are ensemble models. The building block of these models are decision tree method. This model generates a number of classification and regression trees (CART) with different samples and variables to learn the underlying data model.

Recurrent Neural Network-LSTM: In these models the output of their activation functions are propagated in two directions (from input to output and from output to input). This feature creates a loop in the network architecture, which acts as “memory” for neurons. This memory allows the neurons to remember what was learned. But when the number of data and consequently the number of layers increases, learning and adjusting the parameters of the earlier layers becomes even more difficult. To overcome this problem, a new type of RNN network called LSTM was developed [45]. In LSTM networks, information flows through a mechanism called “cell state”. The LSTM network consists of a set of memory blocks called cells. Memory blocks are tasked with the remembering of information and memory manipulations are done through three very important mechanisms called "Gates". Forget Gate: This gate is responsible for removing redundant and unimportant information from cell state. Input Gate: it adds information to the cell state. This gate helps the system remember only the important information. Output Gate: this gate is tasked with selecting important information from the current cell state and showing it out as the output [46, 47].

In addition to the extracted standard features, LSTM is trained using the Word_Embeddings features [46, 47]. These features model the contents of documents using a dense vector representation model. In this model, the position of a word in vector space is learned from textual documents. The learning is based on the co-occurrence words in the context. The Word_Embeddings features can be trained using the input data but it is recommended to use a pre-trained one such as Glove, FastText, or most importantly Word2Vec.

The Performance of the Learning-based approaches and the proposed method on Reuters-21578 and 20Newsgroup datasets are illustrated in Tables 6 and 7. Comparing the results suggest that the proposed method, in most cases, achieves better Recall and Accuracy results. However, machine learning methods produce better or comparable precision results compared with the proposed method. Machine learning methods achieve solid results on 20Newsgroup and Reuters-21578 datasets. However, the LSTM Network method has achieved disappointing results compared with the proposed method. It can be concluded that the knowledge-based methods are better in terms of semantic modeling than the Deep Learning methods. It can be concluded that during the numerical transformation of semantic features, a part of semantics will be lost. This can be a contributing factor in scoring disappointing results.

Table 6-The evaluation results of XGB, Random Forest and RNN-LSTM on 20Newsgroup

Methods	Mean Accuracy	Mean Precision	Mean Recall	Mean F-measure
XGB	*92.0518%*	*96.014%*	*92.484%*	*94.1322%*
Random Forest	*95.764%*	*95.76%*	*95%*	*95.372%*
RNN-LSTM	*91.708%*	*95.43%*	*94.156%*	*94.786%*
Proposed Method	*97.74%*	*93.47%*	*95.42%*	*94.43%*

Table 7-The evaluation results of XGB, Random Forest and RNN-LSTM on Reuters-21578

Methods	Mean Accuracy	Mean Precision	Mean Recall	Mean F-measure
XGB	*91.376%*	*92.998%*	*92.852%*	*92.862%*
Random Forest	*93.98%*	*93.526%*	*92.6584%*	*93.046%*
RNN-LSTM	*90.778%*	*94.926%*	*93.47%*	*94.19%*
Proposed Method	*96.75%*	*90.75%*	*93.3%*	*92.99*

5-3- Evaluating the reliability of the proposed method in identifying the correct topic classification

In the final stage, the goal is to examine the reliability of the proposed method in predicting the correct topic classification of documents. The assessment of the reliability of the proposed method is carried out through "Hypothesis testing". For this purpose, 4,000 documents from the "*20newsgroup*" and "*Reuters-21578*"dataset are randomly selected. In the selected collection of data, there are 800 documents representing each of the five topics. The method of evaluating the reliability of the proposed method in predicting the correct topic classification is described for one of the topics and the evaluation for other topics is done in the same way. Assuming that the user preferences are similar to the content of the documents in the "computer" topic, the semantic representation of user preferences is created using the procedure explained in Section (5.1). In the next step, the documents in "Computer" topic are assigned the label "1" and the documents in other topics are assigned the label "-1". Next, the semantic similarity between document semantic networks and the semantic network representation of user preferences is computed using the introduced document ranking and classification method (see section 5.5). If the similarity of a given document to the "Computer" topic is higher than other topics, the prediction label "1" is assigned to this document, otherwise the prediction label "-1" is assigned. The assigned prediction labels act as the topic prediction for each document. In other words, if the true label of each document is equal to its prediction label, the document is classified in its correct topic, otherwise the topic classification of the document is incorrect.

5-4- Hypothesis Testing for Evaluating the Reliability in Predicting the Correct Topic Classification of:

For this purpose, the two-sample t-test is performed. The optimal value (correct prediction label) for documents relevant to "Computer" topic is "1" and the optimal value of irrelevant ones is "-1". The mean and sample standard deviation of the computed prediction labels is *-0.6005* and *0.7997*, respectively. The purpose of two-sample t-test is to test whether the means of two different populations, the population of true labels and prediction labels are equal or not. The two-sample t-test does not assume the equality of variances. Let the null hypothesis be as follows:

H_0: The data of both populations come from independent random samples of normal distribution with equal means. In other words, the propose method is reliable in predicting the correct topic classification.

H_1: The null hypothesis is rejected. In other word, the proposed method is not reliable in predicting the correct topic classification and results may have been obtained by random chance in sample selection.

The significance level is 5% (0.05). In order to assess whether the null hypothesis should be accepted or rejected, first we need to calculate the t-value as follows:

$$t = (\bar{x}_1 - \bar{x}_2)\Big/\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \tag{12}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, s_1 and s_2 are the sample standard deviation, and n_1 and n_2 are the sample size. Table 8. Shows the results on the "Computer" topic. The illustrated results suggest that the proposed method is reliable in predicting the correct topic classification of documents (the Null hypothesis are accepted in all cases) and the results have not been obtained by random chance. Finally, the reliability of the proposed method in predicting the correct topic classification of documents in the Reuters-21578 dataset is assessed. The results are shown in Table 9.

Table 8- The results of Hypothesis Testing on "*20newsgroup*" dataset

Test	Topics	Mean	STD	*p*-value	Null Hypothesis
Test #1	Computer	*-0.6005*	*0.7997*	*0.9777*	*Accepted*
Test #2	Religion	*-0.5990*	*0.8008*	*0.9554*	*Accepted*
Test #3	Politics	*-0.5919*	*0.8064*	*0.6361*	*Accepted*
Test #4	Recreation	*-0.5825*	*0.8129*	*0.3319*	*Accepted*
Test #5	Science	*-0.5835*	*0.8122*	*0.3601*	*Accepted*

Relevant/irrelevant documents: 800/3200, significance level=5%

Table 9- The results of Hypothesis Testing on "*Reuters-21578*" dataset

Test	Topics	Mean	STD	*p*-value	Null Hypothesis
Test #1	Earn	*-0.5895*	*0.8079*	*0.5592*	*Accepted*
Test #2	Acq	*-0.5825*	*0.8129*	*0.3319*	*Accepted*
Test #3	Interest	*-0.5850*	*0.8111*	*0.4051*	*Accepted*
Test #4	Trade	*-0.5960*	*0.8031*	*0.8234*	*Accepted*
Test #5	Crude	*-0.5905*	*0.8071*	*0.5970*	*Accepted*

Relevant/irrelevant documents: 800/3200, significance level=5%

The results demonstrate the reliability of the proposed method in predicting the correct topic classification of documents in the Reuters-21578 dataset.

6- Discussion

One of the most promising aspects of the proposed method is the fusion of information from different sources in the semantic network. The results indicate that the fusion of information will result in better precision; making our assumption about the information fusion true. Since the information come from different knowledge sources, the generated semantic networks are comprehensive. They cover all the available information about the context. The semantic networks coupled with the enrichment module have a positive impact on the performance of the proposed method. As it is evident from the results, the semantic network yields the best results when we employ the top-50% and top-60% of the concepts for "*20newsgroup*" and "*Reuters-21578*"datasets, respectively. It suggests that the proposed representation model would impose less computational burden on the system. Also, the incorporation of enrichment module into the proposed method has a direct effect on generating fully-connected semantic networks. Fortunately, the results are still satisfactory. Comparing the results of the proposed method with the results of machine learning methods is promising. The proposed method provides better accuracy and recall values than these methods. However, machine learning methods achieve better precision values. The surprising thing about the results is lower than expected performance of the LSTM Network method compared to other methods. The reason for such results is that this method is designed specifically for the image processing tasks and also that a part of the semantic is lost during the numerical transformation of semantic features.

7- Conclusion

In order to overcome the lack of semantics and inherent ambiguity associated with textual resources, the structured knowledge of ontology and KBs is integrated in every component of the proposed method. Coupling the content enrichment with the semantic network generation module contributes to the novelty of the proposed method. In the first stage of evaluation, the validity of the assumption, that documents are better represented by the top-n% of the concepts, is assessed. The evaluation results suggest that using the top-50% and top-60% of the concepts for generating the document semantic networks yield the best results for the system. Examining the effect of the content enrichment module on the overall performance shows that this module has a positive effect in improving the performance and precision of the proposed method. Also, the proposed method yields far better results compared with VSM-based and MCS-mcs methods. Creating a unified and comprehensive representation of the documents, by merging concept maps with the semantic networks, is one of the most important contributions of this paper. The results shows that, when sufficient information is available about the information content of the documents, merging concept maps with documents semantic network will improve the performance and precision of the proposed system. Also, the effectiveness of the proposed method in identifying the most relevant information to user preference is assessed. Also, the results illustrate that the proposed method compared with well-known machine learning methods exhibits better or comparable performance. The evaluation results also suggest that the proposed method is reliable and effective in predicting the correct topic classification of documents. The proposed method can be employed in most text mining applications that require semantic representation of the documents, especially when limited information is available.

References

[1] M. Fernández, I. Cantador, V. López, D. Vallet, Pablo Castells, E. Motta, “Semantically enhanced Information Retrieval: An ontology-based approach”, Web Semantics: Science, Services and Agents on the World Wide Web 9, 434–452, 2011.

[2] M. R. Bouadjeneka, H. Hacidc, M. Bouzeghoubd, “Social networks and information retrieval, how are they converging? A survey, a taxonomyand an analysis of social information retrieval approaches and platforms”, Information Systems, Vol. 56, 1-18, 2016.

[3] B. Steichen, H. Ashman, V. Wade, “A comparative survey of Personalized Information Retrieval and Adaptive Hypermedia techniques”, Information Processing and Management, Vol. 48, 698–724, 2012.

[4] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, F.N. Alpaslan, “An ontology-based retrieval system using semantic indexing”, Information Systems, Vol. 37, 294-305, 2012.

[5] A. N. Jamgade, and J. K. Shivkumar, "Ontology based information retrieval system for Academic Library." International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), IEEE, 2015.

[6] T. Roelleke, “Synthesis Lectures on Information Concepts, Retrieval, and Services”, Morgan & Claypool Publishers, 2013.

[7] Z. Hengxiang, J. Lafferty, “A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval”, SIGIR Forum, Vol. 51, 268-276, 2017.

[8] K.M. Kim, J.H. Hong, S.B. Cho, “A semantic Bayesian network approach to retrieving information with intelligent conversational agents”, Information Processing & Management,Vol.43,225–236, 2007.

[9] Y. Bassil, P. Semaan, “Semantic-Sensitive Web Information Retrieval Model for HTML Documents”, European Journal of Scientific Research, Vol. 69, 1-11, 2012.

[10] S.N. B. Bhushan, A. Danti, “Classification of text documents based on score level fusion approach”, Pattern Recognition Letters, Vol. 94, 118-126, 2017.

[11] F. Ramli, S. A. Noah, T. B. Kurniawan, "Ontology-based information retrieval for historical documents", 2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP), 2016.

[12] M. Daoud, L. Tamine, M. Boughanem, “A personalized search using a semantic distance measure in a graph-based ranking model”, Journal of Information Science, Vol. 37, 614–636, 2011.

[13] D. Laura, A. Kotov, and E. Meij, "Utilizing knowledge bases in text-centric information retrieval", In Proceedings of ACM International Conference on the Theory of Information Retrieval., 2016.

[14] M. Banko, O. Etzioni, “The tradeoffs between open and traditional relation extraction”, in Proceedings of ACL-08: HLT, Association for Computational Linguistics, 2008.

[15] B. Mitra, N. Craswel, “Neural Text Embeddings for Information Retrieval”, In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM '17, 2017.

[16] F. Gutierrez, D. Dejing, F. Stephen, W. Daya, Z. Hui. "A hybrid ontology-based information extraction system", Journal of Information Science, Vol. 42, 798-820, 2016.

[17] Y. Gupta, A. Saini, A.K. Saxena, “A new fuzzy logic based ranking function for efficient Information Retrieval system”, Expert Systems with Applications, Vol. 42, 1223-1234, 2015.

[18] M. Daoud, L. Tamine, M. Boughanem, “Towards a graph based user profile modeling for a session-based personalized search”, Knowledge and Information Systems Vol. 21, 365-398, 2009.

[19] G-J. Hahm, J-H. Lee, H-W. Suh, “Semantic relation based personalized ranking approach for engineering document retrieval”, Advanced Engineering Informatics, Vol. 29, 366-379, 2015.

[20] Z. Wu, H. Zhu, G. Li, Z. Cui, H. Huang, J. Li, E. Chen, G. Xu, “An efficient Wikipedia semantic matching approach to text document classification”, Information Sciences, Vol. 393, 15-28, 2017.

[21] J. Yun, L. Jing, J. Yu, H. Huang, “A multi-layer text classification framework based on two-level representation model”, Expert Systems with Applications, Vol. 39, 2035-2046, 2012.

[22] C. Jiang, F. Coenen, R. Sanderson, M. Zito, “Text classification using graph mining-based feature extraction”, Knowledge Based Systems, vol. 23, 302-308, 2010.

[23] H. K. Kim, H. Kim, and S. Cho, “Bag-of-concepts: Comprehending document representation through clustering words in distributed representation.”, Neurocomputing, vol. 266, 336-352, 2017.

[24] W. Jin, Z. wang, D. zhang, J. Yan, "Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification.", Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI Press, 2017.

[25] Y. Li, B. Wei, Y. Liu, L. Yao, H. Chen, J. Yu, W. Zhu, “Incorporating Knowledge into neural network for text representation”, Expert Systems With Applications, In Press - Accepted Manuscript, 2017.

[26]<http://www.loa.istc.cnr.it/DOLCE.html#OntoWordNet>, “Laboratory for applied ontology - DOLCE”, last visited on 19 Feb 2013.

[27] L. Meng, R. Huang, J. Gu, “A review of semantic similarity measures in wordnet,” International Journal of Hybrid Information Technology, vol. 6, 1-12, 2013.

[28] P. Kolb, “DISCO: A Multilingual Database of Distribution-ally Similar Words”, In Proceedings of 9th Conference in Natural Language, 2008.

[29] B. T. McInnes, T. Pedersen, “Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text”, Journal of Biomedical Informatics, Vol. 46, 1116-1124, 2013.

[30] S. Pyysalo, “Part-of-Speech Tagging”, In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology, Springer, 2013.

[31] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky.,”The Stanford CoreNLP Natural Language Processing Toolkit”, In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, System Demonstrations, 2014.

[32] J. Hakenberg, “Named Entity Recognition”, In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology, Springer, 2013.

[33] B. Mohit, “Named Entity Recognition. In: Zitouni I. (eds) Natural Language Processing of Semitic Languages”, Theory and Applications of Natural Language Processing, Springer, 2014.

[34] J. Vilares, M. A. Alonso, M. Vilares, “Extraction of complex index terms in non-English IR: A shallow parsing based approach”, Information Processing & Management, Vol. 44, 1517-1537, 2008.

[35] S.K. Saritha, R.K. Pateriya, “Rule-Based Shallow Parsing to Identify Comparative Sentences from Text Documents”, In: Shetty N., Prasad N., Nalini N. (eds) Emerging Research in Computing, Information, Communication and Applications, Springer, 2016.

[36] M. Baziz, M. Boughanem, S. Traboulsi, “A Concept-based Approach for Indexing in IR”, in the proceedings of INFORSID05, 2005.

[37] C. Biemann, S. P. Ponzetto, S. Faralli, A. Panchenko, and E. Ruppert, “Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation.,” in EACL, 2017.

[38] W. Cohen , P. Ravikumar, S. Fienberg, “A comparison of string distance metrics for name-matching tasks”, American Association for Artificial Intelligence, 73-78, 2003.

[39] Lang, K. “The 20 Newsgroups data set, version 20news-18828”, [last update on Jan 14, 2017], [Online] Available: <http://www. qwone.com/~jason/20Newsgroups>.

[40] N. Seco, T. Veale, J. Hayes., “An Intrinsic Information Content Metric for Semantic Similarity in WordNet”, In Proceedings of European Chapter of the Association for Computational Linguistics, 2004.

[41] W. Zhang, X. Tang, T. Yoshida, “TESC: An approach to TExt classification using Semi-supervised Clustering”, Knowledge-Based Systems, Vol. 75, pp. 152-160, 2015.

[42] S. Langer, J. Beel, “Apache Lucene as Content-Based-Filtering Recommender System: 3 Lessons Learned.”, 5th International Workshop on Bibliometric-enhanced Information Retrieval, 2017.

[43] R. Song, S. Chen, B. Deng, and L. Li, “eXtreme Gradient Boosting for Identifying Individual Users Across Different Digital Devices”, In Proceedings of WAIM, Vol. 9658, pp. 43–54, 2016.

[44] Q. Wu, Y. Ye, H. Zhang, M. Ng and S. Ho, " ForesTexter: An efficient random forest algorithm for imbalanced text categorization", Knowledge-Based Systems, Vol. 67, pp.105-116, 2014.

[45] G. Rao, W. Huang, Z. Feng and Q. Cong, "LSTM with sentence representations for document-level sentiment classification", Neurocomputing, Vol. 308, pp.49-57, 2018.

[46] C . Olah, "Understanding LSTM Networks", [last update on Aug 27, 2015], [Online] Available: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/s>, [Retrieved on Nov 01, 2018].

[47] P. Srivastava, "Essentials of Deep Learning : Introduction to Long Short Term Memory", [last update on Dec 10, 2017], [Online] Available: < https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/ >, [Retrieved on Nov 01, 2018.

Morteza Jaderyan received his Master degree in computer engineering from Shahid Chamran University of Ahvaz, Iran. He is currently in pursuit of Ph.D. degree from Bu-Ali university of Hamadan. His main Research Interest is Artificial Intelligence, Information Retrieval and Management Systems, Semantic Web and Web Engineering, Knowledge Management Systems, Mobile Robotics, Machine learning and Intelligent Systems.

Hassan Khotanlou received the B.E. and M.E. degrees in Computer Engineering from Shiraz University in 1998 and the Ph.D. degree in Computer Engineering (Machine Vision) from Telecom ParisTech in 2008. Since 2008, He has been with the Bu-Ali Sina University and currently He is an Associate Professor of Computer Engineering and Head of the Robot Intelligence and Vision (RIV) Research Group. His current research interests include evolutionary computation, Medical Image Processing, Statistical Pattern Recognition, Deep Learning and NLP.

Cooperative Game Approach for Mobile Primary User Localization Based on Compressive Sensing in Multi-antenna Cognitive Sensor Networks

Maryam Najimi*
Department of Electrical and Computer Engineering, University of Science and Technology of Mazandaran , Behshahr, Iran
M.najimi@mazust.ac.ir

Received: 04/Jun/2019 Revised: 30/Nov/2019 Accepted: 30/Dec/2019

Abstract

In this paper, the problem of joint energy efficient spectrum sensing and determining the mobile primary user location is proposed based on compressive sensing in cognitive sensor networks. By utilizing compressive sensing, the ratio of measurements for the sensing nodes are considerably reduced. Therefore, energy consumption is improved significantly in spectrum sensing. The multi-antenna sensors is also considered to save more energy. On the other hand, multi-antenna sensor utilization is a proper solution instead of applying more sensors. The problem is formulated to maximize the network lifetime and find the mobile primary user position by sensors selection under the detection performance and accuracy of localization constraints. For this purpose, a cooperative game is proposed to study this problem. It is shown that with the proposed game, the network lifetime is maximized while the proper sensors which participate in spectrum sensing and primary user localization are determined. Simulation results show that the network lifetime is improved while the detection performance constraint is satisfied and the location of the primary user is determined with high accuracy.

Keywords: Cooperative Spectrum Sensing; Compressive Sensing; Mobile Primary User Localization; Detection Performance; Game Theory.

1- Introduction

Recently, cognitive radio (CR) has a lot of attention due to its capability of exploiting white spectrums and improving spectral utilization efficiency [1], [2]. This capability is introduced as spectrum sensing which is a technique for determining the existence of the primary users (PUs). However, in an unauthentic network, malicious users (MUs) may imitate the features of PUs and transmit in the cognitive sensing band by reconfiguring the air interface of CR. These are introduced as primary user emulation attacks (PUEA) [3]. According to this, cognitive radio users mistake the adversary of CRs as primary users. Therefore, this will lead to wastage of spectrum resources and interference to the spectrum management of cognitive radio networks. In this case, information about PUs location could enable several capabilities in cognitive radio networks, including improved spatio-temporal sensing, intelligent location-aware routing, as well as aiding spectrum policy enforcement [4]. In order to obtain the primary user position in cognitive radio networks (CRNs), CRNs can be considered as wireless sensor networks (WSNs) [5]. Therefore, the sensors sense the spectrum band in WSN. However, the limited energy budget and low computational capability of each node are the main constraints of WSNs. Although, the sensor nodes have these constraints, the fusion center (FC) usually has a comparatively high computational capability. In fact, the sensors sense the spectrum band and transmit their results to FC. Then, FC makes a final decision about the channel status using a fusion rule. The fusion rules can be the hard decision rules such as OR, AND or K-out-of-N or soft decision rules such as Maximum Ratio Combining (MRC) for combining the local reports from different sensors.

Due to the limitations of the sensors capabilities, compressive sensing (CS) was introduced. Compressive sensing has a surprising property in which sparse signals can be recovered from far fewer samples than the Nyquist-shannon sampling theorem [6]. In fact, compressing sensing technology can be considered as an important method for spectrum sensing since wireless signals in these networks are typically sparse in frequency domain [7]. CS theory states that a signal can be reconstructed from a smaller number of linear measurements if it is sparse or compressible in a certain basis. Therefore, the conventional and compressing sensing techniques can be compared to illustrate that CS improves the energy efficiency and also the lifetime for cognitive sensor network [8]. In [9], the recent advances of compressive sensing in wireless sensor networks is stated. In this paper,

* Corresponding Author

CS can be efficiently applied to solve the problems specific to WSNs.

In [10], an energy-efficient spectrum sensing scheme is proposed based on game theory for cognitive radio sensor networks to prolong the network lifetime. However, there is not any mathematic analysis and simulation results to show the lifetime improvement in these networks. In [11], the problem of energy efficient cooperative spectrum sensing in multi-antenna cognitive sensor networks is investigated by sensing nodes selection while maintains the detection performance constraints. In [12], received-signal strength (RSS) method is used to estimate the PU position. In [13], primary user localization is also considered by utilization of the received-signal strength (RSS) and direction-of-arrival (DoA) estimation from sectorized antenna. In this case, a sectorized antenna is defined as an antenna that is set to different operating modes which leads to the selection of the signals that arrive from within a certain range of angles. In [14], a sparse vector is obtained by formulating the spectrum sensing and primary user localization problem. The CS technology is applied to reconstruct the information of the primary users. In [15] a decentralized way is proposed to solve the spectrum sensing and primary user tracking problem. However, existing works often investigate the CR network with static primary users and network lifetime improvement is not considered in these papers. In [16], the compressed sensing approach is proposed to overcome hardware limitations and acquire the measurements of the signals at the Nyquist rate when the spectrum is large. In [17], a data gathering algorithm is designed to do compressive sensing and select sensors in temporal and spatial domains, respectively. In [18], a cooperative support identification scheme is proposed for recovery of the compressive sparse signal via resource-constrained wireless sensor networks.

In [19], the authors apply a primary user localization algorithm based on compressive sensing in cognitive radio networks. They use the correlation coefficients between primary signal and secondary users (SUs) to estimate the primary user position. However, they do not consider the lifetime improvement in their work.

Summarily the contributions in this paper are as follows

- The problem is the lifetime maximization of cooperative spectrum sensing in wireless cognitive sensor networks by proper selection of the sensors for spectrum sensing and mobile primary user localization under the constraints on the false alarm and detection probabilities and accuracy of the mobile primary user localization. In this case, the distances between each node and FC are assumed to be known.
- An approach is also used based on compressive sensing (CS) framework to monitor the primary user localization and reduce the number of required samples to reconstruct the sampled signal at the fusion center (FC) and so decrease the energy consumption of the sensors. To save more energy, the multi-antenna structure is considered for each sensor and MRC is utilized as the diversity technique for antenna’s signal combination. Therefore, the network lifetime is improved significantly.
- The optimum solution for the problem is the exhaustive search algorithm. This method cannot be used in practice due to its high computational complexity. Hence, a cooperative game is proposed to maximize the lifetime of the network and find the location of the mobile primary user with high accuracy.
- The numerical results analyze the proposed algorithm to find the solution, energy consumption, detection performance and accuracy of mobile primary user localization in different conditions.

The rest of the paper is organized as follows. The system model of a wireless sensor network is introduced in section 2. The problem is formulated in Section 3. In Section 4, the iterative algorithm is proposed. In section 5, the performance evaluation is stated and conclusions are finally drawn in Section 6.

2- System Model

A grid network is considered involving one primary user, Msensor nodes and one fusion center (FC). Each sensor or primary user locates at the center of one certain gird. It is assumed that each primary user moves in each frame duration as denoted by T. On the other hand, in each frame duration, primary user stays on a new position. Fig.1 shows the spectrum sensing model in a wireless cognitive sensor network. During the sensing time, each node which participates in spectrum sensing, applies the energy detection to detect the primary user existence. Then, each sensor sends its result to the fusion center to make a final decision about the channel state. In fact, cooperative spectrum sensing is used as a solution to alleviate the fading and shadowing effects in wireless channels [20]. In order to determine the PU activity, each observation sample $X_j[k]$, has the data model as

$$H_1: \quad X_j[k] = s_j[k] + u_j[k] \tag{1}$$

$$H_0: \quad X_j[k] = u_j[k] \tag{2}$$

Where, $s_j[k]$ is the received primary user signal at the jth node while $u_j[k]$ is a Gaussian noise with zero mean and variance,σ_u^2.

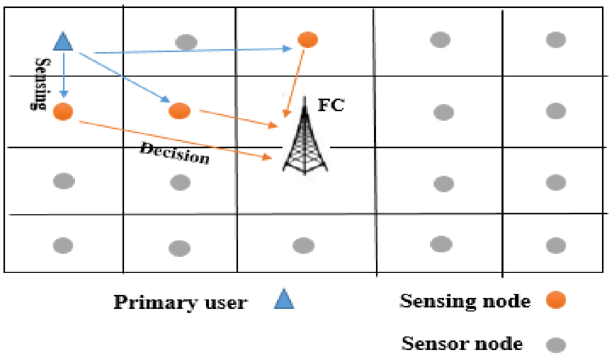


Fig.1 Cooperative spectrum sensing structure using the sensor nodes.

In spectrum sensing, the probability of detection states the protection probability of the primary user transmission from the interference made by the secondary user transmission while the probability of false alarm shows the opportunity of utilizing the idle band by the secondary users. The probability of false alarm in the jth cognitive sensor is given by using the energy detector method as follows [21]

$$P_{f_j} = Q(\left(\frac{\epsilon}{\sigma_u^2} - 1\right)\sqrt{\delta f_s}) \qquad (3)$$

Where, Q is the complementary distribution function of the standard Gaussian, ϵ is the detection threshold, δ is the sensing time and f_s is the sampling frequency. Also, under the hypothesis H_1 , the probability of detection for the jth cognitive sensor is stated by [21]

$$P_{d_j} = Q(\left(\frac{\epsilon}{\sigma_u^2} - \gamma_j - 1\right)\sqrt{\frac{\delta f_s}{2\gamma_j+1}}) \qquad (4)$$

Where γ_j is the received primary user signal to noise ratio at the jth cognitive sensor. Although, increasing the number of sensors helps to have more options for sensing nodes selection and therefore, it improves the detection performance. However, in recent years, multi-antenna utilization leads to decrease the cost of implementation. In fact, the soft decision strategy is utilized for signals combination of the antennas in each sensor. We apply MRC as a combination scheme. The distances between antennas are also considered more than half of the wavelength. In MRC, the effective SNRs for combined signals is defined as [11]

$$\gamma_{j,MRC} = \frac{P_t(\sum_{l=1}^{L}|h_{j,l}|^2)^2}{\sigma_{MRC}^2} \qquad (5)$$

Where, $h_{j,l}$is the channel gain between the lth antenna of the jth node and primary user and P_t is the transmitted power from the primary user. Variances of the effective noise can also be computed as

$$\sigma_{MRC}^2 = \sum_{l=1}^{L}|h_{j,l}|^2\sigma_u^2 \qquad (6)$$

It should be noted that L is the number of antennas in each node. Therefore, the local probabilities of detection and false alarm are obtained by replacing Eq. (5) and Eq.(6) instead of γ_j and σ_u^2 in Eq.(3) and Eq.(4), respectively. However, fading and shadowing effects alleviate the detection performance. Hence, cooperative spectrum sensing is proposed to solve this problem. It means that each sensor sends its decision on the primary user existence to FC to make a final decision about the channel status using a combinational rule. In this paper, OR rule is considered as a combinational rule. It means that if at least one node reports that the spectrum band is busy, the final decision is the primary user activity. According to this definition, the global probabilities of detection and false alarm are obtained as

$$Q_f = 1 - \prod_{j=1}^{M}(1 - P_{f_j}) \qquad (7)$$

$$Q_d = 1 - \prod_{j=1}^{M}(1 - P_{d_j}) \qquad (8)$$

However, in [22], it is shown that participating all nodes in spectrum sensing is not necessary to improve the detection performance. Therefore, probability of participating in spectrum sensing is an important issue for each sensor. So, Eq.(7) and Eq.(8) are modified as follows

$$Q_f = 1 - \prod_{j=1}^{M}(1 - f_i P_{f_j}) \qquad (9)$$

$$Q_d = 1 - \prod_{j=1}^{M}(1 - f_i P_{d_j}) \qquad (10)$$

Where, f_i is the sensing probability, while $f_i \in [0,1]$. However, in an untrusted environment, spectrum sensing in cognitive radio networks not only requires to detect the signal in the spectrum band, but also it should confirm that whether the signal is from legal primary users or malicious users. Therefore, in this paper, the purpose is to find the location of the mobile primary user with high accuracy by selection of the appropriate sensing nodes so that the network lifetime is maximized and the detection performance constraints are satisfied. On the other hand, compressive sensing is introduced into the primary user localization to decrease the number of sensing sensors. In this case, the energy consumption is reduced and therefore the network lifetime is improved. It should be noted that the average energy consumption in spectrum sensing has two parts: the energy consumption for sensing the channel which is assumed to be the same for all sensors and it is denoted by E_s. The second part is the energy consumption for transmission of the sensing results to FC which is denoted by E_{t_j}. Since the transmission energy has an important effect on the battery lifetime, it cannot be ignored. The energy model in [22] is applied for the radio hardware energy dissipation as follows

$$E_{\mathrm{tj}}(d_j) = E_{t-elec} + e_{amp} {d_j}^2 \qquad (11)$$

The first item presents the transmitter electronics energy, while the second part presents the energy consumption for amplifying the radio. d_j is the distance between the jth sensor and FC. Therefore, the total energy consumption for cooperative spectrum sensing is states as

$$E_T = \sum_{j=1}^{M} (E_s + E_{\mathrm{tj}}) \qquad (12)$$

The area which nodes and primary user are distributed, is divided into N girds of the same size. Every node or primary user locates at the center of one certain gird. The received signal energy at grid i from primary user at grid j is considered using the Rayleigh energy decay model in [23] as follows

$$R_{i,j} = \frac{P_0 h_{i,j}}{{d_{i,j}}^2} \qquad (13)$$

Where, P_0 is the power density at primary user. The location of the primary user is denoted by a vector $P_{N\times 1}$. Each element of the vector is zero except the grid which the primary user exists. The value of this element is P_0. It means that P is sparse. In order to localize the primary user, the conventional method is to place the sensor nodes at the monitored environment and obtain the snapshots of RSS. Thus, the received RSS,X, is a $N \times 1$ vector as[19]

$$X = \Psi \mathrm{P} \qquad (14)$$

Where, Ψ is a $N \times N$ matrix represent the primary user energy decay model which is defined as

$$\Psi = \begin{bmatrix} \frac{h_{11}}{{d_{11}}^2} & \frac{h_{12}}{{d_{12}}^2} & \cdots\cdot & \frac{h_{1N}}{{d_{1N}}^2} \\ & & \cdots & \\ \frac{h_{N1}}{{d_{N1}}^2} & \frac{h_{N2}}{{d_{N2}}^2} & \cdots\cdot & \frac{h_{NN}}{{d_{NN}}^2} \end{bmatrix} \qquad (15)$$

Using compressive RSS measurements instead of collecting all measurements, $Y_{M\times N}$ is defined which has the relationship to X. Therefore, we have

$$Y = \Phi X + W \qquad (16)$$

Where,$\Phi_{M\times N}$ is the measurement matrix which $\Phi(i,j)$ represents the probability of sensing the channel by ith sensor in jth grid while W is the additive Gaussian white noise matrix. In order to reconstruct X which is K sparse (in this case, $K = 1$) in Ψ, Mcompressive measurements are required where $M = O(K \log N)$. For this reconstruction, a convex optimization problem should be solved which has the following form

$$\arg\min ||P||_{l_1} \quad , s.t. ||\Phi X - Y||_{l_2} < \epsilon \qquad (17)$$

Where l_1 and l_2 are the corresponding norms in Eq. (17). This convex problem can be solved by linear programming and the global optimal solutions can be achieved.

3- Problem Formulation

As stated before, the problem is the selection of the sensors which determine the primary user existence and find its location, so that the network lifetime is maximized and the detection performance and accuracy of the primary user localization constraints are satisfied. For this purpose, the behavior of sensors is modeled as a cooperative game, in which, the ith sensor (ith player in the game) can determine its sensing probability. Therefore, the utility function of the ith node (lifetime of the ith node) is defined as

$$u_i = \frac{E_{R_i}}{E_{T_i}} \qquad (18)$$

Where, E_{R_i} is the remaining energy after transmission of the result to FC and E_{T_i} is the energy consumption for the ith sensor. It should be noted that maximization of the utility u_i depends not only on f_i, the strategy taken by sensor i , but also on the strategy set of other sensors in the game. The strategy combination space of sensors in the game is considered as follows

$$F = \{f | f = (f_1, f_2, \ldots, f_M), \forall i \in M, 0 \le f_i \le 1\} \quad (19)$$

On the other hand, each sensor for maximization of its utility function has to consider its strategy as well as the strategies taken by the other sensors. On the other hand, all players want to maximize the network lifetime (aggregate utility) while maintaining fairness. Hence, the aggregate utility is given by

$$U_L = \sum_{i=1}^{M} f_i \, u_i \qquad (20)$$

As it is said, our goal is to maximize the aggregate utility in the system which is equal to maximize the network lifetime. It should be noted that in cooperative spectrum sensing, it is desirable to have higher global probability of detection to decrease the interference of the primary user transmission with the secondary users' activities and also lower global probability of false alarm to have more opportunity for spectrum utilization. According to this, the problem can be formulated as

$$P1{:}_{f_i}\, Max\ U_L \qquad (21)$$

$$s.t.\quad Q_d \geq \beta \qquad (21\text{-}1)$$

$$Q_f \leq \alpha \qquad (21\text{-}2)$$

$$\sum_{i=1}^{M} {f_i}^2 \leq M \qquad (21\text{-}3)$$

$$||\Phi X - Y||_{l_2} < \zeta \qquad (21\text{-}4)$$

Eq. (21-1) and Eq.(21-2) show the detection performance constraints while Eq.(21-3) states the probability of participating in spectrum sensing for each sensor node . ζ in Eq.(21-4) is a threshold which shows the accuracy of mobile primary user localization using compressive sensing. The optimal solution for this problem is the exhaustive search method with high complexity with the order of $O(M!)$. Although, heuristic methods are alternative approaches, but they lead to the sub-optimal solutions. Thus, it is desirable to search a distributed approach with linear complexity and optimal solutions for cognitive sensor network. Therefore, the primary user localization (PUL) is modeled as a cooperative game which is defined as

Definition 1: A PUL game can be stated as (M, S, U) in which, M is the number of sensors (players), S is the set of strategies and U is the set of utility functions. Each sensor (player) i determines its strategy $s_i$and gets the utility (payoff)u_i. In fact, u_i is the function of strategy combination set $(f_1, f_2, \dots, f_M)$.

In the game theoretic scenario, it is essential to obtain an equilibrium state for the game which is called Nash equilibrium (NE) [24]. A Nash equilibrium offers a stable solution in which the players achieve a point where no player wants to deviate. On the other hand, a Nash equilibrium states the best status for all players. It should be noted that the efficiency of NE is dependent on utility function of the players. From another respective, for achieving the global optimization, the utility function should be designed so that NE exists. The following conclusion shows the existence of Nash equilibrium in the primary user localization game.

Prposition1: A Nash equilibrium exists in the primary user localization (PUL) game.

Proof: We hope that when all sensors reach the NE, mobile primary user localization and network lifetime maximization are also obtained. To get the NE point, the Lagrangian function is applied as follows

$$L = \sum_{i=1}^{M} f_i\, u_i + \lambda(Q_d - \beta) - \eta\big(Q_f - \alpha\big) - \xi\big(\sum_{i=1}^{M} {f_i}^2 - M\ \big) - \vartheta(\big||\Phi X - Y|\big|_{l_2} - \zeta) \qquad (22)$$

Where λ , η , ξ and γ are the Lagrangian multipliers. It should be noted that for each node, P_{f_j} is not dependent on SNR. On the other hand, all sensors have the same local probability of false alarm. It means that, the global probability of false alarm constraint determines the maximum number of sensing nodes as follows

$$n \leq \frac{\ln(1-\alpha)}{\ln(1-Q_f)} = M\_s \qquad (23)$$

Where, M_s denotes the maximum number of sensing nodes. Therefore, in Eq. (22), the global probability of false alarm constraint is removed, while the maximum number of sensing nodes constraint is considered in sensor selection. Therefore, we have

$$\frac{\partial L}{\partial f_i} = u_i - 2f_i + \lambda P_{d_i} - 2f_i\xi - \vartheta \sum_{i=1}^{M} (f_i \frac{h_{ij}}{{d_{ij}}^2} - y(i,j)) \qquad (24)$$

So, we have

$$f_i = \frac{u_i + \lambda P_{d_i} - \vartheta \sum_{i'=1\neq i}^{M} (f_{i'} \frac{h_{i'j}}{{d_{i'j}}^2} - y(i',j))}{2\xi + \vartheta(\frac{h_{ij}}{{d_{ij}}^2} - y(i,j))} \qquad (25)$$

According to Eq. (25), for any two sensors $i, j \in N$, if $u_i > u_j$, $P_{d_i} > P_{d_j}$ and $d_j < d_i$, then the condition $f_i > f_j$ is satisfied. Therefore, the sensor selection leads to the energy balancing between nodes and therefore, the Nash equilibrium is obtained. In a Nash equilibrium, optimal probability of spectrum sensing is achieved for each player. Due to the existence and uniqueness of the Nash equilibrium, an iterative algorithm is used to find the equilibrium. Therefore, the optimal network lifetime can be obtained by finding the optimal NE point of the game.

4- Proposed Iterative Algorithm for Solving the Problem

In iterative algorithm, the optimum value of λ and ϑ are obtained. At each iteration, first, the sensors with enough energy (i.e., $(E_{R_i} - E_{T_i} > 0)$) are candidates for spectrum sensing. Then, the probability of spectrum sensing (f_i) is calculated for each sensor. The nodes with higher probability of spectrum sensing are selected for spectrum sensing until the detection performance and accuracy of compressive sensing constraints are satisfied. Note that maximum number of sensing nodes is determined using the global probability of false alarm constraint. Then, λ and ϑ are updated according to the subgradient method. Therefore, we have [25]

$$\lambda^{k+1} = \lambda^k - \Gamma_1^k (Q_d - \beta) \qquad (26)$$

And

$$\vartheta^{k+1} = \vartheta^k + \Gamma_2^k \, ||\Phi X - Y||_{l_2} \qquad (27)$$

The step size used in the proposed algorithm is $\Gamma_i^k = \frac{w_i}{\sqrt{k}}$ $i = 1,2$, where $w_i \gg 1$. In each iteration, the total energy consumption is also calculated. This proposed algorithm ends when the convergence metric is satisfied. Then, according to the sensing players, the primary user position is obtained. In fact, in the proposed algorithm, using the optimal value of λ and ϑ, the proper nodes are selected for spectrum sensing and primary user localization. Therefore, the network lifetime and primary user position are determined. Pseudo code for Primary User Localization and Lifetime Maximization Algorithm (PULLM) is shown below.

```
Algorithm1: PULLM Algorithm
----------------------------------------
$\lambda_{min}$=0
$\lambda_{max}$=$\chi$
$\vartheta_{max} = \upsilon$
$\vartheta_{min} = 0$
$\lambda$=$\lambda_{max}$ (input)
$\vartheta = \vartheta_{max}$ (input)
Iteration= $\alpha$(a big number)
    $\varepsilon_1$= small parameter
    $\varepsilon_2$= small parameter
While ($|\lambda^{k+1} - \lambda^k| > \varepsilon_1$ && $|\vartheta^{k+1} - \vartheta^k| > \varepsilon_2$ )
    number of sensing sensors($n$)=0
    Determine the nodes which have enough energy, the number of nodes is $count$
    Compute $f_i$ for each node according to Eq.(25)
    $n = 0$
    While (select $n$ nodes with higher probability of spectrum sensing <min( $count, M\_s$))
        Compute $Q_d$
        If $Q_d > \beta$ , break , End
        $n = n + 1$
    End
    Compute energy consumption according to Eq. (12)
    Compute remaining energy for the sensing nodes
    Compute the accuracy for mobile primary user localization
    Update $\lambda$ and$\gamma$ according toEq. (26) and Eq.(27)
End
The network lifetime and the location of the primary user are obtained (outputs).
```

Fig.2 Pseudo code for the proposed algorithm

5- Performance Evaluation

To evaluate the performance of PULLM game approach, it is assumed that the reports are generated and transmitted per round. In each round, the primary user moves with uniform distribution in a square field with a length of 700 m. FC is located in the center of the environment and the number of nodes is changing from 5 to 50 in values. The nodes which their remaining energy is more than their energy consumption, are supposed to be alive. According to this, the lifetime definition is the time in which more than 25% of sensors are alive. The initial energy for each node is set to be 0.2 mJ. In simulation results, the design parameters are set as $\alpha = 0.1$ and $\beta = 0.9$. Every simulation result in this section is averaged over 10000 realizations.

We use 2.4 GHz IEEE 802.15.4/ZigBee as the communication technology for the cognitive sensor network. The simulation parameters are set in simulation table1 [23], [26].

Table1: Simulation parameters

Parameter	value
Maximum distance between two nodes	700m
Number of Nodes	5-50
Initial energy of the nodes	0.2mJ
E_s	190nJ
E_{t-elec}	80nJ
e_{amp}	40.44pJ/m^2

The proposed algorithm is compared with the following algorithms:

Network Lifetime Improvement with Sensor Selection (NLISS): In this algorithm, the sensors with more remaining energy, probability of detection and less energy consumption are selected for primary user localization. For lifetime maximization in this algorithm, the convex optimization method is used. In this case, the Lagrangian multipliers are updated using subgradient search method [26].

Random Sensor Selection for Network Lifetime (RSSNL): In this algorithm, the sensors are selected randomly for mobile primary user localization and spectrum sensing. This algorithm has the low complexity to solve the problem.

In Fig.3, the minimum remaining energy of the sensors is shown versus different number of nodes. According to Eq. (21), the algorithms attempt to balance the remaining energy of the nodes to maximize the network lifetime.

PULLM algorithm and PULLM algorithm have more minimum remaining energy. Because these algorithms consider local probability of detection, energy consumption, remaining energy and accuracy of primary user localization for sensors selection. RNLSS algorithm has low remaining energy due to the random selection of the sensing nodes. It should be noted that multi-antenna sensors help to balance the remaining energy between sensors. It is noted that the chance of sensor selection increases as the number of nodes is increased. It leads to have more minimum remaining energy in large number of nodes.

Fig.4 shows the energy consumption versus different number of sensors. PULLM using multi-antenna sensors algorithm consumes less energy because multi-antenna sensors are very effective for saving energy especially in large environments. Another important issue is the compressing sensing method which saves the energy consumption. NLISS algorithm consumes more energy than proposed algorithms. This shows that the proposed algorithms are energy efficient in spectrum sensing and primary user localization. It should be noted that the algorithms are compared when they satisfy the detection performance and the accuracy of mobile primary user localization constraints.

Fig.5 shows the successful percent of finding the solution for algorithms in different number of sensors. Howevere, sometimes the problem has no solution. It means that the constraints of the problem are not satisfied by selection of all alive sensors. According to Fig.5, if the problem has the solution, the proposed algorithm with multi-antenna sensors has the most success in finding the solution. NLISS algorithm has less success in finding the solution, because it is not considered the accuracy of the primary user localization in this algorithm. RSSNL algorithm has the minimum percent of success in finding the solution due to the random selection of the sensing nodes. According to this experiment, by increasing the number of nodes, in RSSNL algorithm, it is possible to select the nodes with lower probability of detection and therefore, this metric is decreased. It should be noted that this metric is very important to determine the efficient algorithm.

In Fig.6, the mean error of the network for different number of sensors is illustrated. In fact, this parameter shows the accuracy of the algorithms in mobile primary user localization. According to Fig.6, RSSNL algorithm has the maximum mean error due to the random selection of the nodes for primary user localization and spectrum sensing while the proposed algorithms have the minimum mean error. Because these algorithms consider the accuracy of primary user localization in sensing nodes selection. On the other hand, the sensors which are located near to the primary user have more opportunity for spectrum sensing. However, in NLISS algorithm, decreasing of the mean error of the primary user tracking is not considered for sensor selection.

Fig.7 and Fig.8 are the focused versions of Fig.6. It is obvious that the proposed algorithms have the least mean error in primary user localization due to considering this metric in sensing node selection while in NLISS algorithm, this metric is not important in sensor selection.

Fig.9 shows the utility function of the algorithms versus different number of nodes. According to Eq. (21), the purpose is to maximize the aggregate utility function. In fact, this parameter equals to the network lifetime and increasing of the aggregate utility function improves the network lifetime. It is illustrated that PULLM algorithm with multi-antenna has more utility function than PULLM algorithm. It means that the multi-antenna technique improves the lifetime of the network due to the selection of the sensors based on their energy consumption, remaining energy, global probability of detection and accuracy of primary user localization. It should be noted that by increasing the number of the sensors, the chance of the sensor selection also increases and therefore, the utility functions of the proposed algorithms are improved.

Fig.10 shows the changes of the utility function versus the iterations in which the Lagrangian multipliers are updated. The iterations are changed between 340 and 570. Number of nodes is set to 50. According to Fig.10, in 520th iteration, the utility function converges to the fixed value.

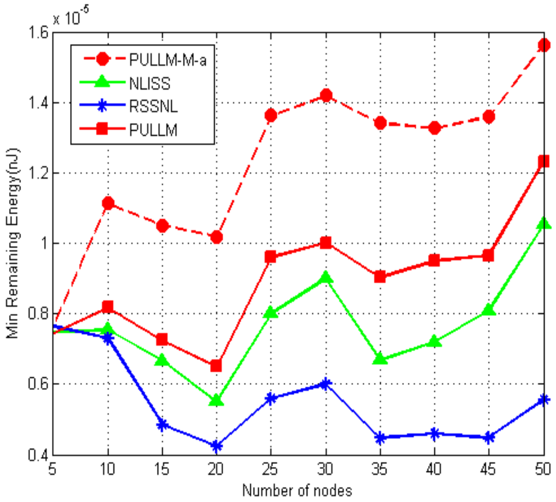


Fig.3 Minimum remaining energy vs. different number of nodes

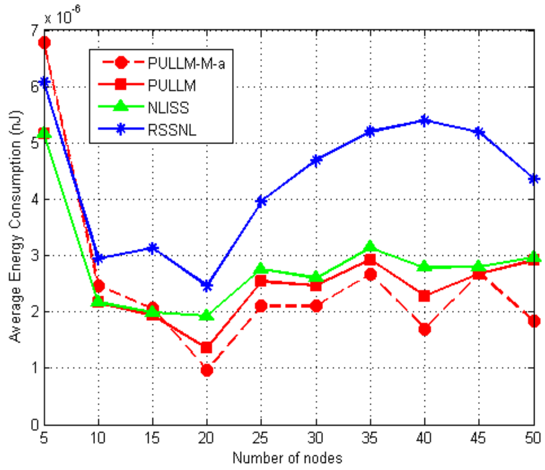


Fig.4. Average energy consumption vs. different number of nodes

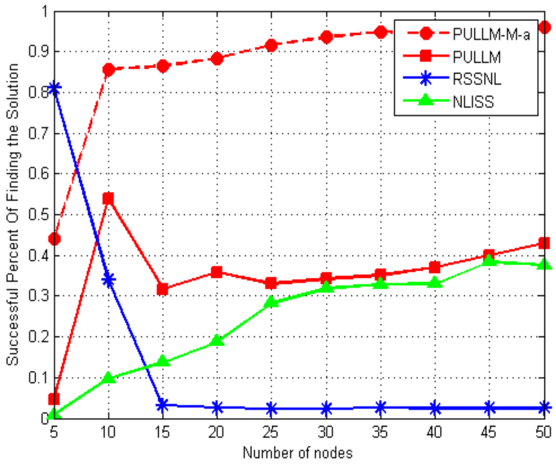


Fig.5 Successful percent of finding the solution vs. different number of nodes

v

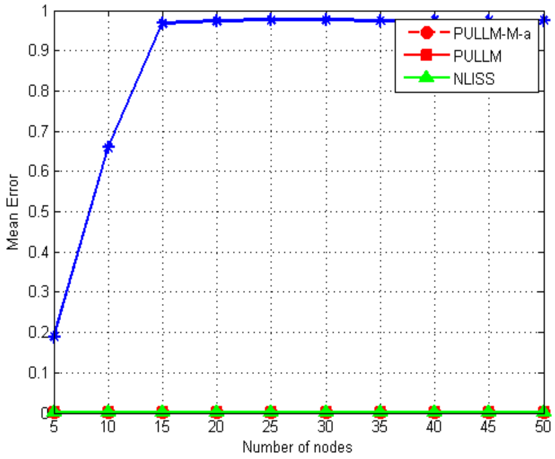


Fig.6 Mean error vs. different number of nodes

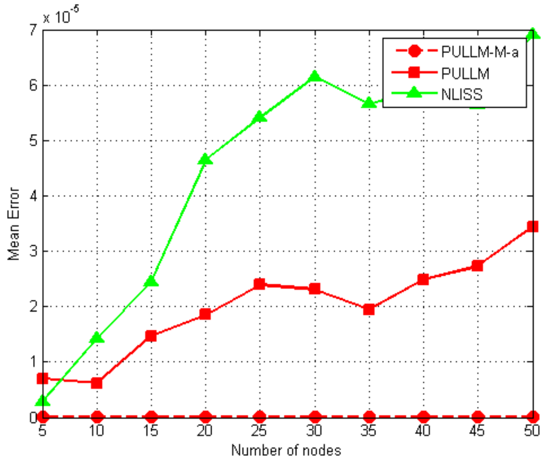


Fig.7 Mean error versus different number of nodes

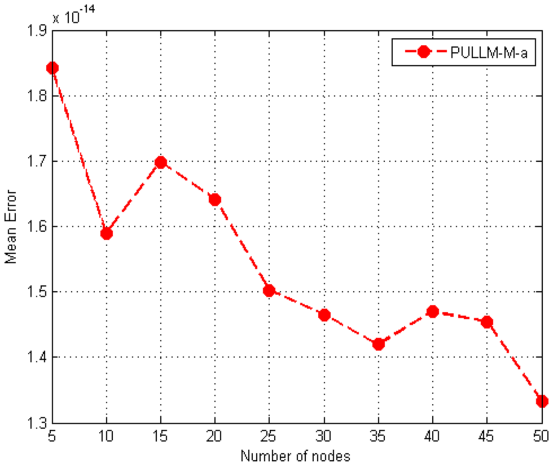


Fig.8 Mean error vs. different number of nodes

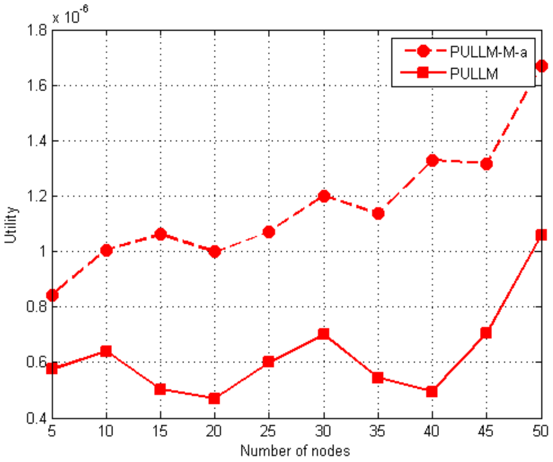


Fig.9 Utility function versus different number of nodes

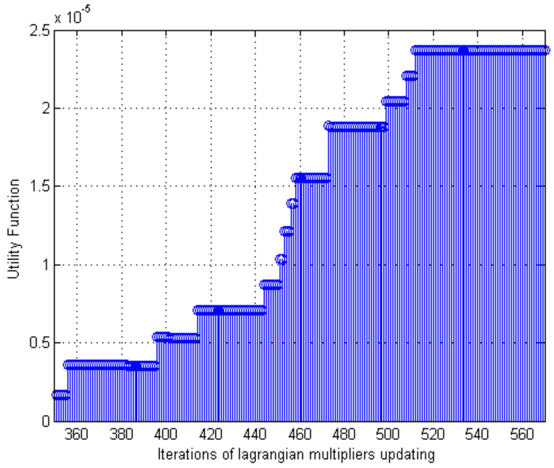


Fig.10. Utility function convergence for different iterations

6- Analysis on Results

In this paper, the purpose is the network lifetime maximization of cooperative spectrum sensing in wireless cognitive sensor networks by proper selection of the sensors for spectrum sensing and mobile primary user localization under the constraints on the false alarm and detection probabilities and accuracy of the mobile primary user localization. To save more energy of the sensors, compressive sensing (CS) framework and multi-antenna structure for each sensor are used to monitor the primary user localization.

Fig.3 and Fig.4 show the effectiveness of the proposed algorithms for improving the network lifetime. In fact, less energy consumption of the network and more remaining energy of the sensors increase the network lifetime. The proposed algorithms improve these parameters and therefore, maximize the network lifetime. Another important parameter is the successful percent of finding the solution which shows the ability of the algorithms in finding the solution. The proposed algorithms have the maximum percent in finding the solution. On the other hand, if the problem has the solution, the proposed algorithms have the most ability to find it. Fig.6, Fig.7 and Fig.8 show the mean error of the network for primary user localization. The proposed algorithms have the minimum mean error. Because these algorithms consider the accuracy of primary user localization in sensing nodes selection while RSSNL algorithm has the maximum mean error due to the random selection of the nodes for primary user localization and spectrum sensing. Fig.9 and Fig.10 show the utility function of the algorithms and changes of the utility function versus the iterations, respectively. In fact, utility function of each sensor is the ratio of its remaining energy to the energy consumption. Our purpose is to maximize the aggregate utility function. In fact, this parameter equals to the network lifetime and increasing of the aggregate utility function improves the network lifetime. According to this figure, the proposed algorithms have the maximum utility function.

7- Conclusion

Cooperative spectrum sensing has an essential role to mitigate the fading and shadowing effects in cognitive sensor networks. However, due to the limited energy budget of the sensors, the lifetime improvement should be considered in these networks.

In this work, the sensing nodes do spectrum sensing to determine the primary user activity and also track the mobile primary user location based on compressive sensing in cognitive sensor networks. This is a capability to detect the primary emulation attacks in cognitive radio networks. For saving more energy, the multi-antenna sensors is also considered. Therefore, the problem is formulated to maximize the lifetime of the network subject to the global detection performance and accuracy of the primary user localization. The problem is investigated using game theoretic solutions. Therefore, the primary user localization (PUL) is proposed as a cooperative game to solve the problem. It means that each node considers the utility function of itself as well as other sensors to improve the network lifetime and find the location of the mobile primary user with high accuracy. It is shown that the proposed algorithms maximize the network lifetime and satisfy the detection performance and accuracy of the mobile primary user localization constraints. In future, the capability of energy harvesting for sensors will be an essential issue for improving the lifetime of the cognitive sensor networks.

References

[1] R. Zhang and Y.-C. Liang, "Investigation on multiuser diversity in spectrum sharing based cognitive radio networks", IEEE Commu. Lett., Vol. 14, No. 2, 2010,pp. 133–135.

[2] R. Zhang, Y.-C. Liang, and S. Cui, "Dynamic resource allocation in cognitive radio networks", IEEE Sig. Proc. Mag., Vol. 27, No. 3,2010,pp. 102–114.

[3] A. Attar, H.Tang, A.V. Vasilakos, F.R, Yu and V.C.M Leung, " A Survey of Security Challenges in Cognitive Radio Networks: Solutions and Future Research Directions", IEEE Proc., 2012, pp. 3172–3186.

[4] J. Wang, J.Chen and D. Cabric, "Cramer-Rao Bounds for Joint RSS/DoA-Based Primary-User Localization in Cognitive Radio Networks", IEEE Trans. Wirel. Commun. Vol.12, 2013, pp.1363–1375.

[5] F.Ye, Y.Li, R.Yang and Z. Sun, "The user requirement based competitive price model for spectrum sharing in cognitive radio networks", Int. J. Distrib. Sens. Netw, 2013, doi:10.1155/2013/724581.

[6] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data", IEEE Signal Processing Magazine, Vol. 25, No. 2, 2008,pp. 92-101.

[7] Z. Tian and G. Giannakis, "Compressed sensing for wideband cognitive radios", IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, Vol. 4, pp. 1357–1360.

[8] F. Salahdine, N. Kaabouch and H.EI.Ghazi, " A survey on compressive sensing techniques for cognitive radio networks", Journal of Physical Commuincations, Vol.20, pp.61-73, 2016.

[9] Th. Wimalajeewa, P. K. Varshney, "Application of compressive sensing techniques in distributed sensor networks:A survey",Electrical Engineering and Systems Science Journal, Last revised 19 Jan 2019.

[10] Sh. Salim and S. Moh, "An Energy-Efficient Game-Theory-Based Spectrum Decision Scheme for Cognitive Radio Sensor Networks", Sensors Journal, 2016 ,pp.1-20.

[11] S.H.Hojjati, A.Ebrahimzadeh, S.M.H.Anadroli and M.Najimi, "Energy efficient cooperative spectrum sensing in wireless multi-antenna sensor networks", Springer Wireless Netw. Journal, 2017, pp. 567–578.

[12] N. Radhi, K. Aziz MIET, S. Hamad 3,H.S.AL-Raweshidy, "Estimate primary user localization using cognitive radio networks", IEEE on International Conference on Innovations in Information Technology,2011,pp.381-385.

[13] J. Werner, J. Wang, A.Hakkarainen, M. Valkama and D. Cabric, "Primary User Localization in Cognitive Radio Networks Using Sectorized Antennas", IEEE on 10th Annual Conference on Wireless On-demand Network Systems and Services (WONS),2013,pp.155-161.

[14] X. Li, V. Chakravarthy, and Z. Wu, "Joint Spectrum Sensing and Primary User Localization for Cognitive Radio via Compressed Sensing", in Proc. of IEEE MILCOM, San Jose, California, 2010.

[15] J. A. Bazerque and G. B. Giannakis, "Distributed spectrum sensing for cognitive radio networks by exploiting sparsity", IEEE Transations on Signal Processing, Vol. 58, No. 3, 2010, pp. 1847-1862.

[16] W. Guibène and D.Slock, "Cooperative spectrum sensing and localization in cognitive radio systems using compressed sensing", Journal of Sensors, dx.doi.org/10.1155/2013/606413,2013.

[17] X. Li , et al., "Spatio-temporal compressive sensing-based data gathering in wireless sensor networks", IEEE Wireless Commun. Lett.Vol.7,No. 2, pp.198-201,2018.

[18] M.-H.Yang, et.al., "Fusion-based cooperative support identification for compressive networked sensing", published in ArXiv, DOI:10.1109/lwc.2019.2946552, 2019.

[19] F. Ye, X. Zhang, Y. Li and H. Huang, "Primary User Localization Algorithm Based on Compressive Sensing in Cognitive Radio Networks", Algorithms Journal, Vo.9,No.12, 2016,pp.1-11.

[20] Y. C. Liang, Y. Zeng, E. C. Peh, and A. T. Hoang, "Sensing-throughput trade off for cognitive radio networks", IEEE Trans.Wireless Commun.Vol.7, No.4, 2008, pp.1326-1337.

[21] E. Peh and Y. Ch. Liang, "Optimization for cooperative sensing in cognitive radio networks", IEEE Commun. Society, WCNC, 2007, pp.27-32.

[22] B. Zhang, "Sparse target counting and localization in sensor networks based on compressive sensing", IEEE Globecom, 2009.

[23] S. Maleki, A. Pandharipande, and G. Leus, "Energy-efficient distributed spectrum sensing for cognitive sensor networks", in Proc. 35th Annu. Conf. IEEE Ind. Electron. Soc., 2009, pp. 2642–2646.

[24] M. J. Osborne and A. Rubinstein, A Course in Game Theory. Cambridge, MA, USA: MIT Press, 1994.

[25] Quan Z., Cui S., Sayed A.H. "linear cooperation for spectrum sensing in cognitive radio networks", IEEE Journal of Selected Topics in Signal Processing,,Vol.2,No.1,2008, pp.28–39

[26] M.Najimi, A. Ebrahimzadeh, S.M.Hosseni Andargoli, A. Fallahi, "A novel sensing nodes and decision node selection method for energy efficiency of cooperative spectrum sensing in cognitive sensor networks", IEEE Sensors Journal, Vol.13,No.5, 2013, pp.1610-1621.

Maryam Najimi received her B.Sc in electronics from Sistan & Baloochestan University, Zahedan, Iran in 2004 and her M.Sc in telecommunication systems engineering from K.N.Toosi University of Technology, Tehran, Iran and Ph.D. degree in communication from Babol University of Technology, Mazandaran, Iran, in 2008 and 2014, respectively. She is currently an assistant professor with department of electrical and computer engineering, University of Science and Technology of Mazandaran, Behshahr, Iran. Her interests include Spectrum sensing in wireless cognitive sensor networks.

An Intelligent Autonomous System for Condition-Based Maintenance- Case Study: Control Valves

Hamidreza Naseri
Department of Computer, Faculty of Engineering, Islamic Azad University, Qom Branch, Qom, Iran
H.Naseri@qom.iau.ac.ir
Ali Shahidinejad*
Department of Computer, Faculty of Engineering, Islamic Azad University, Qom Branch, Qom, Iran
A.shahidinejad@qom-iau.ac.ir
Mostafa Ghobaei-Arani
Department of Computer, Faculty of Engineering, Islamic Azad University, Qom Branch, Qom, Iran
M.ghobaei@qom-iau.ac.ir

Received: 29/Jul/2019 Revised: 19/Nov/2019 Accepted: 29/Dec/2019

Abstract

Maintenance process generally plays a vital role to achieve more benefits to the enterprises. Undoubtedly, this process has a high value-added in oil and gas industries. Process owner expectations and new technology acquisition have been changing the mindset of domain experts to the new maintenance approaches and different newer methods such as condition-based maintenance models for improving the reliability and decreasing the cost of maintenance. Because of the high dynamic behavior of the gas and the instability of the input parameters, the need to apply a model with self-healing behavior is a serious demand in the gas industry. However, to the best of our knowledge, despite its importance, there is not any comprehensive study in the literature. In this paper, we present a new neuro-fuzzy model and a self-management control loop using real world data to meet the mentioned targets for a specified control valve in a gas refinery. ANFIS model is employed for the reasoning process which has six inputs (Inlet/outlet Pressures, temperature, flow rate, controller output and valve rod displacement), and one output that is a type of failure of the control valve and the most failures are considered based on domain expert knowledge. A suitable control loop is used to unceasingly monitor, analyze, plan and finally execute the process of prediction of failures. Due to undertaken improvement, there is a considerable change in reliability and financial indices. Moreover, the proposed approach is compared with two different methods. The results show that our proposed model comprehensively improves accuracy by 24%.

Keywords: Condition-Based Maintenance, Neuro-fuzzy, Autonomic computing, control valve.

1- Introduction

Maintenance is an aggregation of technical and related administrative actions intended to maintain an item/system in, or to restore it to a stable functioning state [1]. There are generally four main maintenance approaches: corrective maintenance (CM), preventive maintenance (PM), and condition-based maintenance (CBM) and the hybrid approach [2, 15]. In the CM approach, known as "run to failure" strategy, a device is allowed to fail. Then, the repair process is performed. This approach is suitable when the consequences and impacts of failures are small. In the PM approach, maintenance is scheduled in advance to prevent failures. It focuses on avoiding failures through replacing components at a particular time [2]. Then, in the CBM approach, the decision is made depending on the measured data. Based on data analysis, whenever monitoring level value exceeds a standard amount, a component is either replaced or repaired. Applying this strategy may lead to a considerable reduction in production cost. The optimization of the maintenance strategy is a trade-off between the cost of planned (PM/CBM) and unplanned (CM) maintenance interventions [3]. Finally, there is also a hybrid approach where CBM incorporated into one component with preventive maintenance (PM) and corrective maintenance (CM) on the other components of the same machine.

As a different stochastic model, Markov decision process and Bayesian networks are mostly used in maintenance prediction. One practical use is to employ a Bayesian network with conjugate priors to provide exact expressions for the remaining useful lifetime and obtain a set of expected operational cycle cost rate functions [18]. In recent years neuro-fuzzy vastly used in CBM since the algorithms are an assimilation of neural networks and FIS and can override the disadvantages of them. Neuro-fuzzy logic is easy and rapid to apply and particularly adaptive, lucid and highly flexible [19].

*Corresponding Author

Because of the high dynamic behavior of gas flowing in the equipment and the instability of input parameters of these equipment, applying a model with self-management behavior to meet the predefined targets is seriously demanding in the gas industry. However, to the best of our knowledge, despite its importance, none of the research studies cover the literature comprehensively, professionally, and precisely for a specified equipment.

Control valves are an integral device in oil and gas refineries. Also, they are one of the most overlooked devices in terms of maintenance. However, if not provided with proper maintenance, these units fail to operate to their maximum efficiency, which in turn leads to several performance issues. Generally, the most industries use corrective approach where a routine maintenance practice is conducted and scheduled maintenance program is considered. In this case, since the time of maintenance is the only criteria for the repair, so some equipment may be repaired although they don't require it. However, this approach usually leads to cost of downtime and maintenance. As a solution, condition based maintenance is used in this research to overcome to this problem. Here, maintenance of the control valves is rendered on the basis of the evaluation results of monitoring and testing equipment. Immediate maintenance or repairing action is taken in case of any discrepancies.

In this research, a new CBM model is designed and implemented for detecting and preventing some often failures in control valves using real data in a gas refinery and then extracting the frequent failures of the control valve, investigate the valve behavior. The data is collected from CMMS (Computerized Maintenance Management System) software running at the refinery for six months. Afterward, using the data and expert knowledge, we design and implement a self-management predictive maintenance system for the control valve. The proposed solution is based on the MAPE-K control loop to automatically analyze the behavior of a control valve and detect predefined failures in earlier condition before the defect causes complete disruption.

The main contributions of this paper are described as follows:

-Designing a self-management solution for predictive condition-based maintenance based on the MAPE-K control loop for the specified control valve.

-Utilizing the neuro-fuzzy technique as a decision maker in the MAPE-K control loop.

The rest of the paper is organized as follows: In section 2, related works and relevant literature is reviewed. Then there are some notations and definitions in section 3. The proposed framework and algorithm is described in detail in section 4. In Section 5, our solution is evaluated by comparing it with two other methods. Finally, in Section 6 and 7 we discuss all the results and conclude the paper.

2- Related Works And Literature Review

Several scientists have employed neuro-fuzzy to carry out a variety of tasks such as prediction and optimization and several models have been placed regarding various applications [4]. Although many investigators have extensively studied predictive maintenance in recent years, however, there are few studies about the control valves failures in gas refineries. In the rest of this section, the most important and relevant researches are reviewed in this field. Then, the most relevant studies have been assessed from various points of view [5-10].

Keizer et al. [8] prepared a broad literature overview of CBM policies for multi-component systems and occurrence of the dependencies. Friedrich et al. [5] presented an introduction into the possibilities to automate the maintenance process. Furthermore, another surveys [6, 7] also reviewed recent papers in successfulness of the CBM in manufacturing companies using vibration measurement and signal processing. Carnero [10] introduced the results of a survey of 35 Small and Medium Enterprises (SME's) in a region of Spain. Though some companies didn't have CBM, so this study assessed the level of the predictive program and where there was CBM, it analyzed its characteristics. They compared the results of this survey with the best practice as set out in the current literature and resulted that assessed SME's had very different results in almost all indicators used. Ahmad et al. [9] presented an overview of two maintenance techniques to determine how the TBM and CBM work toward maintenance decision making. Zhu et al. [11] introduced a new model for a single CBM component which gave opportunities for preventive maintenance. But for the deterioration process, they used both a random coefficient model and a Gamma process and developed an accurate approximate evaluation procedure for control limit policies. The key idea behind this approximation was that they pretended that the time points, at which preventive or corrective maintenance is executed, constitute renewal points as defined within renewal theory. Furthermore, Keizer et al. [12] proposed a multi-component, parallel condition-based maintenance (CBM) system with economic dependence and load sharing. They concluded that the redundancy resulting from the identical setting permits corrective replacements to be postponed without a system performance.

In a different approach, Do et al. [13] implemented a proactive model for perfect and imperfect actions on condition-based maintenance to check and consider the impacts of imperfect maintenance actions and to improve an adaptive maintenance policy, which can help to select optimally suitable actions at each inspection period. Liu et al. [14] incorporated side effect of the Wiener degradation process with linear drift into a condition-based maintenance policies with age- and state-dependent operating cost, which can be applied in various real systems to gain economic and societal benefits. Guariente et al. [16] studied the implementation of the autonomous maintenance in a company which supplies air conditioning tubes to the automotive section using Lean philosophy and by deploying a seven stage process. Furthermore, Lewandowski et al. [17] studied the way of handling the complex situation for the operational maintenance processes using an autonomous system as well as related spare part logistics in general and illustrated first concrete approach.

Table 1 summarizes some of the most relevant works related to the condition-based maintenance approaches based on applied techniques, performance metrics, case study, and reactive/proactive or autonomous policy.

Table 1: Statistics of the most relevant works related to CBM approaches

Ref.	***Applied Technique(s)***	***Performance Metrics***	***Case study***	***Reactive/ Proactive/ Autonomous***	***Single/Multiple components***	***Published Year***
[11]	Random coefficient mode Gamma process	Cost, Average Absolute Difference (AAD), Maximum Absolute Difference (MAD)	Lithography machines	Proactive	Single	2017
[15]	Threshold-based	Maintenance cost	Compressed air, generator and pumps	Hybrid	Multiple	2018
[13]	Gamma stochastic process	Maintenance cost, unavailability cost rate, inspection cost	Logistic industry	Proactive	Single	2015
[19]	Neuro-fuzzy system	Response surfaces	Aircraft engine	Autonomous	Single	2007
[18]	Bayesian Survival signature	System reliability, the unit cost rate	Manufacturing companies	Proactive	Multiple	2017
[12]	Markov Decision Process	optimal policy structure	Gas company	Proactive	Multiple	2018
[16]	Autonomous Maintenance Lean philosophy tools	Equipment Availability Overall Equipment Effectiveness	Automotive Component Manufacturer	Autonomous	Multiple	2017
[20]	Parameter Identification	Determine the remaining the lifetime of a component.	Industrial Machines	Proactive	Multiple	2016
[14]	Wiener process with linear drift	Optimum maintenance decisions	N/A	Proactive	Multiple	2017
[17]	CBM	Quality Enhancement	N/A	Autonomous	Multiple	2014
Our Proposed research	ANIFIS/ MAPE-K	Reliability and Financial	Control valve in gas industry	Autonomous	Multiple	------

Most of these studies mainly attempt to provide a model for CBM, employing techniques such as Gamma process [11] or Bayesian [18] which have proactive approach to the problem. In a different approach, some papers use neural network [16], [19] and [20], but they propose a general model to the CBM in industrial equipment.

In this study, a new model is proposed for a specified device so it can detect any abnormal condition of a control valve and therefore preventing it from any breakdown. Due to instability of gas entity and variety of input parameters, MAPE-K control loop is applied which according to obtained results we experience a considerable improvement.

3- Problem Formulation

The main goal in this research is identifying and analyzing the behavior of control valves in the gas industry so that it will be possible to detect failures in the

initial state. It should be noted that in all circumstances it has been assumed that the system is initially in a normal situation, then a failure occurs within an interval of time, and eventually the failure is detected and fixed. This sequence of generating and fixing the failure improves and enriches failure detection signal and also training of neuro-fuzzy network. In this section, equations and notations are described as shown in Table 2.
There are six parameters as inputs to the system, which are as follows:

- Inlet pressure: This is the pressure at valve input in the bar scale.
- Outlet pressure: This is the pressure at valve output in the bar scale.
- Controller output: This is a mini voltage from the controller.
- Flow rate: The amount of gas which passes through the control valve in a second.
- Temperature: that is the gas temperature in centigrade degree.
- Rod Displacement: this is the amount of valve stem movement.

Table 2: Notations and definitions

Notation	***Definition***
P_i	Valve Inlet pressure
P_o	Valve Outlet pressure
C	Valve Controller output
T	Gas Temperature
μ	Membership Function for the input value
Wi	Ignition Strength of the rules
$\overline{w_\imath}$	Normalized Firing strength
pi, qi, ri	Consequent Parameters
F_T	Failure Type as Output
S	The standard deviation of data
TA_T	Vector of Initial Learning Rates to be tested
R_T	The vector of rules to be tested
η	number of tests
E_P	number of epochs
Ai, Bi	Fuzzy sets
Y_i	Input value

4- Proposed Framework And Algorithms

In this section, our proposed framework and algorithms are explained in more detail. The proposed autonomous framework for condition-based maintenance makes use of ANFIS. The conceptual framework shows which components our system has and how they relate to each other, as shown in Figure 1.

4-1- Autonomous Systems

Autonomous systems act as a self-management component to detect errors preventively and make the appropriate decision to repair automatically. As the mastermind, MAPE-K control cycle make an autonomous system that consists of four components, as it is depicted in Figure 1; Monitoring, Analyzing, Planning, and executing Phases, which are described briefly in the following:
- **Monitor Phase**: In this phase, data that are produced by the sensor, meters, and gauges are gathered and filtered.
- **Analysis phase**: This phase acts as a signal modulating and transmitting component of the information obtained from the monitor phase.
- **Planning phase**: As the brain of the autonomous system, in this phase ANFIS is applied to utilize and make the required decisions.
- **Execution phase**: In this phase, the decisions made in the planning phase are performed.

All the above mentioned phases store their results in the Knowledge to make further use in the future by the other components.

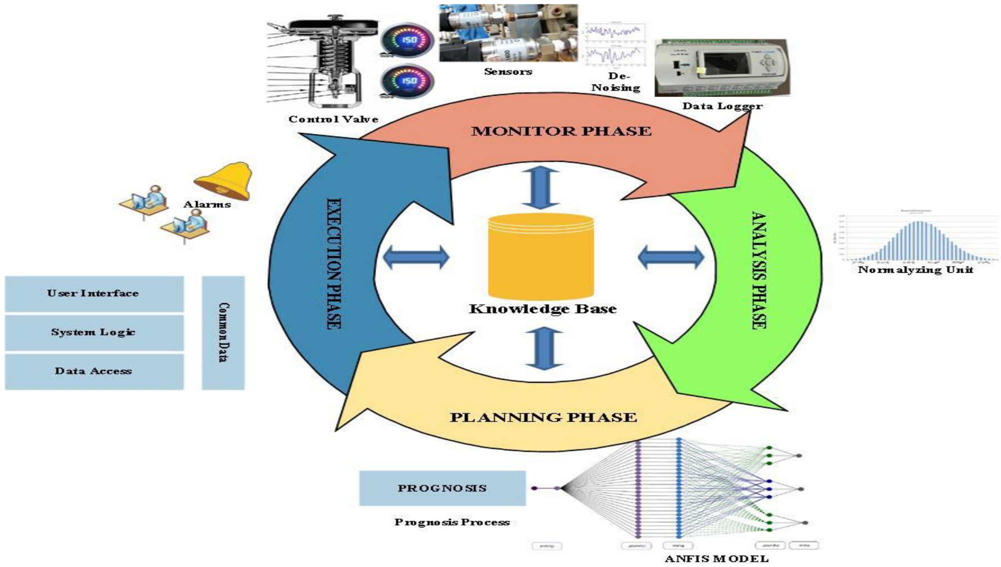


Fig. 1 The proposed autonomous framework for condition-based maintenance

4-2- Proposed Algorithms

In this subsection, the algorithms that are designed to predict, detect, and display predefined failure types are described. As mentioned earlier, the model uses the MAPE-K as control loop and ANFIS as a prediction tool to increase accuracy. In this research, it is assumed that the system (control valve) is initially in normal condition and then a failure occurs within a time interval. Therefore, at each time interval Δt, monitoring, analysis, planning, and execution are continuously repeated for existing data (algorithm 1- line 2-4).

In the monitor phase, sensors for temperature, pressure, flow, controller output voltage and displacement of the stem are used to receive and collect raw data. Then, these data will be stored in knowledge base engine that is a SQL database in this study. This data will be used as input to the next phase (line 6).

Afterward, in the analyzer component, aggregated raw data will be normalized (line 7). In the planning, which is the core phase, the prediction of the failure is done. In this step, using ANFIS model and training data, failure prediction is done (line 8). In the execution phase, in addition to setting up new basic data, the main outputs are displayed as the failure type (line 9).

Algorithm 1 shows the pseudo code for the four main phases of the MAPE-K for each time interval Δt.

Algorithm 1. Pseudo code for Autonomous predictive maintenance

```
1: Initialization
2: while (the system is running and in the beginning of interval Δt) do
3: begin
4:     for (every gas particle in control valve flow rate Fv at interval Δt) do
5:     begin
6:         Monitor (T, $P_i$ , $P_o$, C, F, X at interval Δt)
7:         Analyze (Input data as to be normal at interval Δt)
8:         Plan (Inference using ANFIS as prediction of failures at interval Δt)
9:         Execute (software components to retreive data and alert failure types at interval Δt)
10:    end for
11: end while
```

A- Monitor Phase

This component is responsible to collect and aggregate all the necessary data as the system inputs. In this research, controller output voltage, temperature, Inlet and outlet pressures, flow rate and rod displacement of a typical control valve are considered as the system inputs. A data logger has to be employed (Figure 2- a) for data gathering and homogenizing inputs of the system. Figure 2-b, c, d show a real picture of the control valve associated with all sensors and cables after installation. The data logger uses XML format (although it can store data in some other forms e.g. JSON). It has Analogue and digital ports as input/output ports and then can transfer streams via a RS-485 serial port using Modbus protocol. An embedded Ethernet port may also be used for FTP/SNMP/SMTP protocols.

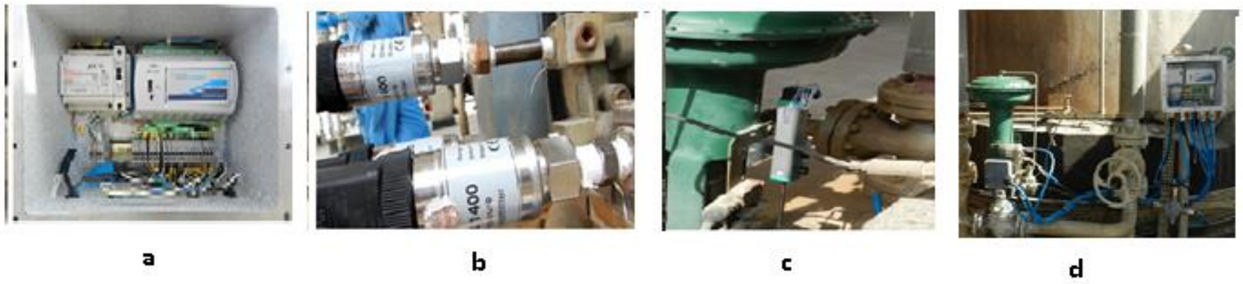


Fig. 2 a: Data logger b: Pressure/Flow transducer c: Rod displacement d: real control valve with all associated sensors

Generally, in this phase, the input parameters are aggregated and stored. To this purpose, some equipment such as pressure and temperature transmitters as well as all installed sensors and resistances are used, so that changes can be made to the valve without affecting the valve stem. These instruments are ultimately connected to the data logger observing the safety cautions and using the armored cables to the anti-explosion box. Afterwards, set up a SQL server to manage to store and to retrieve data and expert knowledge. Algorithm 2 shows the pseudo code for the monitoring phase of MAPE-K for each time interval.

Algorithm 2. Pseudo code for Monitoring Phase

```
1: Begin
2: Read (Controller output C in interval Δt); /*controller output voltage Monitoring*/
3: Read (Temperature T in interval Δt); /*Temperature Monitoring*/
4: Read (Inlet Pressure $P_i$ in interval Δt); /*Inlet Pressure Monitoring*/
5: Read (Outlet Pressure $P_o$ in interval Δt); /* OutletPressure Monitoring*/
6:Read (Flow rate F in interval Δt); /* Flow rate Monitoring*/
7:Read (Rod Displacement X in interval Δt); /*Rod Displacement Monitoring*/
8: store (gathered data); /* storing in SQL database*/
9: return gathered value for (C, T, $P_i$ , $P_o$, F, X)
10: End
```

B- Analysis Phase

This phase consists of one subcomponent, namely the normalizing unit. There are four transmitters (e.g. two Pressures, temperature and flow) to modulate and transmit data signal in the form of electrical signal. In data logger then all data will be demodulated into a digital signal using. To do this, the device is configured so it can gather input parameters at desired time intervals.

Among the various normalizing methods, the model uses the Standardization method, which was first introduced by Saati to scale up the data in artificial neural networks, using equation (1)[25].

$$Normalized\ variable = \frac{x-mean}{S} \qquad (1)$$

In this equation, S is the standard deviation, and the variable X can be any of the input parameters (temperature, pressure, flow, and displacement rate of the stem).

Due to the existence of different scales for the received parameters (temperature: centigrade - pressure: Bar - flow: cubic meter per second and valve stem displacement: mm), it must be normalized (line 3).

Algorithm 3 shows pseudo code for this phase.

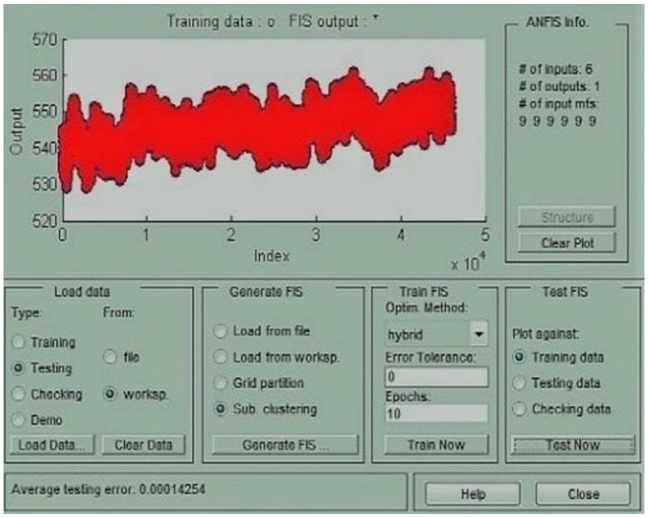
Algorithm 3. Pseudo code for Analysis Phase

```
1: Begin
2: Normalize data (C, T, $P_i$ , $P_o$, F, X) /* using Eq. 2*/
3:store Normalized value /*using SQl database*/
4: return Normalized value
```

C- Planning Phase

The output of the previous phase is considered as the input of this phase. In this study, the ANFIS is employed to train the network, generalize the results to different states, and predict the failures. ANFIS as a hybrid method uses a parallel-distributed processing model and learning ability feature of ANN. ANFIS comprises of mainly two parts. The first one is the antecedent and the second part is the conclusion section, which is connected by rules [21].

ANFIS network is the feed-forward type which makes this adaptive network employ in a wide variety of applications of modeling, decision-making, and control [22]. For the learning process, a dataset is prepared which has been gathered from a gas refinery for six months. A total number of 57450 data records were considered in which 80% of the data were used for training and the remaining was used for testing (line 4 to 5). Figure 3 presents the train data in our ANFIS model.



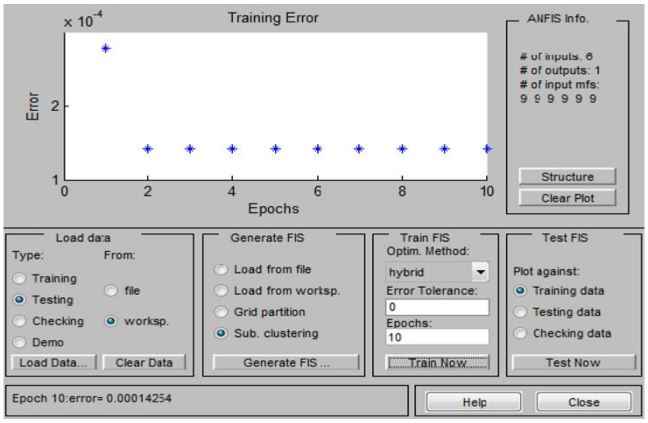
ANFIS Info.
# of inputs: 6
# of outputs: 1
# of input mfs:
9 9 9 9 9 9

Load data — Type: ○ Training ● Testing ○ Checking ○ Demo; From: ○ file ● worksp.
Generate FIS — ○ Load from file ○ Load from worksp. ○ Grid partition ● Sub. clustering
Train FIS — Optim. Method: hybrid; Error Tolerance: 0; Epochs: 10
Test FIS — Plot against: ● Training data ○ Testing data ○ Checking data

Average testing error: 0.00014254

Fig. 3 Training data for the ANFIS model

FIS is generated by taking advantage of the sub-clustering method that is suitable for applications with a small number of inputs variables (less than 8). Training process fulfilled in 10 epochs. Error tolerance was kept zero for this process. After running the system, the value of 0.0014254 was obtained for the MSE variable, which was an acceptable value (see Figure 4)**.** Algorithm 4 shows pseudo-code for this step.



ANFIS Info.
# of inputs: 6
# of outputs: 1
# of input mfs:
9 9 9 9 9 9

Load data — Type: ○ Training ● Testing ○ Checking ○ Demo; From: ○ file ● worksp.
Generate FIS — ○ Load from file ○ Load from worksp. ○ Grid partition ● Sub. clustering
Train FIS — Optim. Method: hybrid; Error Tolerance: 0; Epochs: 10
Test FIS — Plot against: ● Training data ○ Testing data ○ Checking data

Epoch 10:error= 0.00014254

Fig. 4 System running and MSE obtained for ANFIS

Algorithm 4. Pseudo code for the Planning Phase

```
Begin
Input: aggregated normalized data
Output: Failure type
Training = 80% of Base
Testing = 20% of Base
Vector of rules to be tested $R_T$
Vector of Initial Learning Rates to be tested $TA_T$
$\eta$ = number of tests
$E_P$ = number of epochs
for i ϵ $TA_T$ do
    for j ϵ $R_T$ do
        for l = 1 to $\eta$ do
            fuzzify inputs using trangular membership function by Eq. 2 /*Layer 1*/
            check weights of each membership function (T-norm) by Eq. 3 /*Layer 2*/
            perform pre-condition matching of fuzzy rules by Eq. 4 /*Layer 3*/
            inference of rules and produce normalized firing rule strength by Eq. 5 /*Layer 4*/
            sum up all the weights and defuzzification by Eq. 6 /*Layer 5*/
            save the FIS with lowest validation error, the training error, output vector validation
        end for
    end for
end for
End
```

D- Execution Phase

This phase deals with a software component that can predict the occurrence of some pre-defined defects based on artificial intelligence and then alerts the user for suitable feedback. Algorithm 5 shows the Pseudo code for this phase [23-25].

Algorithm 5. Pseudo code for the execution Phase

```
Begin
Input: Failure type
Output: Approprate reports regarding deduction and domain knowledge
fetch domain knowledge and rules and Common data /*from ANFIS model / Database/ common data */
show failure type as report or list
```

5- Performance Evaluation

In this section, the effectiveness of our model is evaluated. In addition, in this study, MSE is used to determine accuracy and then as performance indices. Furthermore, some reliability and cost (financial) metrics are used that are explained at the end of this section.

Appraisal of the model is fulfilled by simulating the framework with two different methods (fuzzy SUGENO and neural network). To this purpose, at first the performance metrics are introduced which employed and then simulate the system with two various methods. Noteworthy, these experiments were all carried out on the same database though they may have had different training and test sets.

5-1- Performance Metrics

In this study two different approaches are employed for evaluation purposes. The MSE is used as a leading indicator, and MTBF, MTTR, and FR as a lagging indicator and then describe their short definition in the following.

- Mean Square Error (MSE):

This indicator measures the average of the squares of the errors that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss.

- MTBF

This is a reliability term used to provide the number of failures per hours for equipment and is the most common inquiry about equipment's life span, and is important in the decision-making process of the end users as shown in the equation (2)[25].

$$MTBF = \frac{T}{R} \qquad (2)$$

where T is total operation time, and R is the number of failures.

- MTTR

MTTR is the time needed to repair failed equipment. In an operational system, repair generally means replacing a failed part. Equation 3 may calculate this indicator [25].

$$MTTR = \frac{Total\ Maintenance\ Time}{Number\ of\ repairs} \qquad (3)$$

-

- FR

Failure rate (FR) is another way of reporting MTBF, and it indicates the number of expected failures per (one million) hours of operation for a device.

- Maintenance Cost

Maintenance expenses are the costs incurred to keep an item in good condition or good working order. This includes maintenance materials, contract maintenance labor, and equipment rental.

- Downtime Cost

Downtime cost refers to the economic loss arising from downtime. The downtime attributes the financial damage to equipment maintenance or replacement at the refineries.

Figure 5 shows the results of the simulation done in the neural network. As it can be seen MSE is 0.026 that was measured by this method.

5-2- Experimental Analysis

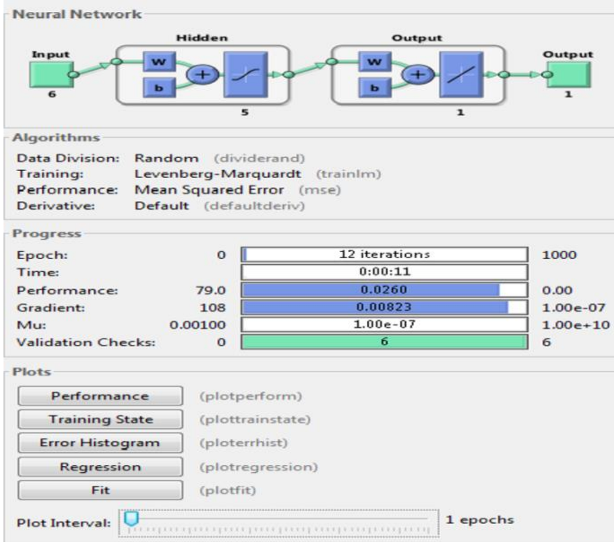
In this section the effectiveness of the proposed framework is compared with two other approaches. i) At first, calculation of MSE is done for evaluating the prediction accuracy of the planner component by simulating the model with two different methods (fuzzy SUGENO and neural network). ii) Then calculate some reliability and finance metrics after implementation. Here we explain how to simulate the model, and afterward, describe performance indices.

- Neural Network Simulation

Using neural network fitting tool, 80% of data is input as training data, 10% for test and the remaining 10% for validation data. Table 3 shows the set parameters in the MATLAB environment.

As described early, neural networks are suitable candidates to fulfill CBM objectives. These algorithms have been validated on various data sets and were shown to possess good accuracy.

After simulating by this method, calculation of MSE equals to 0.0015, which shows a better value than neural network simulation.



Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Epoch: 0 | 12 iterations | 1000
Time: 0:00:11
Performance: 79.0 | 0.0260 | 0.00
Gradient: 108 | 0.00823 | 1.00e-07
Mu: 0.00100 | 1.00e-07 | 1.00e+10
Validation Checks: 0 | 6 | 6

(plotperform)
(plottrainstate)
(ploterrhist)
(plotregression)
(plotfit)

Plot Interval: 1 epochs

Fig. 5 Simulation result for the NN method

Table 3: Separation of data in ANFIS toolbox

Training Data	80%	45958 Samples
Validation Data	10%	5745 Samples
Test Data	10%	5745 Samples

- Fuzzy SUGENO simulation

As a second method for evaluating the model, fuzzy SUGENO is used. Sugeno is efficient for diagnosis applications, and the knowledge (rules) it extracts would be abstract for a domain expert as they are in a linguistic format. However, Sugeno uses a defuzzification strategy that limits the membership functions to obtain a Gaussian functional form. Table 4 denotes the set parameters for the simulation.

Table 4: Set parameters in Sugeno

Type	Sugeno
orMethod	prod
DE fuzzy Method	wtaver
impMethod	Prod
aggMethod	Sum
Input	6
Output	1
Training Data	80%
Validation Data	10%
Test Data	10%

6- Results

As it can be seen earlier in the proposed algorithm section, this error was 0.0014254 for MSE that shows the best performance among the other simulated methods. Table 5 shows the MSE value for the three methods.

Table 5**:** MSE value for different methods

Method	***MSE Value***
Neural Network	0.0260
Fuzzy SUGENO	0.0015
ANFIS	0.0014254

In a different approach and as lagging indicators, we validated our model by reliability and financial indicators. Reliability indices are those to monitor performance and efficiency. By implementing the proposed model, an improvement in all indices was observed. As it can be seen in Table 6, the overall outcome of implementing the proposed model results in an increase in MTBF and decrease in all other criteria that in refineries may have a significant benefit.

Table 6: Improvement rate in indices

Category	***Indices***	***Before Implementation***	***After Implementation***	***Improvement% (APP.)***
Reliabil ity	MTBF(h)	11520	13090	14%
	MTTR(h)	282	218	22%
	FR	0.0505	0.0229	54%
Financial	Maintenance Cost(USD)	350	280	20%
	Downtime Cost(USD)	900	700	22%

It should be considered that this value is calculated for the specified control valve at the same time in the pilot plan.

7- Conclusion

In this paper, an intelligent autonomous system for condition-based maintenance is introduced which employed on control vales in a gas refinery. This model benefited from the real dataset as input gathered from various sensors (for more than six months) to run ANFIS as an analyzer, planner and decision maker and also MAPE-K as a control loop to present and implement a new model for condition-based maintenance in a gas refinery. Furthermore, this model utilizes expert knowledge to tune up the rules. This model is successfully implemented in a gas refinery as a pilot plan, and therefore it can be extended to other units and even other industries. There are some limitations to our research, many related to the sensitivity of the gas industry. Although this study is limited in specific equipment, it may be addressed by other researches in the future.

The proposed approach was evaluated using the real trending data of a control valve in the gas refinery. We concluded that our proposed model has many advantages compared to other policy since real monitoring of equipment’s condition helps to cope better with the uncertainty. There are two groups of indices for evaluation in this study. Results show that there is a considerable improvement in all indices (from 14 to 54%).

References

[1]"ISO 14224, Petroleum, petrochemical and natural gas industries -- Collection and exchange of reliability and maintenance data for equipment." https://www.iso.org/standard/64076.html.

[2] P. Agarwal, M. Sahai, V. Mishra, M. Bag, and V. Singh, "A review of multi-criteria decision making techniques for supplier evaluation and selection," *International journal of industrial engineering computations,* vol. 2, no. 4, pp. 801-810, 2011.

[3] A. Garg and S. Deshmxukh, "Maintenance management: literature review and directions," *Journal of quality in maintenance engineering,* vol. 12, no. 3, pp. 205-238, 2006.

[4] M. Jain and K. Pathak, "Applications of artificial neural network in construction engineering and management-a review," International Journal of Engineering Technology, Management and Applied Sciences, vol. 2, no. 3, pp. 134-142, 2014.

[5] C. Friedrich, A. Lechler, and A. Verl, "Autonomous Systems for Maintenance Tasks–Requirements and Design of a Control Architecture," Procedia Technology, vol. 15, pp. 595-604, 2014.

[6] D. Goyal and B. Pabla, "Condition based maintenance of machine tools—A review," CIRP Journal of Manufacturing Science and Technology, vol. 10, pp. 24-35, 2015.

[7]A. Mérigaud and J. V. Ringwood, "Condition-based maintenance methods for marine renewable energy," Renewable and Sustainable Energy Reviews, vol. 66, pp. 53-78, 2016.

[8] M. C. O. Keizer, S. D. P. Flapper, and R. H. Teunter, "Condition-based maintenance policies for systems with multiple dependent components: A review," European Journal of Operational Research, vol. 261, no. 2, pp. 405-420, 2017.

[9] R. Ahmad and S. Kamaruddin, "An overview of time-based and condition-based maintenance in industrial application," Computers & Industrial Engineering, vol. 63, no. 1, pp. 135-149, 2012.

[10] M. C. Carnero, "Condition Based Maintenance in small industries," IFAC Proceedings Volumes, vol. 45, no. 31, pp. 199-204, 2012.

[11] Q. Zhu, H. Peng, B. Timmermans, and G.-J. van Houtum, "A condition-based maintenance model for a single component in a system with scheduled and unscheduled downs," International Journal of Production Economics, vol. 193, pp. 365-380, 2017.

[12] M. C. O. Keizer, R. H. Teunter, J. Veldman, and M. Z. Babai, "Condition-based maintenance for systems with economic dependence and load sharing," International Journal of Production Economics, vol. 195, pp. 319-327, 2018.

[13] P. Do, A. Voisin, E. Levrat, and B. Iung, "A proactive condition-based maintenance strategy with both perfect and imperfect maintenance actions," Reliability Engineering & System Safety, vol. 133, pp. 22-32, 2015.

[14] B. Liu, S. Wu, M. Xie, and W. Kuo, "A condition-based maintenance policy for degrading systems with age-and state-dependent operating cost," European Journal of Operational Research, vol. 263, no. 3, pp. 879-887, 2017.

[15] J. Poppe, R. N. Boute, and M. R. Lambrecht, "A hybrid condition-based maintenance policy for continuously monitored components with two degradation thresholds," European Journal of Operational Research, vol. 268, no. 2, pp. 515-532, 2018.

[16] P. Guariente, I. Antoniolli, L. P. Ferreira, T. Pereira, and F. Silva, "Implementing autonomous maintenance in an automotive components manufacturer," Procedia Manufacturing, vol. 13, pp. 1128-1134, 2017.

[17] M. Lewandowski and S. Oelker, "Towards autonomous control in maintenance and spare part logistics–challenges and opportunities for preacting maintenance concepts," Procedia Technology, vol. 15, pp. 333-340, 2014.

[18] G. Walter and S. D. Flapper, "Condition-based maintenance for complex systems based on current component status and Bayesian updating of component reliability," Reliability Engineering & System Safety, vol. 168, pp. 227-239, 2017.

[19] R. Kothamasu and S. H. Huang, "Adaptive Mamdani fuzzy model for condition-based maintenance," Fuzzy Sets and Systems, vol. 158, no. 24, pp. 2715-2733, 2007.

[20] M. Engeler, D. Treyer, D. Zogg, K. Wegener, and A. Kunz, "Condition-based Maintenance: Model vs. Statistics a Performance Comparison," Procedia CIRP, vol. 57, pp. 253-258, 2016.

[21] M. Ghobaei-Arani, R. Khorsand and M. Ramezanpour, "An autonomous resource provisioning framework for massively multiplayer online games in cloud environment", Journal of Network and Computer Applications, Volume 142, pp. 76-97, 2019.

[22] M. Aslanpour, S. Dashti, M. Ghobaei-Arani and A. Rahmanian, "Resource provisioning for cloud applications: a 3-D, provident and flexible approach", The Journal of Supercomputing, vol. 74, no. 12, pp. 6470-6501, 2018.

[23] M. Ghobaei-Arani, A. Rahmanian, M. Shamsi and A. Rasouli-Kenari, "A learning-based approach for virtual machine placement in cloud data centers", International Journal of Communication Systems, vol. 31, no. 8, p. e3537, 2018.

[24] A. Sharifi, A. Vosolipour, M. A. Sh, and M. Teshnehlab, "Hierarchical Takagi-Sugeno type fuzzy system for diabetes mellitus forecasting," in *Machine Learning and Cybernetics, 2008 International Conference on,* 2008, vol. 3: IEEE, pp. 1265-1270.

[25] N. Walia, H. Singh, and A. Sharma, "ANFIS: Adaptive neuro-fuzzy inference system-a survey," International Journal of Computer Applications, vol. 123, no. 13, 2015.

Hamidreza Naseri received the B.S degree in 2003 and M.S degree in Computer/ Information technology administration from Shahid Beheshti University, Tehran, Iran in 2011. He is now PhD candidate in information technology management in Azad University, Qom Branch, Qom, Iran. He has more than 19 years experience in design and implementation of ICT frameworks and standards, network administration, Software development and service management.
His current research interests are Software development, intelligent systems, IT frameworks and standards (ISMS, COBIT, DevOps, ITIL, EFQM) and also SDN. (https://orcid.org/0000-0002-4790-3368).

Ali Shahidinejad* received the B.S. degree in computer hardware engineering from Islamic Azad University of Kashan, Iran in 2008, the M.S. degree in computer architecture from Islamic Azad University of Arak, Iran, in 2010 and Ph.D. degree in Computer Networks at the University of Technology of Aachen University, Malaysia/Germany, in 2015. He joined the Department of Computer Engineering, Islamic Azad University of Qom, as an Assistant Professor. He is currently the head of the department of higher educations in computer engineering at the Islamic Azad University of Qom .His research interests include cloud computing, fog computing, edge computing, internet of things, optical wireless communications (ORCID: https://orcid.org/0000-0003-4856-9119).

Mostafa Ghobaei-Arani received the Ph.D. degree in software engineering from Islamic Azad University, Science and Research Branch, Tehran, Iran. He has published more than 50 journal and conference papers in the area of distributed computing. His research interests include distributed computing, cloud computing, autonomic computing, edge/fog computing, exascale computing, soft computing, and the IoT. He has served as a member of editorial board and review committee for a number of peer-reviewed international journals and PC member of various international conferences. (https://publons.com/researcher/1267819/mostafa-ghobaei-arani/).

The Innovation Roadmap and Value Creation for Information Goods Pricing as an Economic Commodity

Hekmat Adelnia Najafabadi
Department of Management, Najafabad branch, Islamic Azad University, Najafabad, Iran
hek.adel@gmail.com
Ahmadreza Shekarchizadeh*
Department of Management, Najafabad branch, Islamic Azad University, Najafabad, Iran
ahmad_shekar2@hotmail.com
Akbar Nabiollahi
Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad, Iran
a.nabi@pco.iaun.ac.ir
Naser Khani
Department of Management, Najafabad branch, Islamic Azad University, Najafabad, Iran
naserkhani@phu.iaun.ac.ir
Hamid Rastegari
Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad, Iran
rastegari@iaun.ac.ir

Received: 17/Aug/2019 Revised: 16/Nov/2019 Accepted: 29/Dec/2019

Abstract

Nowadays, most books and information resources or even movies and application programs are produced and reproduced as information goods. Regarding characteristics of information goods, its cost structure and market, the usual and traditional pricing methods for such commodity are not useful and the information goods pricing has undergone innovative approaches. The purpose of product pricing is to find an optimal spot for maximizing manufacturers' profits and consumers' desirability. Undoubtedly, in order to achieve this goal, it is necessary to adopt appropriate strategies and implement innovative strategies. Innovative strategies and tactics reflect the analysis of market share, customer behavior change, pattern of cost, customer preferences, quick response to customer needs, market forecast, appropriate response to market changes, customer retention, discovery of their specific requirements, cost reduction and customer satisfaction increase. In this research, 32 papers have been selected among 540 prestigious articles to create a canvas containing more than 20 possible avenues for innovations in the field of information goods pricing, which can be used in the companies producing information goods, regardless of their size, nationality, and type of information goods they produce. Introduction of some key ideas on how to increase both profits and customer satisfaction and also three open issues for future research in the field of information goods pricing is one of the achievements of this research.

Keywords: Innovation; Pricing; Information goods; Customer satisfaction.

1. Introduction

An exciting and rapid increase in global competition in production, marketing and sale of information products, along with a change in economy into a knowledge-based economy, creates a renewed emphasis on innovation in these scopes. This new economy, especially by those who are innovate, is driven faster than other competitors, but with respect to the growing number of types of information and the immediate exchange of it in electronic world, there has not been such a considerable need for innovations in the pricing of information goods in no time like today. Therefore, in order to be able to survive in the thriving market of information products and overcome competitors, improve customer retention and attract the new ones, traditional methods of marketing and pricing of these commodities must be abandoned and the innovative methods should be invented.

The term "innovation" is a broad concept as a process for using knowledge or information for the purpose of creating or introducing new and useful things. In other words, innovation refers to everything which is revised to become reality, and strengthens the position of the organization against competitors, and provides a long-term competitive advantage [1]. Market innovation involves up-

* Corresponding Author

to-date knowledge in distribution channels, products and its usage to meet expectations, value, and demands of customers which its main goal is to improve marketing-mix (product, price, place, and promotion). On the other hand, an appropriate and innovative pricing strategy will lead to the creation of competitive advantage for the company, as well as among the factors of marketing-mix, price is the only element of revenue-generating value which has more flexibility, because it can be quickly changed [2].

With the advent of the Internet and the creation of information products, the concept of pricing for information goods has become one of the most interesting and innovative topics in information technology management and economy. On the other hand, regarding characteristics of information goods, its cost structure and market, the usual and traditional pricing methods for such commodity are not useful and the information goods pricing has undergone innovative approaches. With this view, perhaps the most important gap in the research literature on information goods pricing is the lack of detailed and extensive research on innovations in pricing of these commodities in terms of customer satisfaction and more profit for sellers.

The purpose of product pricing is to find an optimal spot for maximizing manufacturers' profits and consumers' desirability. Undoubtedly, in order to achieve this goal, it is necessary to adopt appropriate strategies and implement innovative strategies and tactics. Accordingly, the main issue of this article is to explore innovations in the field of information goods pricing with providing several key ideas on how to increase profits and customer satisfaction. Therefore providing a marketing canvas, including more than 20 solutions for innovation in the field of information goods pricing is one of the results of this study that can be used in organizations producing information goods regardless of size and nationality, and the type of information product. The rest of this paper is organized as follows. Section 2 describes the theoretical basis of research used throughout this paper. Section 3 contains the research background and shows categorizing method for innovations that made by previous researchers in the field of pricing of information goods. Sections 4 to 5 describe research method and introduces innovation road map framework for information goods pricing along with three open issues for innovation and value creation in this field. Finally, Section 6 presents our conclusions.

2. Theoretical Basis of Research

Pricing is simply means placing a value on a product or service [3]. Pricing is an activity that needs to be repeated and is a continuous process [4]. This continuity is due to environmental and volatile market conditions that necessitate price adjustments. An innovation in pricing, for example, includes this issue that organizations in choosing pricing strategies, pricing tactics, and organizational factors act in a way that using consumers' psychology leads to a change in their perceptions of the value and price of the purchased item [5].

Information or digital goods are defined as a commodity that can be digitally transmitted through information networks such as the Internet. Examples of information goods include computer software, e-books, online magazine, databases, music, movies, television programs and search engines such as Google and Yahoo [6].

Kai and Patrick [7] have presented interesting ranking of information goods. They have divided information goods into tools and services (such as anti viruses) of content-based digital products (such as books and magazines) and on-line consulting services, and mentioned some of the unique features of information products: lack of erosion, easy copying, share ability, network effect (refers to the more people use the information, the more people will want to use it) and trial ability use. The cost structure of information goods is also such that the cost of reproducing it from the original version is negligible, and policies must be adopted so that the cost of the original version is abated according to the amount of demand on the reproduced copy.

3. Research Background

In today's complex and competitive world, innovation is vital for every type of business in the economy and technology of any society, and it can be said that without innovation, any business is condemned to destruction. The importance of innovation in the pricing of information goods due to its specific characteristics has led to valuable studies in this context. Hence, with the provision of access levels and licenses, Steele (2003) [8] has created an innovative three-dimensional model base on following dimensions; (what is sold, such as goods and services), time frame such as (One-time, Perpetual or Subscription), a pricing policy (such as Per user, Per CPU or Per use). Dixit et al. (2008) [9] introduce advances in technology, real-time computing capability, and quick access to marketing databases as key elements that have led to smart factors technology to use in performing for any innovative pricing policy. They praised the capabilities of smart mobile agents for bargaining, because they believe that intelligent agents can access competitive information and use them creatively for competitive advantage.

The wider range of innovations in information goods pricing shows that among researchers in this area, some have specially introduced innovations in retail and e-commerce. For example, [10] have grouped innovation in retail by answering the three main questions (whom to target, what promotions and pricing models to use, how design elements can increase the effectiveness of these promotions?). In order to answer the first question, they have mentioned pricing on the basis of the value of customers, and in order to answer the second question, dynamic pricing is specifically proposed. They also suggest the use of RFID technology and trading stalls in response to the third question. Andreas Hinterhuber and his colleague Stephen Liozu [5] are among other researchers who have specially categorized innovations in pricing into three categories: Innovation in Strategy, Innovation in Tactics, and Organizational Innovations.

Since the present article introduces innovations in information goods pricing by considering its specific features based on [5] division (strategies, techniques, and organizational factors), Table 1 summarizes only three examples of research conducted. It addresses each category and the rest will be discussed in Section 5, depending on the structure of the article

Table 1-Example researches on the field of pricing information goods

Research / study	Subject/Context	Innovation category
Pascual-Miguel et al(2015)	Market segmentation based on demographic information	Strategy
Larson(2014)	Psychological pricing	Strategy
Yao(2012)	Dynamic pricing	Strategy
Carlson &Kukar-Kinney(2018)	Using creative discounts	Tactic
Morrison(2016)	Bundling for information goods pricing	Tactic
Massoud & Aboriz-ka(2012)	Price personalization	Tactic
Laatikainen & Ojala(2018)	Using Pricing Committee	Organizational factors
Wikner(2018)	Using CRM software in the organization	Organizational factors
Ahmed et al(2018)	Using key performance indicators for pricing	Organizational factors

4. Research Method

This paper is a kind of review research that in its first step, as shown in Fig. 1, the process of data collection has been done through the Internet search of (Science Direct, Google, Magiran [1] and Irandoc [2]) databases limited to Persian and English and without time limitation, with the key words of pricing information goods, innovation in commodity pricing, and their Persian equivalent, with a preliminary selection of 540 articles. In the second step, by reviewing the title and abstracts of all articles related to the field of research, the rest were excluded from the study list. In the third step, by analyzing the section of the introduction and conclusion, articles that somehow included innovations in commodity pricing, especially information commodity have been chosen. In the fourth step, with a full review of the refined articles, 21 articles completely related to the topic of the article have been selected, and in the final step, by returning to the references, among the books and the selected articles, 32 cases of authoritative and experienced authors which repeatedly were cited, were selected to provide the results of the research.

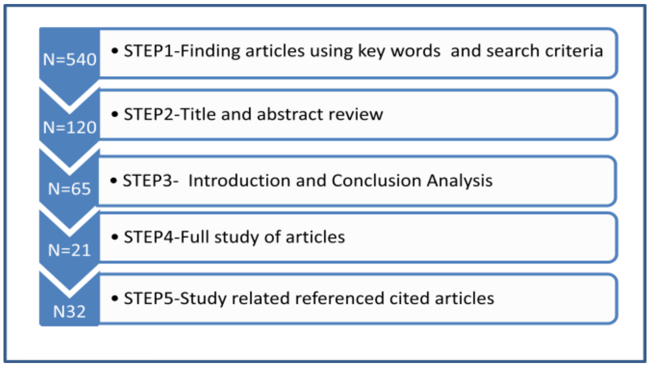

N=540
• STEP1-Finding articles using key words and search criteria
N=120
• STEP2-Title and abstract review
N=65
• STEP3- Introduction and Conclusion Analysis
N=21
• STEP4-Full study of articles
N32
• STEP5-Study related referenced cited articles

Fig 1-The process of data collection

5. The Conceptual Model of The research

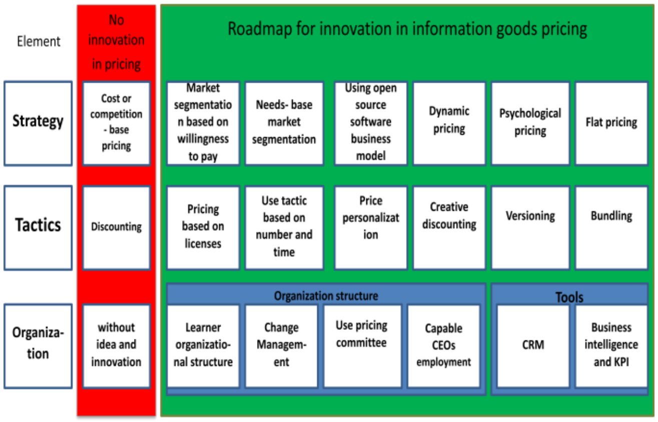
The main contribution of this paper is to provide roadmap for information goods pricing so the framework presented in this study (Figure 2) has been developed with the expansion of the Hinterhuber and Liozu (2014) [5] categories for pricing innovations, considering the specific characteristics of information product for this type of commodity. These two researchers (Hinterhuber and Liozu) believe pricing strategies in organizations that do not use innovation are largely based on competitive or cost-based pricing and pricing tactics are limited to discounting. In addition, they believe that organizations that do not use innovation in pricing do not have a specific procedure to apply discounts on pricing and a sales manager is likely in charge of pricing and price regulation. With this view, the present paper, by expanding the framework provided by Hinterhuber and Liozu, categorizes information pricing innovations into three key domains of strategy, tactics, and organizational factors for successful companies in the field of information products pricing. This framework can be used as a road-map for facilitating the choice of pricing strategy and tactic, as well as creating an organizational mechanism for the proper pricing process by the managers of information-producing companies and encourage them to choose a right alternative for their current pricing strategy or tactic. In the following, the dimensions of the proposed framework will be described and while presenting objective examples, we take a glance at the innovations presented in these areas for information products pricing and then introduce some open research issues for future studies in this field.

[1] -www.Magiran.com(All Iranian Magazines & Scientific Journals)

[2] -https://irandoc.ac.ir/(Iranian Research Institute for Information Science and Technology (IranDoc))

5.1 Innovations in The field of Information Goods Pricing Strategies

Pricing with the goal of survival and subsistence of the company, to maximize the company's current profits and market share, and moving forward are done in terms of quality [11]. Therefore, selective strategies for pricing should also be chosen according to the company's goals. In the following, some of the most important innovative strategies for pricing information products will be discussed.



Element	No innovation in pricing	Roadmap for innovation in information goods pricing					
Strategy	Cost or competition - base pricing	Market segmentation based on willingness to pay	Needs- base market segmentation	Using open source software business model	Dynamic pricing	Psychological pricing	Flat pricing
Tactics	Discounting	Pricing based on licenses	Use tactic based on number and time	Price personalization	Creative discounting	Versioning	Bundling
		Organization structure				Tools	
Organization	without idea and innovation	Learner organizational structure	Change Management	Use pricing committee	Capable CEOs employment	CRM	Business intelligence and KPI

Fig 1-The framework for selecting innovations in the field of information goods pricing

5.1.1 Market Segmentation Based on Customer Willingness to Pay

Good market segmentation provides customers with products at different prices to attract them with a willingness to pay. The ability to test the information product (demo version) as a trial version, as well as the ability to restrict access to information product at times and locations makes these goods have potential to market segmentation based on customer preferences. This strategy is also called "price discrimination" in some sources. The basis of the price discrimination strategy is that the willingness to pay by different customers for a particular item is not the same. Thus, with the help of discriminatory prices, consumer groups tend to be absorbed with a willingness to pay and compared to standard prices, discriminatory prices can increase revenue and profits for the producer and utility for the consumer [6]. Pigou in [6] about price discrimination creates a useful framework for developing pricing strategies. He has divided the price discrimination into three degrees; first, second and third. In recent years, especially after Shapiro and Varian studies in 1998, these pricing methods have also been known as "full price discrimination", "menu pricing" and group pricing. First-degree price discrimination occurs when sellers receive different prices for a single item, that is, the price received, such as auctioning for each commodity unit, is equal to the maximum willingness of individuals to pay [6]. Of course, auctioning often involves the sale of antique or special goods, but nowadays companies such as eBay and Amazon have implemented the auction process with the help of Internet technology for all the products, including information goods.

Second-degree price discrimination also implies the same price of goods among different consumers. In this type of price discrimination, every consumer encounters the same price plans, and the consumer chooses which level to use [6]. Because of their own choices, this price discrimination is also called a menu pricing (such as choosing high quality music versus choosing music with lower quality). In third-degree price discrimination, the same information product is presented to different groups at different prices, so that the price is the same within the groups. For this reason, third-degree price discrimination is also called group pricing. This strategy is used when the customer groups can be easily identified [6].

5.1.2 Market Segmentation Based on Customer Needs

Market segmentation based on customer needs is defined as a process in which the market is divided into different sectors of potential customers, with the needs and / or similar characteristics. The characteristic of customers in each section is the similar shopping behaviors. The dividing the market based on customer needs is usually done by segmentation based on demographic characteristics, behavioral characteristics, psychological characteristics and geographic features.

Demographic segmentation is widely used in software corporations because information product can be produced according to age, gender, and even occupation and field of study [12]. In market segmentation based on a customer's behavior pattern, the product is actually marketed according to customer behavior. Psychological division also emphasizes the psychological aspects of consumer shopping behavior. These psychological aspects may be consumers' lifestyles, social status, activities, interests and beliefs [13]. People are also divided according to geographic area in geographical segmentation. This means that customers have different needs based on the geographical area they are in. Geographical segmentation in the field of production and pricing of information goods is very crucial because businesses are expanding locally and internationally. Hence, based on what has been said, it is possible to identify the opportunities. Moreover, information products can be produced ad priced in a way that ultimately results in more profits for sellers and more satisfaction for customers.

5.1.3 Using Open Source Software Business Model

Offering a free of charge product will increase market share, brand popularity among customers, and attract customers to other products or services provided by a company [14]. Hence, offering free products is one of the successful strategies of information-producing companies. The free search service and advertising, internet calling service and fixed line calls at the price set by Skype, and the provision and production of open source software are examples of offering free of charge information product. Hence, offering complimentary products is one of the successful strategies of information-producing companies.

Open source software is referred to the one that people can use to copy, modify, or publish its source code using a copyright license. Awareness of open source business models and the innovative and intelligent use of these strategies can be used to price this segment of information products. Table 2 is Onetti and Verma (2009) [15] Research and includes the most prominent open source business models that are used to adopt an appropriate strategy in pricing information products.

Table 2-The Most Important Business Models for Open Source Software (Onetti & Verma, 2009)

Model	Brief description/examples of strategy choices for pricing
Independent software sales	Distribution of open source software (sales on CD instead of download), membership in software, proprietary software-based software, dual licensing (free and paid licenses), franchising, delayed presentation of successful publication
Service providing	Support Services, Maintenance Services, Database categorizing including permit, Training, Certification issuance for individuals, Open source Software components selection
Brand sales	When a company creates a well-known brand, it can be given over to another person in order to marketing in return for receiving some amount of money.
Advertising	At every step of distributing an open source component, online advertising can be done by adding links to main pages, download pages, support pages, and ...

5.1.4 Dynamic Pricing

Dynamic pricing is pricing in an environment where prices are not fixed but flexible [16]. Dynamic pricing is a strategy in which prices are changed at any given time for customers and consumers or because of the set of products and services provided. In other words, the price is not fixed in this strategy and varies in response to supply and demand conditions. Dynamic pricing in the field of information goods has many examples, such as [17] that has classified penetration price (start with low prices and rise prices according to the conditions) and price skimming (start at high prices and reduce prices according to the circumstances) as dynamic pricing strategy, but because of the particular potentiality of information goods for dynamic pricing, this strategy seriously needs innovations in this regard.

5.1.5 Psychological Pricing

Psychological pricing is a strategy in which the psychological aspects of the price that stimulate customers to buy products and services are considered. Larson (2014) [18], in a research entitled "Psychology Pricing Principle for Organization with Market Power", argues the pricing principles which are originated from the principles of psychology. He has described 51 psychological principles used in pricing and concluded that these principles would result in more profit. He also categorizes the principles of psycho-pricing into four principles including highlighting (such as price comparison, for example, less than the cost of a cup of coffee per day), proportionality (such as fairness guaranteed, for example, prices remain fixed over three months), background (such as product order, for example, attractive products are displayed at first, and the signaling theory (such as displaying small numbers in the pricing, for example, using number 9 on the right side of the price).

5.1.6 Flat Pricing

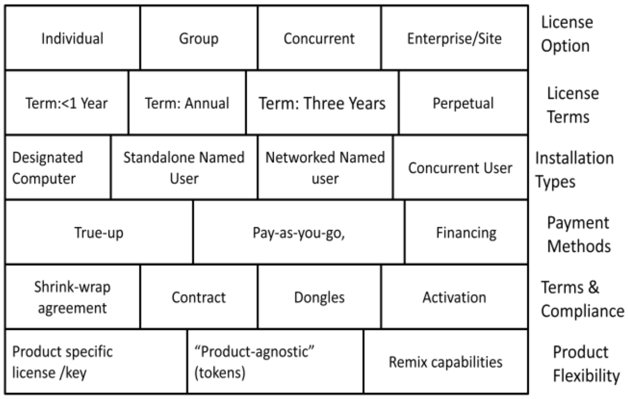
Flat pricing is a pricing strategy in which prices are offered to customers without any exceptions. This type of pricing can be popular for customers and can dramatically improve the sales of a product or service. Flat pricing is easy to be managed and advertised. Furthermore, it results in management and operational costs reduction. In addition, it can cause word-of-mouth advertisement leading to more product sales. For example, some phone operators offer the same unlimited bandwidth prices for all users. Customers also benefit from flat pricing, which can predict the exact cost of a seller's product or service, regardless of usage, which may lead to more sales. However, when the product is consumed at very different rates, the seller must have a thorough understanding of the costs for consumer goods and raise flat price high enough that consumers with high consumption do not reduce the profit [19].

5.2 Innovation Tactics Used for Information Goods Pricing

Making money from customers may be the same as creating an artwork. The art of goods pricing, especially information products, is revealed with innovative tactics. As you know, there are numerous pricing tactics for these kinds of goods that can be used. Using these tactics depends on the factors such as marketing budget, purchase behavior, perceived value, or customer life expectancy. In the following, some of the most important tactics of information goods pricing are described with respect to the specific features of this product.

5.2.1 License-Based Pricing

A license is an official permission or a software protection tool that grants users the legal right to use the software [20]. The cryptographic feature and the ability to control licenses for having access to the content of the information product have made these kinds of goods potential to use the selection of license-based pricing tactics. Figure 3 shows license attributes for information products, especially software provided by [21] for software and information goods pricing.



Individual	Group	Concurrent	Enterprise/Site	License Option
Term:<1 Year	Term: Annual	Term: Three Years	Perpetual	License Terms
Designated Computer	Standalone Named User	Networked Named user	Concurrent User	Installation Types
True-up	Pay-as-you-go,	Financing		Payment Methods
Shrink-wrap agreement	Contract	Dongles	Activation	Terms & Compliance
Product specific license /key	"Product-agnostic" (tokens)	Remix capabilities		Product Flexibility

Fig 2-License attributes (Nayak, 2006)

5.2.2 Using Time-Frame and Counter-Base Tricks

The ability to control the number of allowed accesses as well as the time access control of the information goods, and also the ability to distribute information goods on the network, leads to creation some tricks for pricing information goods based on the counting and timing. Economists use the term "price elasticity" to illustrate the extent of correlation between price and demand for a commodity. In the case of information commodity, the degree of price elasticity depends on the nature of the information goods, the purpose of its creation, the circumstances in which the need is created, the person who uses it, the level or amount of processing necessary to make it useful, and the accessible time available to users. Hence, tricks have been developed to use these flexibilities. Figure 4 illustrates intellectual property that can be licensed along three dimensions for using these tricks (Steele, 2003) [8]. These dimensions include the nature of the information goods (what is being sold?) the pricing policy and strategy (such as pricing in terms of the number of clients), and the timing of the sale or possession (such as the permanent purchase or time subscription) of the information product.

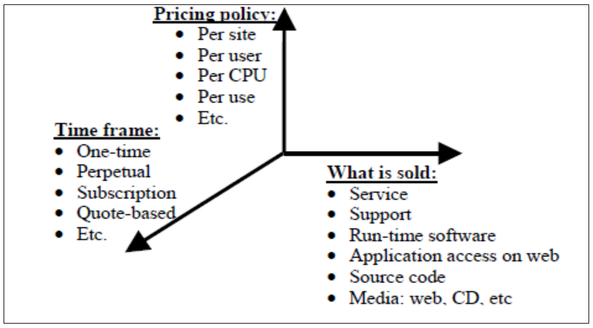


Fig 3-Intellectual property that can be licensed along three dimensions: (1) what is sold, (2) pricing policy, and (3) time frame [8].

5.2.3 Price Personalization

Price personalization can usually be done by having enough information from the customers, namely having the customer's history of purchasing, along with their profile information. The research by Massoud and Abo-Rizka (2012) [22] is a comprehensive example of price personalization by using recommender systems for customizing information goods pricing. The two Egyptian researchers have developed a three-part conceptual model consisting price controller, the customer valuation and discount category, in which users are evaluated online, and then the appropriate product is offered to customers for purchasing. A remarkable point in this research is the customer valuing segment where customer is valued based on considering sales records and user profiles, then price is personalized for customer.

5.2.4 Creative Discounting

In addition to increasing sales (i.e. massive sales and packaging several products together and making customers to buy more products), offering discount is a tactic that can gain other benefits such as the positive feelings of customers because of dealing with sellers, as well as helping them to choose sellers' products among competitors and creating a new opportunity and reaching new customers. Some of the creative discounts used for information goods include non-linear pricing (buy one for get two), permanent discounts, prize packs, purchase coupons, discounts for student or seniors [23].

5.2.5 Versioning

With this tactic, the company offers its products in various versions, among which, customers choose the right editorial version they are willing to pay for it [24]. This tactic is very effective in situations that customers are dispersed in different demographic groups and are not identifiable by the producers [6]. In this tactic, versions that are designed for people tending to pay higher prices should have advantages to versions designed for people willing to pay lower. Creating different versions of products can be based on product quality. This method is useful especially when reducing the quality of the information goods to produce a version with lower quality is not expensive [25].

5.2.6 Packaging

Packaging means selling two or more products or services as a pack, which means selling two or more products at a price lower than the total price of each item when they are sold individually which is attractive for customers and increases profits for sellers [26]. This method allows companies to have more market share through packaging products. The positive impact of packaging or selling information goods has been proven by researchers such as [27], and is particularly common for information goods. Selling and distributing Windows software with Windows operating system on a CD is one example of using packaging tactic and successful sales of information goods.

Since most of information goods are accessible through Internet, it appears that analysis of user interactions with Internet and using filters in advisory systems can be helpful in identifying the components that should be included in a package. Whether components of the packages should be reviewed over time, as well as the number of optimal packages that the producers have to offer, are questions that need to be investigated.

5.3 Innovations Offered in the Context of the Organizational Factors of Information Goods Pricing

An innovative organization is the one that has institutionalized innovation in all parts of the organization, in all aspects of business and among all members of the work teams. An innovative organization creates an environment that is fully engaged in positive change and a culture that is rich in creativity and recreation [28]. The classification of organizational factors in the field of products pricing is one that has not been ignored. Liozu and Ecker [29] considering organizational structure, has categorized innovation for pricing into centralized, decentralized, based on pricing and support by the leader. In this paper, considering the specific features of the information goods, as well as research and analysis of the articles and selected books of organizational innovations in two areas of organizational structure, such as creating an organizational structure of the learner in the organization and tools used in the organization for pricing, such as the use of business intelligence tools will be discussed.

5.1.3 Creating a learning Organization Structure

A learning organization is the one that helps improve organizational learning through creating structures and strategies. The organization has the skill and ability to create, acquire and transfer knowledge, and modify its behavior so that it reflects up-to-date knowledge and insights. According to Peter Senge [30], the learning organization is an organization that uses individuals, values, and other sub-systems to continuously modify and improve performance by relying on the lessons and experiences it gains. Therefore, learning organizations should integrate pricing innovation with organizational learning and distributing knowledge effectively, solving issues systematically, learning from organization experiences, past events, and best actions by others in order to empower members to learn and develop their skills in pricing. Expanding the learning culture will lead the organization to choose the best pricing strategies and tactics [5].

5.3.2 Change Management

Innovation in pricing basically leads the organization to manage changes. Choosing a pricing strategy or tactic is not just a change in marketing and sales, and the complexity of its activities is more than the change in price list. Therefore, new pricing methods often require change management with new capabilities, a new organizational structure, different target and motivational systems, new processes and tools, and new organizational prerequisites, as well as a welcome to constant change in organizational structure with open arms lead the organization toward the best choice for pricing strategy and tactic. Hence, innovation in pricing should be viewed from a organizational perspective as a continuous change, rather than a short-term project process [31].

5.3.3 Using a Pricing Committee

The digitization, servicing and cloud computing, and the specific characteristics of the information goods and its related business models are shaped so that the pricing committee of these types of goods should have an acceptable understanding of its resources and pricing capabilities. Hence, the positive impact of having a pricing team for information goods has been proven not only by the investigators but also the number of members of the pricing committee, the duties and skills of each member; including technical skills such as data collection skills, software skills, analytical skills, skill and experience in pricing, management skills, creativity and risk-taking against pricing are explained in details for choosing a pricing team [32].

5.3.4 Capable CEOs Employment

Innovation is one of the key abilities of any organization's top managers which allow them to help the growth and profitability of the organization. Innovation, and consequently innovation in commodity pricing, is highly-challenging due to its closeness to risk and profitability.

Hence, the research conducted in this context shows that the positive impact of CEOs on pricing is undeniable, and their decisions significantly affect the pricing and performance capabilities of the company [33]. CEOs need to understand the importance of pricing and be enthusiastic about pricing function and providing resources to support it. Paying attention to the pricing committee, showing strength when faced with pricing barriers, identifying key players and authorities responsible for solving pricing problems and giving them confidence when these problems arise, as well as selecting and adopting an appropriate pricing strategy in a way that makes profit for company and leads to customer satisfaction is one of the major issues that is partly related to the organizational and individual innovations of CEOs.

5.3.5 Business Intelligence and KPIs

Key Performance Indicators (KPIs) are powerful and vital tools for managers and organizational leaders to understand the success rate and adopt with strategic direction of the program. Some of the indicators used for pricing goods and services, particularly related to information products, are: the number of sales, net profit, customer retention rate, customer satisfaction index, customer life span, market share, market penetration rate, Web search ranking on the Internet, click rate and customer site viewing, online customer engagement levels, software downtime, average revenue per user, support rate [34,35]. Key performance indicators are typically displayed by the Business Intelligence Dashboard which is an information management tool. Hence, successful pricing organizations can use a BI-Dashboard to view a wide range of complex data in a simple way with easy-to-understand dashboards, and make decisions on choosing a specific tactic or pricing strategy.

5.3.6 Using CRM

Having comprehensive CRM (Customer Relationship Management) software can lead CEOs to optimize pricing [36]. Microsoft defines CRM as the software used for the automation of sales and marketing activities, as well as the management of sales-related activities and services within an organization. Gartner, a giant research company in information technology, believes CRM is a business strategy that optimizes revenue and profitability while promoting customer satisfaction and loyalty [37]. CRM software provides functionality to companies in four segments: sales, marketing, customer service and digital commerce. Hence, many authors such as [38] emphasize the positive impact of CRM on business management in an organization, and have looked at it in terms of marketing mixing elements, ie, product, price, place and promotion.

5.4 Open Issues for Innovation and Value Creation in Information Commodity Pricing

Specific characteristics of information goods such as conversion of knowledge and technology to information goods, easy access, ability to transfer over the Internet and social networks, time and space restrictions removal, reduction of transaction costs and time, unnecessary intermediaries elimination and many other benefits, has increasingly led to the growth of information goods. After reviewing the compiled articles, despite the numerous needs for innovation in the field of information goods pricing, three areas of studies have been identified for pricing such goods, including; target market identification, demand estimation and using revenue management systems which will be briefly mentioned following.

5.4.1 Target Market Identification and Market Segmentation

In the past, the market was referred to a place where goods were traded. Philip kotler [39], in his marketing book, describes the market as a potential and realistic set of buyers that exists for the product. With the advent of the Internet, as well as the creation of information goods, this concept has been modified considerably. Researchers of this study believe that the target market for information goods includes those who, while interested in the product, have enough resources to buy the product and can have access to the product and are not legally prohibited to buy it. Finding such individuals in the information goods market can be done using veritable and innovative methods. Determining the target market can, in addition to generating motivation for investment, production and sales, is also a measure for decision makers of manufacturing companies in the field of information goods pricing. Therefore, according to studies conducted in this paper, research to identify the target market of information goods that may have global reach and be transmitted with one simple click to the rest of the world is open issues for future research.

5.4.2 Demand Estimation

If the sale of a new product is unlikely to reach the break-even point, the company should not bother manufacturing it. Unfortunately, there is no way to know the future status of a product sale precisely. Therefore, the only possible option is prediction and estimation. Since each information goods is designed and manufactured for its target community, estimating the demand and determining the target market plays an important role for decision making by producers as well as produced commodity pricing. Fast transfer, unrestricted access without time limitations on the Internet at any time, the impact of internet browsing by Internet users, the possibility of market segmentation based on demographic characteristics (such as gender, age, field of study) is only one distinguishing difference in demand estimates for information goods. Hence, with regard to the domain of the subject, research is open to identifying and developing appropriate ways to estimate the demand for information goods.

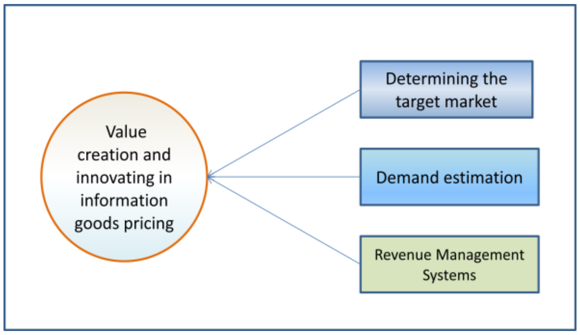

Value creation and innovating in information goods pricing
Determining the target market
Demand estimation
Revenue Management Systems

Fig 4-Open issues for innovation and value creation in information goods pricing

5.4.3 Revenue Management Systems

Revenue management systems have been used since the mid-1980s in the airline ticket sales industry. This tactic is mostly used for perishable goods or the ones with expiration date. For example, airplane seats are not worthwhile after flying. The classic model of revenue management systems is a sort of capacity allocation, which seeks to find the best way to allocate capacity based on the assumption of the definite demand. The dynamic pricing model of this solution involves selling the right product to the right buyer at the right time and price, in which the customer's desire is to pay for the issue which is in the center of attention [40,41]. This means that customers should not be alike in their desire to pay, that is, they usually offer different prices for the same goods and services. On the other hand, in this model, different groups of buyers should be identified. For example, when booking a plane ticket, a price insensitive customer to the price is only sensitive to the time and date of the flight, while the price sensitive customer seeks to minimize the cost and price. Since for information goods, a good segmentation can be done for selling right product to the right buyer at the right time and at the right price and based on customer's willingness, pricing models according to this issue, is another open issue for future studies has been neglected by the researchers of this realm.

6. Conclusion

As business environment is extremely dynamic, and only the companies that are superior in terms of competitiveness, can succeed in achieving a sustainable market, Information technology companies must seek innovative solutions for pricing their goods along with customer satisfaction and more profit. Innovative solutions that reflect the analysis of market share, customer behavior change, cost pattern, customer preferences, quick response to customer needs, market situation forecast, proper response to market changes, customer retention, discovery of specific requirements, cost reduction and increase customer satisfaction.

This article explores innovative strategies in the field of strategies, tactics and organizational factors for information goods pricing, and ideas have been presented to get more profit by sellers and increase customer satisfaction. The presentation of a canvas with more than 20 possible avenues for innovation in the field of information goods pricing, which can be used in information-producing companies, regardless of size, nationality, and type of information commodity produced. Also, the introduction of open issues for future research, including demand estimates, dynamic pricing and revenue management systems in the field of information goods pricing, are other achievements of this study which outlines a clear vision for future research. In the end it should be said, innovation in pricing process is not a phenomenon that occurs only once, but a continuous process consisting of innovation in strategies, tactics and organizational factors that must be invented and implemented at all stages, from the production stage to the sale, and even its support.

References

[1] K Holt, *Product innovation Management*. London: The University Press, 1998.

[2] Shahriyar Azizi, "Pricing in Electronic Markets," *Tadbir*, no. 186, p. [In Persion], 2008

[3] V. H Spingies and A. S. A du Toit, "Pricing of information products: three scenarios," *Library Management*, vol. 18, no. 5, pp. 301-304, 1997

[4] D Shipley and D Jobber, "Integrative Pricing Via The Pricing Wheel," *Industrial Marketing Management*, vol. 3, no. 30, pp. 301-314, 2001

[5] A Hinterhuber and S.M Liozu, "Is Innovation in Pricing Your Next Source of Competitive Advantage?," *Business Horizons*, vol. 57, no. 3, pp. 413-423, 2014

[6] K. L Wang, "Pricing strategies for information products: A review," in *In IEEE International conference on e-commerce technology for dynamic E-business*, 2004

[7] L Kai and Y.K. C Patrick, "Classifying digital products," *Communications of the ACM*, vol. 45, no. 6, pp. 73-79, 2002

[8] R Steele. (2003) Software business models. [Online]. http://www.corp21.com/download/SWBusinessModels.pdf

[9] T Dixit, G Whipple, E Zinkhan, and A Gailey, "A Taxonomy of Information Technology-Enhanced Pricing Strategies," *Journal of Business Research*, vol. 61, no. 4, pp. 275-283, 2008

[10] Dhruv Grewal et al., "Innovations in Retail Pricing and Promotions," *Journal of Retailing*, pp. 43-52, 2011

[11] Amir Baghtaee and Shadi Golchinfar, *Tadbir*, 2008

[12] F Pascual-Miguel and et all, "Influences of gender and product type on online purchasing," *Journal of Business Research*, vol. 68, pp. 1550-1556, 2015

[13] C Antoun, "Who Are the Internet Users, Mobile Internet Users, and Mobile-Mostly Internet Users?: Demographic Differences across Internet-Use Subgroups in the U.S," in *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies*, D Toninelli, B Pinter, and P Pedraza, Eds. London: Ubiquity Press, 2015, pp. 99-117

[14] K Shampanier, N Mazar, and D Ariely, "Zero as a special price: The true value of free products," *Marketing Science*, vol. 26, no. 6, pp. 742-757, 2007

[15] A Onetti and S Verma, "Open source licensing and business models," *Journal of Knowledge Management*, no. 7, pp. 68-94, 2009

[16] R Mohammed, R. J Fisher, and B. J Jaworski, *Internet Marketing: Building Advantage in the Networked Economy*.: McGraw – Hill press, 2002

[17] S Lehmann and P Buxmann, "Pricing Strategies of Software Vendors," *Business & Information Systems Engineering*, vol. 6, no. 1, pp. 452–462, 2009

[18] B. R Larson, "Psychology Pricing Principle for Organization with Market Power," *Journal of Applied Business*, 2014

[19] L. W McKnight and J Boroumand, "Pricing Internet services: Approaches and challenges," *Computer*, vol. 32, no. 2, pp. 128-129, 2000

[20] A Morin, J Urban, and P Sliz, "A quick guide to software licensing for the scientist programmer," *PLoS Comput Biol*, vol. 8, no. 1, pp. 1-7, 2012

[21] Shivashis Nayak, "Pricing and licensing of software products and services: A study on industry trend," System Design and Management Program, MSc 2006

[22] M Massoud and M Abo-Rizka, "A Conceptual Model of Personalized Pricing Recommender System Based on Customer Online Behavior," *IJCSNS International Journal of Computer Science and Network Security*, vol. 12, no. 6, 2012

[23] J. R Carlson and M Kukar-Kinney, "Investigating discounting of discounts in an online context: The mediating effect of discount credibility and moderating effect of online daily deal promotions," *Journal of Retailing and Consumer Services*, no. 41, pp. 153-160, 2018

[24] C Shapiro and H. R Varian, "Versioning: The smart way to sell information," *Harvard Business Review*, pp. 106-114, 1988

[25] Kang Bae Lee, Sung-Yeol Yun, and Seong Jun Kim, "Analysis of pricing strategies for e-business companies providing information goods and services," *Computers & Industrial Engineering*, vol. 1, no. 51, pp. 72-78, 2006

[26] F Linde, "Pricing information goods," *Journal of Product & Brand Management*, vol. 18, no. 5, pp. 379-384, 2009

[27] W Hui, B Yoo, V Choudhary, and K. Y Tam, "Sell by bundle or unit?: Pure bundling versus mixed bundling of information goods," *Decision Support Systems*, vol. 53, no. 3, pp. 517-525, 2012

[28] J Rowley, A Baregheh, and S Sambrook, "Toward an innovation type mapping tool," *Journal of management decision*, pp. 73-86, 2011

[29] S Liozu and K Ecker, "The organizational design of the pricing function in firms," in *Innovation in Pricing: Contemporary Theories and Best Practices*. New York: Routledge, 2013, pp. 27-46

[30] P. M Senge, "The art and practice of the learning organization," *The new paradigm in business: Emerging strategies for leadership and organizational change*, pp. 126-138, 1990

[31] A Jerome, Colletti, and B Lawrence, "Change Management Initiatives: Moving Sales Organizations from Obsolescence to High Performance," *Journal of Personal Selling & Sales Management*, vol. 17, no. 2, pp. 1-30, 1997

[32] G Laatikainen and A Ojala, "Pricing of digital goods and services," in *Information Systems Research Conference in Scandinavia*, 2018

[33] S Liozu and A Hinterhuber, "CEO championing of pricing, pricing capabilities, and firm performance in industrial firms," *Industrial Marketing Management*, vol. 42, no. 4, p. 633—643, 2013a

[34] Haris Ahmed, Tahseen Ahmed Jilani, Waleej Haider, and Mohammad Asad Abbasi, "Establishing Standard Rules for Choosing Best KPIs for an E-Commerce Business based on Google Analytics and Machine Learning Technique," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 5, pp. 12-24, 2017

[35] C Aydede and T Turkoglu, "How to utilize a value-based pricing strategy in service contracts: A descriptive case study of how a Swedish pricing consultancy company optimizes pricing of services for its customers," Master thesis 2017

[36] Sarah Wikner, "The Next Generation CRM Tools: Bridging the Gaps between Sales Needs and CRM Tools Architecture," in *Organizing Marketing and Sales*., 2018, pp. 195-205

[37] Gartner Group. (2018) Gartner IT Glossary - Customer Relationship Management (CRM). [Online]. https://www.gartner.com/it-glossary/customer-relationship-management-crm/

[38] Hamed Tohidi and Mohammad Mehdi Jabbari, "CRM as a Marketing Attitude Based on Customer's Information," *Procedia Technology*, vol. 1, pp. 565-569, 2012

[39] Philip T Kotler, *Marketing Management*, 14th ed.: Pearson, 2011

[40] T Fiig, O Goyons, R Adelving, and B Smith, "Dynamic pricing – The next revolution in RM?," *Journal of Revenue and Pricing Management*, vol. 15, no. 5, pp. 360–379, 2016

[41] Alessandro Capocchi, *Economic Value and Revenue Management Systems*.: Palgrave Macmillan, 2019

[42] H R Varian, "Pricing Information Goods," *Research Libraries Group Symposium on Scholarship in the New Information Environment*, 1995

[43] Roy Jones and Haim Mendelson, "Information Goods vs. Industrial Goods: Cost Structure and Competition," *Management Science*, vol. 57, no. 1, pp. 164-176, 2011

[44] Alper Şen, "A comparison of fixed and dynamic pricing policies in revenue management," *Omega*, vol. 41, no. 3, pp. 586-597, 2013

[45] E. T Anderson and D. I Simester, "Effects of $9 price endings on retail sales: Evidence from field experiments," *Quantitative Marketing and Economics*, vol. 1, no. 1, pp. 93-110, 2003

[46] T O'Keeffe, "Organizational Learning: a new perspective," *Journal of European Industrial Training*, vol. 26, no. 2, pp. 130-141, 2002

[47] P Rikhardsson and O Yigitbasioglu, "Business intelligence & analytics in management accounting research: Status and

future focus," *International Journal of Accounting Information Systems*, pp. 37-58, 2018

[48] Dong-Qing Yao, Ziping Wang, Samar K Mukhopadhyay, and Yu Cong, "Dynamic pricing strategy for subscription-based information goods," *Journal of Revenue and Pricing Management*, vol. 11, no. 2, pp. 210-224, 2012

[49] William G. Morrison, "Product bundling and shared information goods: A pricing exercise," *The Journal of Economic Education*, vol. 47, no. 1, 2016

[50] N Biehn and C Zawada, "Innovations in Determining Willingness-to-Pay for B2B Companies," in *Innovation in Pricing: Contemporary Theories and Best Practices*. New York: Routledge, 2013, pp. 288-297

Hekmat Adelnia Najafabadi received the B.S. degree in Computer Engineering from Shahid Bahonar Technical and Engineering College, Shiraz, Iran in 2003, and M.S. degree in Artificial Intelligence from Isfahan University of Technology (IUT), Esfahan, Iran, in 2012. Currently he is Ph.D. Candidate in Azad University, Najafabad Branch, Iran. Her research interests include Information Retrieval, Recommendation Systems, Revenue Management system, Web Mining and Data Mining

Ahmadreza Shekarchizadeh has been assistant professor in Islamic Azad University, Najafabad Branch, Iran. Recently has been retired. His research interests include Digital marketing, Service Marketing and Strategic Marketing Management.

Akbar Nabiollahi is a full-time Assistant Professor in the Faculty of computer engineering at Islamic Azad University, Najafabad Branch, Iran. He received the B.S degree in Computer Engineering from Isfahan University of Technology (IUT), Esfahan, Iran in 1994, then he has experienced for ten years in a large IT enterprise in development of Information systems. Meanwhile he has graduated in M.S degree of software engineering from Islamic Azad University, Najafabad branch in 2004. Finally he received his Ph.D. degree in Computer Science from University of Technology Malaysia (UTM), Johor Bahru, Malaysia in 2011. Now he is the head of Big Data Research Center of Islamic Azad University, Najafabad Branch, Najafabad, Iran. He is conducting research activities in the areas of Information Technology Management, IT Service Management, Enterprise Architecture, Big Data and Business Intelligence and Agile Software development.

Naser Khani is an assistant Prof. in Management and Director of management department at Najafabad Branch of Islamic Azad University (IAUN), Iran. His research interests include Strategic Management, and Management Information Systems. His research has been published in international journals and conferences.

Hamid Rastegari received his Ph.D. in Computer Science (Soft Computing) from UTM in 2011. He is currently Assistant Professor on faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Iran. His research interests include Natural Language Processing, Information Retrieval, and Semantic Web.