ISSN 1800-4156

Volume 09 Number 1 and 2 2016

International Journal on Advances in ICT for Emerging Regions

www.icter.org

International Journal on Advances in ICT for Emerging Regions

This journal is also available on-line at https://icter.sljol.info/

Publication:

Published quarterly on-line and yearly in print by University of Colombo School of Computing in collaboration with CodeGen International and the Sri Lanka Chapter of the IEEE Computer Society.

Price:

Foreign: US $ 15, Local: Rs.1000.00
Rate per issue includes postage and airmail

Disclaimer:

No responsibility is assumed by the University of Colombo School of Computing for statements and opinions expressed by the contributors to this journal.

Manuscripts:

Manuscripts and all correspondence relating to the Journal should be submitted via www.icter.org

International Journal on Advances in ICT for Emerging Regions

CONTENTS

RESEARCH ARTICLES

Volume 09 Number 1 2016
(Special Issue on 15th IEEE Conference on ICT for Emerging Regions)

Volume 09 Number 2 2016

Editorial

Abhaya Induruwa

Computers of whatever hue, from silicon based to quantum, bio to optical, and the associated act of computation have now become indispensable for the creative acts of humans as amply demonstrated by the emerging developments in the ICT field. Information, be it a creation of humans, or one that is naturally occurring and is captured or presented in voluminous content and velocity is the ideal raw material for computers for which computer scientists are constantly on the lookout for efficient breakthroughs in processing approaches. In this spirit, Volume No 9 of *The International Journal on Advances in ICT for Emerging Regions 2016* consists of a special issue comprising extended versions of three of the best papers on cutting edge research in computing presented at the sister conference ICTER 2015 held in December 2015, in Colombo and two more articles selected from regular submissions.

The five papers in this issue broadly cover areas of graph analytics in citation ranking, semantic web search, nature inspired optimisation in numerical computing, machine learning for data mining, and human motion modelling with sensor data.

The paper on *Improving Citation Network Scoring by Incorporating Author and Program Committee Reputation* looks at how to select the best venue for your publications. In the academic world much emphasis is placed on publications of research outcome, and in the field of Computing, deciding on an appropriate publication venue is a hard task given the proliferation of such venues. The authors argue that it is important to take the quality of citations over a publication history in deciding the venue. Towards this the authors modify the well-known page rank algorithm and also compensate for new venues by taking into account of the reputation of program committee members.

The paper on *Path Index Based Keywords to SPARQL Query Transformation for Semantic Data Federations* addresses key word search in the newly emerging semantic web. SPARQL is the natural query language for semantic web, and the authors argue that a more user friendly key word based search if mapped to SPARQL would be of immense value, and towards this proposes a path index based approach for the transformation.

The paper on *An Archived firefly algorithm; A mathematical software to solve univariate nonlinear equations* is an interesting application of a natural heuristic for optimal resolution of the roots of a non-linear equation. The recently proposed firefly algorithm emulates the inherent ability of swarms to converge on optimal solutions to hard problems, and given an initial guess at the roots of a nonlinear equation, the authors show that the modified firefly inspired algorithm results in a better quality solution than most other competing techniques.

Given that Sri Lanka boasts a system of free education, extending to university level, it is imperative that the Government makes proper decisions in regulating the country's university education system such that employable graduates are produced. The paper on *Employability and Related Context Prediction Framework for University Graduands: A Machine Learning Approach* attempts to predict a graduate's employability given the background data. Authors evaluate several machine learning approaches on census data and come out with most suitable models for the prediction.

The last paper which is on *Motion Tracking by Sensors for Real-time Human Skeleton Animation* is proposing an alternative sensor based approach for human motion capture to the widely used Microsoft Kinect that showed the proposed method has improved motion capturing properties. This line of work has diverse applications ranging from applications in robotics to 3D animation to orthopaedics.

I feel privileged to have the opportunity to write this editorial when the ICTer publications are enjoying their success in their 10th year. The untiring efforts and commitment of those associated with the ICTer during this period have helped to firmly establish both the annual ICTer conference and related publications. Had it not been for the authors who have contributed by documenting their excellent research; the efforts of the reviewers in selecting suitable papers to publish, and more importantly, extending their help to authors where needed to ensure a high quality and standard of published papers; and the hard work of the members of the editorial panel led by my predecessor Editor-in-Chief Prof. Athula Ginige, ICTer would not have been recognition as a worthy platform for research and the ICT community in the Global South. I would like to record my sincere thanks to them all.

Last but not the least I would like to take this opportunity to congratulate the authors whose papers appear here and to thank the local organizing committee for all their efforts in publishing this special issue to coincide with the 17th Annual ICTer Conference in September 2017.

August 2017

Improving Citation Network Scoring by Incorporating Author and Program Committee Reputation

Dineshi Peiris, Ruvan Weerasinghe

***Abstract*— Publication venues play an important role in the scholarly communication process. The number of publication venues has been increasing yearly, making it difficult for researchers to determine the most suitable venue for their publication. Most existing methods use citation count as the metric to measure the reputation of publication venues. However, this does not take into account the quality of citations. Therefore, it is vital to have a publication venue quality estimation mechanism. The ultimate goal of this research is to develop a novel approach for ranking publication venues by considering publication history especially to identify the key Computer Science journals and conferences from various fields of research. Our approach is completely based on the citation network represented by publications. A modified version of the PageRank algorithm is used to compute the ranking scores for each publication. In our publication ranking method, there are many aspects that contribute to the importance of a publication, including the number of citations, the rating of the citing publications, the time metric and the authors' reputation. Known publication venue scores have been formulated by using the scores of the publications. New publication venue ranking is taken care of by the scores of Program Committee members which derive from their ranking scores as authors. Experimental results show that our publication ranking method reduces the bias against more recent publications, while also providing a more accurate way to determine publication quality.**

***Keywords*—Ranking, Citation Network, Publication Venues, Publications, Publication Authors**

I. Introduction

The Internet has opened up new ways for researchers to demonstrate research results and share their research findings at a rapid pace than the traditional methods. Today, researchers tend to submit their findings to a wide variety of publication venues such as conferences, journals, and seminars. These publication venues play an important role in the scholarly communication process and the visibility that their work receives. Often researchers might be concerned in knowing about the most important publication venues for publishing their research [1]. However, the selection of publication venues is usually based on the researcher's existing knowledge of the field of his/her discipline [2, 3]. As a result, researchers may not be aware of more appropriate publication venues to which their publications could be submitted.

On the other hand, Computer Science (CS) is a highly active research area that brings together multiple disciplines such as physics, mathematics, and Life Sciences. The number of publication venues has been increasing continuously, making it difficult for researchers to be fully aware about the appropriateness of such publication venues [4]. With an abundance of available publication venues, it becomes a very sdifficult task for new researchers to find exactly what they are looking for or for researchers to keep up to date on all the information [2].

Most of the existing methods to measure the reputation of publication venues use citation count as their chief metric [1]. For journals, among existing methods the most popular citation analysis method is Garfield's Impact Factor (IF) which itself is based on citation counts [5]. The number of citations is not a good individual indicator to measure the quality of publications, since it does not take into account the quality of the citations [6, 7, 8]. In the case of conferences, there are no criteria or consolidated metrics for measuring impact. Unlike some other fields, conferences are essential instruments for the timely dissemination of Computer Science research [9]. As demonstrated in [10], the Computer Science programs follow publication ratio of more than two conference papers per journal paper. In addition, conferences have the precise benefits of giving rapid publication of papers [11]. Therefore, the impact of a publication venue is a key consideration for researchers whether the venue is a journal or a conference [3].

Selecting the most appropriate venue to which to submit a new paper minimizes the risk of publishing in disreputable or fake publication venues. On the other hand, the quality of a publication venue is also important in helping with decisions about awards as well for deciding about scholarships funded by research institutions [12]. If publication venue ranking scores are measured successfully, then researchers can make better decisions about a particular publication venue much quicker based on such a mechanism. There is a significant requirement for an automated process of measuring the publication venue scores to support researchers, so that they can easily recognize the venues in which to publish their research. The findings of this research will definitely be beneficial for the researchers and in return it gives this research a great importance.

In our research, we propose a novel approach for ranking publication venues by considering publication history. We have used a modified version of the PageRank algorithm [13] to generate the scores for publications. We have considered two types of publication venues for which we normally need such information:

1. Known publication venues about which we have historical data

Manuscript received on 23 Nov 2015. Recommended by Prof. K. P. Hewagamage on 16 June 2016.
This paper is an extended version of the paper "Citation Network Based Framework for Ranking Academic Publications and Venues" presented at the ICTer 2015.

Dineshi Peiris holds a B.Sc. (Honours) in Computer Science from the University of Colombo School of Computing, Sri Lanka. (*dineshi.peiris89@gmail.com*).
Dr. Ruvan Weerasinghe is a Senior Lecturer at the University of Colombo School of Computing. (*arw@ucsc.cmb.ac.lk*).

doi: 10.4038/icter.v9i1.7167

- For example, publications of previous conference venues in the series with citation and author data

2. New publication venues about which we have little information

   - For such new conferences, we often only have information about the Program Committee (PC). For new journals, we often only have information about the editorial board.

The paper is organized as follows: first, we briefly describe our data sets. Then the major modules of the conceptual approach - citation network construction, out-links and in-links creation, publication score generation, author score estimation, lower citation counts of recent publications smoothing and publication venue ranking are presented in Section II. The results of our experiments on the real datasets obtained from DBLP and EventSeer.net are presented in Section III. Then a survey of the existing approaches which perform academic publication analysis is conducted. The strengths and weaknesses of these approaches are also given in Section IV. Finally, a conclusion is provided in Section V. Some directions for the future research work are also suggested in.

II. Our Approach

Fig. 1 illustrates the architecture of our proposed citation network-based publication venue ranking approach, which consists of an academic database and six major modules: citation network construction, out-links and in-links creation, publication score generation, author score estimation, lower citation counts of recent publications smoothing and publication venue ranking. First of all, data preparation is discussed in detail. Then we explain the design of each module in our proposed approach.

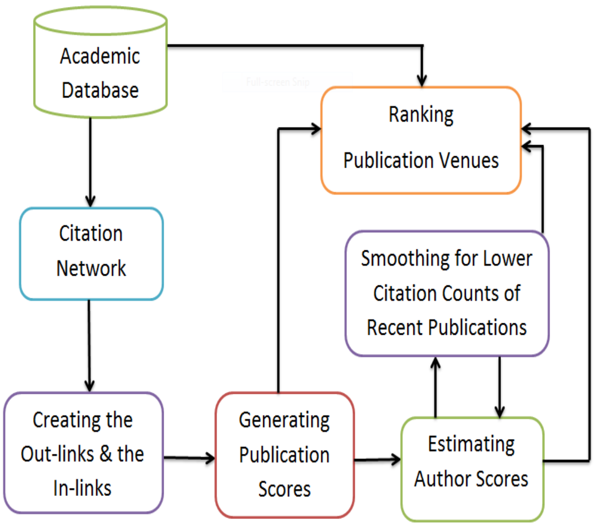

Academic Database
Ranking Publication Venues
Citation Network
Smoothing for Lower Citation Counts of Recent Publications
Creating the Out-links & the In-links
Generating Publication Scores
Estimating Author Scores

Fig. 1. Architecture of the proposed approach

A. Datasets

We had to work with data that we have access to, which generally are the citation data and the PC/Editorial Board data which may be not easy to get directly through sources such as Google Scholar[1]. Besides, there are diverse digital repositories freely available to the general public. DBLP[2], ACM[3], Microsoft Academic Search[4], and CiteSeer[5] digital libraries are vast collections of citations of past publications. DBWorld[6] and EventSeer.net[7] contain most of CFPs for conferences in Computer Science. From these sources we can collect a list of upcoming and past publication venues with the information about topics and organizations among others.

Our approach used data from two primary sources: DBLP and EventSeer.net. These data sources offer different data services: from DBLP we got XML records, while data from EventSeer.net can only be extracted from its website using a HTML parser. DBLP offers XML records for its dataset which can be download from its website. The DBLP dataset contains information about publications from the numerous fields published over the years. This stores a set of metadata for each publication, including publication title, author(s), type of publication, the year of publication and citations. Each publication is represented by the unique key from DBLP. They did a lot of work into resolving the names problem the same person referenced with many names. Because of that the work in this study relied on the DBLP dataset for author and citation data.

EventSeer.net contains most of the Call for Papers (CFPs) for conferences in Computer Science. Therefore this dataset is essential for our work since it is required in ranking upcoming new publication venues. From EventSeer.net, we collected a list of five new conferences with this information from the listed PC members. Summary statistics of the collected data is shown in Table I.

TABLE I.
SUMMARY STATISTICS OF THE COLLECTED DATA

Data	Quantity
Publications	2645295
Unique Authors	1441289
Conferences	3642
Journals	1345
PC members(within five new conferences)	177

B. Citation Network Construction

Citation networks help in evaluating the impact of publication venues, publications and authors [14]. Citation networks are directed networks in which one publication cites another publication. In most cases, authors cite older publications in order to identify the related body of work or to critically analyze earlier work. Hence citation networks are networks of relatedness on subject matter [15]. On the other hand, publications are well defined units of work and accepted papers play an important part in the success of a publication venue [16]. Our approach is completely based on the citation network represented by publications.

We built the citation network defined as a directed graph, with each publication representing a vertex and the citations representing the edges in the graph; the edges being directed ones, directed from the citing vertex to the cited vertex [1].

[1] http://scholar.google.com/

[2] http://www.informatik.uni-trier.de/~ley/db/

[3] http://portal.acm.org/dl.cfm

[4] http://academic.research.microsoft.com/

[5] http://citeseerx.ist.psu.edu/

[6] http://www.cs.wisc.edu/dbworld/

[7] http://eventseer.net/

Each vertex has several attributes, including publication title, conference/journal of publication, year of publication, author(s) and a unique key from the DBLP dataset.

C. The Ranking Method

The method for ranking publications consists of four phases:

- Creating the publication in-links and out-links
- Using a modified version of the iterative PageRank algorithm to calculate the ranking score for each publication
- Estimating ranking scores for authors using the ranking scores of *stable* publications.
- Smoothing for lower citation counts of recent publications

1) Creating the Publication In-links and Out-links

The method for ranking publications based on the citation network uses the two forms of edges: out-links and in-links.

Definition 1. Out-links: *From a given publication p, link all the publications p_i that the publication p cites.*

Definition 2. In-links: *To a given publication p, link all the publications p_j that cite the publication p.*

2) Generating Publication Scores

According to a class of publication-based ranking methods, the graph vertices represent publications, whereas an edge from node n_i to node n_j represents a citation from publication p_i to publication p_j. Computing the ranking at the publication level has the benefit that only a single procedure is performed to evaluate more than one entity: the publication itself, the publication venue it belongs to, and the author(s) [14]. PageRank offers a computationally simple and effective way to assess the relative importance of publications beyond mere citation counts [6]. Unlike other methods, PageRank constructs a single model that integrates both out-links and in-links [17]. PageRank of a publication is defined as follows [18]:

Definition 3. *Assume publication P has publications R1...Rn which point to it. The parameter d is a damping factor which can be set between 0 and 1. People usually set d to 0.85. C(P) is defined as the number of out-links of publication P. The PageRank of a publication P is given as follows:*

$$PR(P) = (1-d) + d\left(\frac{PR(R_1)}{C(R_1)} + .. + \frac{PR(R_n)}{C(R_n)}\right) \quad (1)$$

PageRank is calculated using an iterative algorithm, and corresponds to the eigenvector of the normalized link matrix [18]. However, damping factor allows for personalization and can make it nearly impossible to deliberately mislead the calculations in order to get a higher ranking score [18]. PageRank extends the idea by not counting citations from all publications equally, and by normalizing its rank by the number of citations on a publication [18]. Another important justification is that a publication can have a high PageRank value if there are many publications that point to it, or if there are publications that point to it which themselves have high PageRank values [18]. PageRank handles both these cases by recursively using the link structure of the citation network.

There are many aspects that contribute to the importance of a publication, such as the number of citations it has received, the rating of the citing publications, the time metric of these citations and its author(s). PageRank only includes the first two factors. The publication environment is not static but changes continuously. PageRank favors older publications because older publications have many citations accumulated over time. Bringing emerging publications to researchers are very important since most of them want the latest valuable information. On the other hand, Aditya Pratap Singh et al. [1] have introduced the timing factor in the PageRank algorithm [13] to reduce the bias against more recent publications which have less time than the older publications to get cited. To make the algorithm time-independent, the metric Aditya Pratap Singh et al. [1] proposed to use is the average of the total number of citations of the publications published in each year. We have also modified the formula for calculation of the PageRank of a publication *P*, to make the algorithm time-independent in this way. This Timed PageRank value of a publication *P* is given by the following:

$$yearY = PY[P]$$

$$TPR(P) = (1-d) + \frac{d * \sum \frac{TPR(R_n)}{C(R_n)}}{AYCC[Y]} \quad (2)$$

where *PY[P]* is the year of publication *P*, *TPR(P)* is the Timed PageRank of *P*, *TPR(R_n)* is the Timed PageRank of publication R_n that links to publication *P*, *C(R_n)* is the number of out-links of publication R_n, *AYCC[Y]* is the average number of citations in the year *Y*, and *d* is a damping factor, which is set to 0.85.

3) Alternative Method to Smooth for Lower Citation Counts

Summarizing the weaknesses of the ranking methods we observe that:

- Citation count does not take into account the quality of the citing publications.
- PageRank does not capture the fact that an older publication has more time to be cited in comparison to the recent publications.
- Timed PageRank is able to adjust the rank of emerging quality publications. But it is not sufficient for all the publications since new publications of recent years only have a few or zero citations.

Timed PageRank algorithm is adequate for ranking the publications as it captures the important aspect that an older publication has more time to be cited in comparison to the recent publications. But it is not sufficient for all the publications since new publications of recent years only have a few or no in-links. New publications, which may be of high quality, have a few or no in-links are left behind in this aspect. It is possible the time independent metric of recent publications to become zero.

A study of conferences and journals indicates that many of the references reach back five and more years giving newer publications comparatively little opportunity to get cited [19]. It is possible that the time independent metric of recent publications is zero. Having in mind the above weakness, an alternative method was defined to smooth for lower citation counts of recent publications by modifying the Timed PageRank method.

a) Ranking Scores for Authors

To address the weakness of the Timed PageRank algorithm, we have proposed an alternative metric which uses an author score derived from citations received for publications of that author for previous publications. To assess the quality of a recent publication, its author(s) are useful [20]. It is important to use *stable* publications for calculating scores for authors since we use these author scores for smoothing the lower citation counts of recent publications.

It is to be noted that the DBLP dataset that we have used only has less citation data after the year 1999 (see Table II). Thus we have taken the year 1999 as the *margin year* to demarcate the *stable* publications and the recent publications. An author score is computed by averaging the Timed PageRank values of all the past publications a given author has written till the year 1999.

The equation for the score of an author A_i is:

$$ARS_{A_i} = \frac{\sum TPRS_{P_{A_i}}}{APC[A_i]} \qquad (3)$$

where ARS_{Ai} is the author ranking score, $TPRS_{PAi}$ is the Timed PageRank score of a publication P_{Ai} written by the author A_i and $APC[A_i]$ is the number of publications written by A_i.

TABLE II.
AVERAGE NUMBER OF CITATIONS PER PUBLICATION FROM 1999 TO 2014

Year	Average year citation count	Year	Average year citation count
1999	$1.547473\mathrm{x}10^{-2}$	2007	$1.231\mathrm{x}10^{-5}$
2000	$2.53011\mathrm{x}10^{-3}$	2008	$5.81\mathrm{x}10^{-6}$
2001	$7.080\mathrm{x}10^{-5}$	2009	$1.073\mathrm{x}10^{-5}$
2002	0	2010	$1.541\mathrm{x}10^{-5}$
2003	$1.034\mathrm{x}10^{-5}$	2011	$2.431\mathrm{x}10^{-5}$
2004	$1.752\mathrm{x}10^{-5}$	2012	$4.68\mathrm{x}10^{-6}$
2005	$1.49\mathrm{x}10^{-5}$	2013	0
2006	0	2014	0

b) Smoothing for Lower Citation Counts of Recent Publications

Using these *authoritativ*e scores of authors, we adjust the publication scores after the year 1999. Thus, the score for a new publication is the average score of all the authors of that publication. If this newly calculated publication ranking score is less than the Timed PageRank score of that publication, we will take the Timed PageRank score as the score of the publication.

The equation for the score of a publication P_i is:

$$NPRS_{P_i} = \frac{\sum ARS_{AP_i}}{PAC[P_i]} \qquad (4)$$

where $NPRS_{Pi}$ is the new publication ranking score of lower citation count, ARS_{APi} is the author ranking score of an author A_{Pi} who has written the publication P_i and $PAC[P_i]$ is the number of authors who have written the publication P_i.

D. Ranking Publication Venues

In our Adjusted PageRank method, there are many aspects that contribute to the importance of a publication, including the number of citations it has received, the rating of the citing publications, the time metric of these citations and the authors' prior reputation. Besides, computing the ranking at the publication level has the benefit that only a single procedure is performed to evaluate more than one entity: the publication itself, the publication venue it belongs to, as well as the authors of such publications [14]. Hence we can evaluate publication venues based on this Adjusted PageRank scores.

1) Type I: Generating the Scores for Known Publication Venues

The quality of accepted papers plays an important part in determining the success of a publication venue [16]. The ranking score of a publication venue depends on the quality of research papers it publishes [1]. This is the key behind our approach for ranking known publication venues. We have adjusted the publication ranking scores to deal with Computer Science publication venues. Using the Timed PageRank Scores of the publications and the new publication ranking scores of the publications, we formulate scores for publication venues. Known publication venue scores have been formulated by using the scores of the publications.

The equation for the score of a publication venue V_j is:

$$PVRS_{V_j} = \frac{\sum APRS_{P_{V_j}}}{VPC[V_j]} \qquad (5)$$

where $PVRS_{Vj}$ is the publication venue ranking score, $APRS_{PVj}$ is the adjusted PageRank score of a publication P_{Vj} in the venue V_j and $V\,PC[V_j\,]$ is the venue publication count in V_j.

$APRS_{PVj}$ can be either $TPRS_{PVj}$ Timed PageRank score of a stable publication or $NPRS_{PVj}$ New Publication Ranking Score of lower citation count of a publication.

2) Type II: Generating the Scores for New Publication Venues

Adjusted publication ranking scores are not sufficient for all venues because new venues only have PC/Editorial Board data. Research indicates that the quality of a conference is related to that of its PC members [4]. To assess the importance of a new conference, its PC members are useful. As a proof-of-concept, new publication venue scores are generated only for selected conferences. A recent study of PC candidate recommendation shows that the publication history is the strongest indicator for being invited as PC members [16]. New publication venue ranking is taken care by the scores of PC members which derive from their ranking scores as authors.

The score for a PC member is the author score of this person as an author. Earlier we used the publications the author has written till the year 1999 for calculating the author score. Then we adjusted the publication scores. Now we can calculate the author score of the PC using the Adjusted PageRank scores of each of its members as authors. The score for an author is the average score of the Adjusted PageRank values of the publications the author has written. The ranking score for a new conference is the average score of all the PC members of that conference.

The equation for the score of an author A_i is:

$$ARS_{A_i} = \frac{\sum APRS_{P_{A_i}}}{APC[A_i]} \qquad (6)$$

where ARS_{Ai} is the author ranking score, $APRS_{PAi}$ is the Adjusted PageRank score of a publication P_{Ai} written by the author A_i and $APC[A_i]$ is the number of publications written by A_i.

The equation for the score of a new conference C_j is:

$$NCRS_{C_j} = \frac{\sum ARS_{C_j}}{CPC[C_j]} \qquad (7)$$

where $NCRS_{Cj}$ is the new conference ranking score, ARS_{Cj} is the author ranking score of a PC member in the conference C_j and $CPC[C_j\,]$ is the conference program committee member count in C_j .

III. Experiments and Results

A. Ranking Publications

We carried out our comparative study mainly based on the studies on academic publication analysis [1, 6, 14, 21]. Most of the existing methods use *Citation Count (CC)* to determine the impact of publications [5, 22, 23]. On the other hand, there has been some work done on academic research using the *PageRank (PR)* algorithm [1, 6, 21], which considers the importance of the citing publication to rank the publication being cited. To integrate the time measurement, we have added a timing factor in the PageRank algorithm named *Timed PageRank (TPR)*. Since our approach has been derived through above mentioned methods, we were able to make a comparison between our *Adjusted PageRank (APR)* method and other mentioned methods.

TABLE III.
Ranking methods

Method	Notation
Citation Count	CC
PageRank	PR
Timed PageRank	TPR
Adjusted PageRank	APR

TABLE IV.
Summary of publication ranking methods

Method/Factor	CC	PR	TPR	APR
Number of citations	X	X	X	X
Rating of the citing publications		X	X	X
Time metric			X	X
Smoothing for lower citation counts of recent publications				X

1) Comparison between APR and CC

The following table shows the top 10 publications as determined by our method. Along with the publication APR rank and score, we also show its citation count and its citation rank.

TABLE V.
Top 10 publications in APR method and their citation ranks

Title	APR		CC	
	Rank	*Score*	*Rank*	*Count*
Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total.	1	1	79	90
Implementing Data Cubes Efficiently.	2	0.93830579	71	95
A Relational Model of Data for Large Shared Data Banks.	3	0.88883433	2	580
Mining Association Rules between Sets of Items in Large Databases.	4	0.77657482	45	111
Fast Algorithms for Mining Association Rules in Large Databases.	5	0.71465106	62	100
Object Exchange Across Heterogeneous Information Sources.	6	0.71062948	108	77
The Entity-Relationship Model - Toward a Unified View of Data.	7	0.59574822	1	604
Relational Completeness of Data Base Sublanguages.	8	0.52836066	18	170
Query Evaluation Techniques for Large Databases.	9	0.51691711	70	95
Organization and Maintenance of Large Ordered Indices.	10	0.50069555	21	153

On analyzing the table, the following key observations were made:

- Citation count, the most common measure of publications, is based on mere citation counts that do not account for the quality of the publications where the citations originate. This table illustrates how accounting for citation origin affects the citation ranking of publications.
- Adjusting for citation origin provided a more refined measure of publication status and changed the publication rankings.

2) Comparison between APR and PR

The following table shows the year-wise contribution in the top 100 publications from both the PageRank and the Adjusted PageRank methods.

TABLE VI.
Comparison between PageRank and Adjusted PageRank method distributions of the top 100 publications

Year	PR	APR	Year	PR	APR
1965	0	1	1992	1	5
1970	1	1	1993	1	9
1971	3	2	1994	1	7
1972	2	2	1995	0	21

doi: 10.4038/icter.v9i1.7167

Year	PR	APR	Year	PR	APR
1973	0	0	1996	2	13
1974	4	1	1997	0	10
1975	13	0	1998	0	0
1976	9	2	1999	0	1
1977	11	1	2000	0	0
1978	6	1	2001	0	1
1979	9	1	2002	0	1
1980	2	0	2003	0	1
1981	8	0	2004	0	1
1982	4	0	2005	0	2
1983	3	0	2006	0	1
1984	6	2	2007	0	0
1985	1	0	2008	0	0
1986	5	0	2009	0	0
1987	5	0	2010	0	1
1988	0	0	2011	0	4
1989	1	1	2012	0	2
1990	2	2	2013	0	1
1991	0	0	2014	0	2

Fig. 2 shows the variation of the number of publications in the top 100 in both the APR and the PR methods over the years spanning from 1965 to 2014.

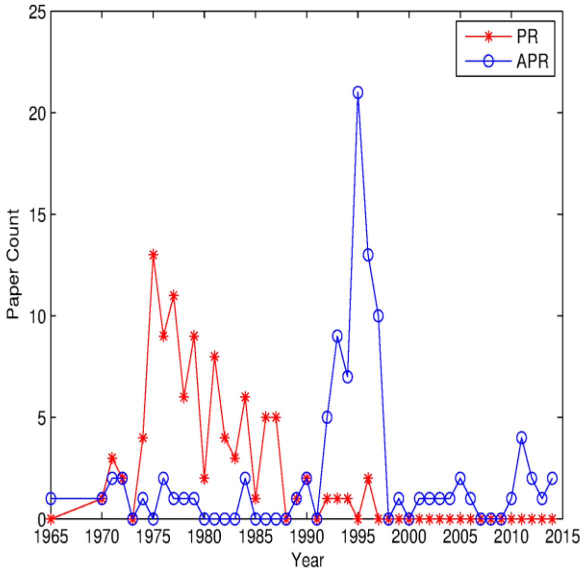


Fig. 2. The number of publications distributed over the years in Adjusted PageRank and PageRank methods

On analyzing the graph, the following key observations were made:

- The top publications in the PR are mostly from 1970s and 1980s whereas in the APR, the top publications are mostly from 1990s and 2000s. This shows that PR favors older publications because older publications have many citations accumulated over time.
- This shows that our method reduces the bias against the recent publications which have less time than older publications to get referenced. Hence it is able to adjust the rank of emerging quality publications.

3) *Comparison between APR and TPR*

The following table shows the year-wise contribution in the top 100 publications from both the Timed PageRank and the Adjusted PageRank methods.

TABLE VII.
COMPARISON BETWEEN TIMED PAGERANK AND ADJUSTED PAGERANK METHOD DISTRIBUTIONS OF THE TOP 100 PUBLICATIONS

Year	TPR	APR	Year	TPR	APR
1965	1	1	1992	7	5
1970	1	1	1993	10	9
1971	2	2	1994	11	7
1972	2	2	1995	24	21
1973	1	0	1996	14	13
1974	2	1	1997	10	10
1975	0	0	1998	0	0
1976	4	2	1999	1	1
1977	1	1	2000	0	0
1978	1	1	2001	0	1
1979	1	1	2002	0	1
1980	0	0	2003	0	1
1981	1	0	2004	0	1
1982	0	0	2005	0	2
1983	0	0	2006	0	1
1984	2	2	2007	0	0
1985	0	0	2008	0	0
1986	0	0	2009	0	0
1987	0	0	2010	0	1
1988	0	0	2011	0	4
1989	1	1	2012	0	2
1990	2	2	2013	0	1
1991	1	0	2014	0	2

Fig. 3 shows the variation of the number of publications in the top 100 in both the APR and the TPR methods over the years spanning from 1965 to 2014.

doi: 10.4038/icter.v9i1.7167

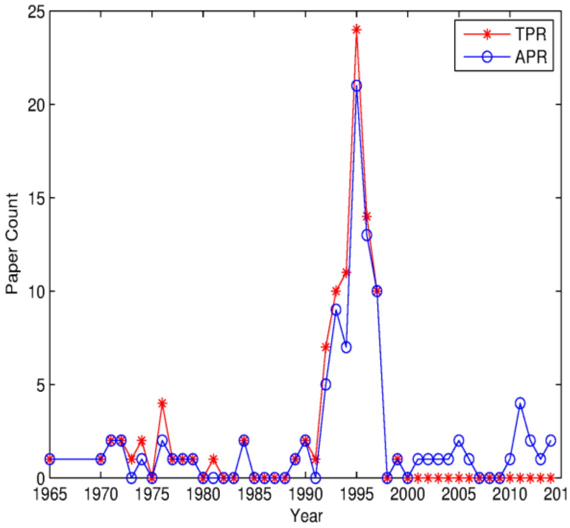


Fig. 3. The number of publications distributed over the years in Adjusted PageRank and Timed PageRank methods

On analyzing the graph, the following key observations were made:

- In the APR method, the publications are distributed over the years as compared to that in the TPR method.
- TPR is able to adjust the rank of emerging quality publications. But it is not sufficient for recent publications which only have a few or zero citations (after 1999). This is clearly visible in the graph as the TPR method is not able to assess the importance of recent publications whereas the APR method is able to assess the importance of recent publications based on their authors.
- To assess the importance of a recent publication, its authors are useful.

For better analysis, we selected few young publications for which citation statistics are not readily available in our dataset, and analyzed them by using their normalized ranking scores in the Time PageRank and the Adjusted PageRank methods over the recent years as shown in Table VIII.

TABLE VIII.
NORMALIZED PUBLICATION SCORES IN THE ADJUSTED PAGERANK AND THE TIMED PAGERANK METHODS

Year	Title	Ranking Score	
		TPR Score	APR Score
2006	A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations.	0.06421025	0.34819944
2007	Distributed Resource Management and Admission Control of Stream Processing Systems with Max Utility.	0.06421025	0.20661683
2008	A low-power RF front-end of passive UHF RFID transponders.	0.06421025	0.20620484
2009	A revised r*-tree in comparison with related index structures.	0.06421025	0.24453887
2010	Fair power control for wireless ad hoc networks using game theory with pricing scheme	0.06421025	0.24960894
2011	Permission Re-Delegation: Attacks and Defenses.	0.06421025	0.19908869
2012	Anatomy of a gift recommendation engine powered by social media.	0.06421025	0.24091758
2012	Clickjacking: Attacks and Defenses.	0.06421025	0.19908869
2013	Optimizing budget constrained spend in search advertising.	0.06421025	0.17896617
2014	Thermal design and simulation of automotive headlamps using white LEDs.	0.06421025	0.20620484

On analyzing the table, the following key observations were made:

- Every publication has a PageRank of 0.15 value even though no-one is referencing for it. In the Timed PageRank method, every publication has a ranking score of 0.06420859 when normalizing the 0.15 to scale down the value within the range (0, 1).
- Recent publications, which may be of high quality, have no in-links climbed up their ranking scores when switched from Timed PageRank to Adjusted PageRank method.
- This table shows that our method reduces the bias against the recent publications, which have no in-links.

B. Ranking Publication Venues

1) Type I: Known Publication Venues

We relied on our publication ranking method to compute publication venue ranking scores. One approach could be to compute the average score of all their publications. The following table shows the top 10 publication venues by averaging of all their publications.

TABLE IX.
TOP 10 PUBLICATION VENUES BY AVERAGING OF ALL THEIR PUBLICATIONS. **TYPE**: WHETHER THE VENUE IS A JOURNAL (J) OR A CONFERENCE (C)

Publication Venue	Type	Score
IPSJ	C	0.28796447
Electronic Networking: Research, Applications and Policy	J	0.16220377
VLDB workshop on Management of Uncertain Data (MUD)	C	0.08146009
Foundations and Trends in Databases (FTDB)	J	0.08097285
ACM Trans. Database Syst. (TODS)	J	0.08017983
Conference on Very Large Data Bases (VLDB)	C	0.07990088
ACM SIGMOD International Conference on Management of Data (SIGMOD)	C	0.07961964

doi: 10.4038/icter.v9i1.7167

Publication Venue	Type	Score
Science	J	0.07911258
VIEWS	C	0.07818149
Performance and Evaluation of DataManagement Systems (ExpDB)	C	0.07746338

For instance, publication venue A has 30 publications with only 20 being top ranking publications. Assume that these high quality publications have a score of 10 points each, where the remaining ones have a score of 1 point. Publication venue B has in total 5 publications, with 4 publications of them being top ranking publications. It is reasonable to consider that publication venue A should be ranked higher than publication venue B for their scientific contribution, because A has 5 times the number of top ranking publications than publication venue B. If we compute the average of all publication scores, then publication venues A and B would have 7 and 8.2 points respectively. It is not fair to take that approach to compute venue scores.

In order to deal with this problem, we have taken into account the top *n%* of publications to calculate publication venue score. Therefore, our problem was to choose the *n%* of publications of each publication venue that should be considered in the ranking. We performed the following experiment to determine the number *n*. We computed the average score for each publication venue by using their top *n%* publications, $\forall\ n \in \{25, 50, 75\}$. Thus, we produced 3 ranking lists for our publication venue ranking task. As a test bed we used the CORE 2013 Conference Ranking list[8]. In CORE conference ranking, conferences are allocated a rank of A*[9], A[10], B[11] or C[12]. The ratios of A* and A conferences within the top 10 publication venues were calculated, the better the evaluation was considered as the publication venue ranking list. It is to be noted that we have only considered the conferences within the top 10 publication venues to compute the ratio.

The following tables show the top 10 publication venues by averaging the top 25%, 50%, and 75% of publications respectively. Along with our publication venue rank and score, we also show its CORE 2013 Ranking.

TABLE X.
TOP 10 PUBLICATION VENUES BY AVERAGING THE TOP 25% OF PUBLICATIONS.
TYPE: WHETHER THE VENUE IS A JOURNAL (J) OR A CONFERENCE (C)

Publication Venue	Type	Score	CORE Ranking
Foundations and Trends in Databases (FTDB)	J	0.12188608	-
VIEWS	C	0.11511190	-
ACM SIGMOD International Conference on Management of Data (SIGMOD)	C	0.11356147	A*
VLDB workshop on Management of Uncertain Data (MUD)	C	0.11322808	-
Conference on Very Large Data Bases (VLDB)	C	0.11307256	A*
ACM Trans. Database Syst. (TODS)	J	0.11122050	-
ACM SIGMOD Digital Symposium Collection (DISC)	J	0.10751429	-
Performance and Evaluation of Data Management Systems (ExpDB)	C	0.10126516	-
Conference on Parallel and Distributed Information Systems (PDIS)	C	0.10032141	C
Conference on Innovative Data Systems Research (CIDR)	C	0.09893309	A

TABLE XI.
TOP 10 PUBLICATION VENUES BY AVERAGING THE TOP 50% OF PUBLICATIONS.
TYPE: WHETHER THE VENUE IS A JOURNAL (J) OR A CONFERENCE (C)

Publication Venue	Type	Score	CORE Ranking
Foundations and Trends in Databases (FTDB)	J	0.09897220	-
VLDB workshop on Management of Uncertain Data (MUD)	C	0.09815774	-
ACM SIGMOD International Conference on Management of Data (SIGMOD)	C	0.09391697	A*
Conference on Very Large Data Bases (VLDB)	C	09386330	A*
ACM Trans. Database Syst. (TODS)	J	0.09375841	-
VIEWS	C	0.09076796	-
Performance and Evaluation of Data Management Systems (ExpDB)	C	0.09055314	-
Conference on Innovative Data Systems Research (CIDR)	C	0.08719047	A
ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)	C	0.08603544	A*
ACM SIGMOD Digital Symposium Collection (DISC)	J	0.08586227	-

TABLE XII.
TOP 10 PUBLICATION VENUES BY AVERAGING THE TOP 75% OF PUBLICATIONS.
TYPE: WHETHER THE VENUE IS A JOURNAL (J) OR A CONFERENCE (C)

Publication Venue	Type	Score	CORE Ranking
VLDB workshop on Management of Uncertain Data (MUD)	C	0.08721005	-
Foundations and Trends in Databases (FTDB)	J	0.08706834	-
ACM Trans. Database Syst. (TODS)	J	0.08532985	-
Conference on Very Large Data Bases (VLDB)	C	0.08508535	A*

[8] http://www.core.edu.au/

[9] flagship conference

[10] excellent conference

[11] good conference

[12] other ranked conference venues

Publication Venue	Type	Score	CORE Ranking
ACM SIGMOD International Conference on Management of Data (SIGMOD)	C	0.08475611	A*
Performance and Evaluation of Data Management Systems (ExpDB)	C	0.08391501	-
VIEWS	C	0.08360480	-
Conference on Innovative Data Systems Research (CIDR)	C	0.08026346	A
Workshop on Data Management on New Hardware (DaMoN)	C	0.07970309	-
ACM SIGACT-SIGMOD Symposium on Principles of Database Systems (PODS)	C	0.07965752	A*

According to the tables, Table XIII shows the ratios of A* and A conferences within the top 10 venues by averaging the top 25%, 50% and 75% of publications respectively. Based on this experiment, we concluded that the average of top 50% publications is the most appropriate publication venue ranking list.

The equation for the ratio is:

$$Ratio = \frac{X}{Y} \tag{8}$$

where X is the number of A* and A conferences within the top 10 publication venues, and Y is the number of conferences within the top 10 publication venues

TABLE XIII.
THE RATIO OF A* AND A CONFERENCES WITHIN THE TOP 10 PUBLICATION VENUES

	25%	50%	75%
Number of conferences within the top 10 publication venues	7	7	8
Number of A* and A conferences within the top 10 publication venues	3	4	4
Ratio	0.4286	0.5714	0.5

Furthermore, we produce a venue ranking list by using a cut-off of 50 publications to indicate statistical significance. The following table shows the top 10 publication venues which have higher than 50 publications.

TABLE XIV.
TOP 10 PUBLICATION VENUES WHICH HAVE HIGHER THAN 50 PUBLICATIONS
TYPE: WHETHER THE VENUE IS A JOURNAL (J) OR A CONFERENCE (C)

Publication Venue	Type	Score	CORE Ranking
ACM Trans. Database Syst.	J	0.08017982	-
Conference on Very Large Data Bases (VLDB)	C	0.07990088	A*
ACM SIGMOD International Conference on Management of Data (SIGMOD)	C	0.07961964	A*
Conference on Innovative Data Systems Research (CIDR)	C	0.07623364	A
ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)	C	0.07582618	A*
Parallel and Distributed Information Systems (PDIS)	C	0.07480646	C
Journal on Very Large Data Bases (VLDB J.)	J	0.07449637	-
International Workshop on the Web and Databases (WebDB)	C	0.07401554	C
International Conference on Database Theory (ICDT)	C	0.07394727	A
International Conference on Data Engineering (ICDE)	C	0.07238653	A*

2) Type II: New Publication Venues

As a proof-of-concept, new publication venue scores were generated only for following five conferences. Among recent CFPs we have only taken conferences which were stated as their first conference. The following table shows the selected new conference venue details along with their ranking scores.

TABLE XV.
NEW CONFERENCE VENUES AND CORRESPONDING RANKING SCORES

Conference Venue	Year	Score
1st International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM)	2015	0.15460470
1st IEEE International Conference on Multimedia Big Data (BIGMM)	2015	0.15156164
1st Biomedical Linked Annotation Hackathon (BLAH)	2015	0.15059667
1st International Conference on Fundamentals and Advances in Software Systems Integration (FASSI)	2015	0.15053273
1st International Conference on Decision Support System Technology (ICDSST)	2015	0.15007116

IV. RELATED WORK

There has been considerable work in the field of academic research. Among existing methods, the most widely adopted method for measuring the quality of publication venues is to use Garfield's IF. This metric uses the publication citations from only the last two years, which neglects the importance of older papers that they cite. On the other hand, it has been criticized for its only dependency on citation counts [7]. As a result, many alternative methods, e.g., h-index [22], g-index [23], and PageRank algorithm [13], have been used to rank venues [24].

Most research work on academic publications uses citation count as the metric. However, metrics like IF, h-index and g-index are based on the citation count, and hence would not give accurate results in all scenarios [1]. The number of citations is not a good individual indicator to measure quality of publications, since it does not calculate the importance of the quality of citations [6]. It is important to look at a metric which considers the importance of the citing publications to rank the publication being cited.

There has been much interest in applying social network-based methods for generating recommendation and measuring conference quality. A recommender system for academic events and scientific communities based on Social Network Analysis (SNA) is presented in [25]. This work regards on co-authorship and citation networks. The system constructs an academic event participating matrix, based on which similarity between any two researchers is computed. To make recommendations to a target researcher, a group of the most similar researchers is first selected and then the rank of upcoming events is determined by their aggregating ratings.

Zhuang et al. [4] have identified a set of heuristics to automatically determine the quality of the conferences based on characteristics of the PC. This research is completely based on a hypothesis, where the quality of a conference is closely correlated to the reputation of its PC members. The study was unique in the way the authors have brought their views. The heuristics both in combination and isolation have been examined under a classification scheme. In [4], when combined under this scheme, these proposed heuristics achieved a satisfying accuracy in differentiating conferences. These heuristics are also used to rank and recommend conferences. The proposed heuristics rely on the completeness of the list of the PC members. One issue is that a small number of CFPs do not have an entire list of PC members.

There has been some work done on academic research using the PageRank algorithm [1, 6, 21]. Ding et al. [6] used PageRank to rank authors based on the co-citation network. The closest to our work is research work in [1] which uses an efficient approach to rank the papers in various conferences. A modified version of PageRank has been used to rank papers as well as conferences. An important metric in the algorithm which takes the time factor in ranking the papers has been introduced to minimize the bias against new papers which get little time for being cited. Using the year of publication of the papers, the year-wise score for each conference venue has been calculated.

However, the timing factor is not sufficient for all the publications since new publications only have a few or zero citations. Another issue is how to estimate scores for new venues for which citation data are not available using this method. Our work is motivated by this work and takes two further steps. To address the weakness of this method, we have proposed an alternative metric which uses an author score derived from citations received for previous publications of the author. We have also introduced a new way to assess the importance of old and new publication venues.

V. Conclusion and Future Work

We proposed a novel approach for ranking publication venues by considering publication history. The Timed PageRank algorithm is not sufficient for all the publications since new publications of recent years only have a few citations. New publications, which may be of high quality and have a few citations, are left behind in this aspect. To assess the relative importance of recent publications, we have adjusted the Timed PageRank values with its authors' past publication scores. In our approach, there are many aspects that contributed to the importance of a publication, including the number of citations it has received, the rating of the citing publications, the time metric and its authors' reputations. The experimental results indicate that our method reduces the bias against more recent publications, which only have a few citations. The researchers can make better decisions about a particular venue much quicker and easier based on this mechanism.

The DBLP dataset that we have used only have a few or no citation data after the year 1999. Thus we have taken the year 1999 as the margin year to separate the *stable* publications and the new publications. There is definitely room for improvement on the margin year. The proposed margin year relies on the completeness of the citation data. One issue is that our database does not have a complete list of citations. For example, a quality publication may get a lot of citations from scientific domains that are not included in the DBLP dataset. In such cases, it requires further action to harvest citation data before the proposed approach can be applied.

The ranking scores for authors were derived from the publication ranking scores till the year 1999 only. Using the scores of authors, the lower citation counts of recent publications were adjusted by calculating an average score for each publication after the year 1999. The score for a lower citation count publication was taken as the average score of all the authors who have written that research paper. If there was no author score for a particular author, then we would have ignored that author score and take the average score of other authors. On the other hand, if there were no author scores for all the authors of a particular paper then solution would not have been given. Thus we have taken previously measured Timed PageRank value as the score of the publication. Our smoothing method relies on the generated scores of the authors. In such cases, as mentioned earlier, it requires further action to harvest the citation data as well as author data before proposed approach can be applied.

Currently, five CFPs from EventSeer.net were imported into our database. In the future, it would be of interest to add other CFPs for venue ranking problem. EventSeer.net does not offer a structured dataset like that of the DBLP dataset; we have to parse its website to extract the relevant information. Regular expressions could be used to process aspects of the CFPs text. DBLP XML records and EventSeer.net need to be combined in one unique dataset. The problem is to connect these two data sources to provide a unique data repository for publications. Regular expression could be used to match authors' names and join PC members' names in EventSeer.net to DBLP dataset. Various data refining techniques could be applied to make the analysis more precise.

Some other data sources like Google Scholar, CiteSeer and ACM could be integrated into our data repository to make it more complete. Currently, data from DBLP and EventSeer.net is imported into our database. To have better ranking results, we need data from other sources. Publication data gathered from the web by a web crawler is also an interesting development direction.

Acknowledgment

We would like to thank all the staff members of the Network Operating Center of the UCSC for providing a virtual server instance to facilitate our research activities.

References

[1] Singh A. P., Shubhankar K. and Pudi V. (2011). An efficient algorithm for ranking research papers based on citation network. *Data Mining and Optimization (DMO), 2011 3rd Conference on*. IEEE, pp. 88-95.

[2] Luong H., Huynh T., Gauch S., Do P. and Hoang K. (2012). Publication venue recommendation using author network's publication history. *Intelligent Information and Database Systems*. Springer, pp. 426-435.

[3] Luong H. P., Huynh T., Gauch S. and Hoang K. (2012). Exploiting social networks for publication venue recommendations. *KDIR*, pp.239-245.

[4] Zhuang Z., Elmacioglu E., Lee D. and Giles C. L. (2007). Measuring conference quality by mining program committee characteristics. *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM, pp. 225-234.

[5] Garfield E. (1999). Journal impact factor: a brief review. *Canadian Medical Association Journal*, 161(8): 979-980.

[6] Ding Y., Yan E., Frazho A. and Caverlee J. (2009). Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11): 2229-2243.

[7] Saha S., Saint S. and Christakis D. A. (2003). Impact factor: a valid measure of journal quality? *Journal of the Medical Library Association*, 91(1): 42.

[8] Seglen P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314(7079): 497.

[9] Patterson D. A. (2004). The health of research conferences and the dearth of big idea papers. *Communications of the ACM*, 47(12): 23-24.

[10] Laender A. H. F., de Lucena C. J. P., Maldonado J. C., de Souza e Silva E. and Ziviani N. (2008). Assessing the research and education quality of the top brazilian computer science graduate programs. *ACM SIGCSE Bulletin*, 40(2): 135-145.

[11] Franceschet M. (2010). The role of conference publications in cs. *Communications of the ACM*, 53(12): 129-132.

[12] Martins W. S., Goncalves M. A., Laender A. H. and Pappa G. L. (2009). Learning to assess the quality of scientific conferences: a case study in computer science. *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. ACM, pp. 193-202.

[13] Page L., Brin S., Motwani R. and Winograd T. (1999). The pagerank citation ranking: Bringing order to the web. Stanford InfoLab, Tech. Rep.

[14] Sidiropoulos A. and Manolopoulos Y. (2006). Generalized comparison of graph-based ranking algorithms for publications and authors. *Journal of Systems and Software*, 79(12): 1679-1700.

[15] Valmarska A. (2014). *Analysis of citation networks*. Diploma Thesis, Faculty of Computer and Information Science, University of Ljubljana.

[16] Han S., Jiang J., Yue Z. and He D. (2013). Recommending program committee candidates for academic conferences. *Proceedings of the 2013 workshop on Computational scientometrics*. ACM, pp. 1-6.

[17] Mihalcea R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, pp. 20-23.

[18] Brin S. and Page L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1): 107-117.

[19] Rahm E. and Thor A. (2005). Citation analysis of database publications. *ACM Sigmod Record*, 34(4): 48-53.

[20] Yu P. S., Li X. and Liu B. (2004). On the temporal dimension of search. *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. ACM, pp. 448-449.

[21] Dellavalle R. P., Schilling L. M., Rodriguez M. A., Van de Sompel H., and Bollen J. (2007). Refining dermatology journal impact factors using pagerank. *Journal of the American Academy of Dermatology*, 57(1): 116-119.

[22] Hirsch J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46): 16 569-16 572.

[23] Egghe L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1): 131-152.

[24] Katerattanakul P., Han B. and Hong S. (2003). Objective quality ranking of computing journals. *Communications of the ACM*, 46(10): 111-114.

[25] Klamma R., Cuong P. M. and Cao Y. (2009). You never walk alone: Recommending academic events based on social network analysis. *Complex Sciences*. Springer, pp. 657-670.

Path Index Based Keywords to SPARQL Query Transformation for Semantic Data Federations

Thilini Cooray, Gihan Wikramanayake

***Abstract*— Semantic web is an emerging research domain. Enhancing the ability of keyword query processing on Semantic Web data provides a huge support for familiarizing the usefulness of Semantic Web to the general public. Most of the existing approaches focus on just user keyword matching to RDF graphs and output the connecting elements as results. Semantic Web consists of SPARQL query language which can process queries more accurately and efficiently than general keyword matching. There are only about a couple of approaches available for transforming keyword queries to SPARQL. They basically rely on real time graph traversals for identifying sub-graphs which can connect user keywords. Those approaches are either limited to query processing on a single data store or a set of interlinked data sets. They have not focused on query processing on a federation of independent data sets which belongs to the same domain. This research proposes a Path Index based approach eliminating real time graph traversal for transforming keyword queries to SPARQL. We have introduced an ontology alignment based approach for keyword query transforming on a federation of RDF data stored using multiple heterogeneous vocabularies. Evaluation shows that the proposed approach has the ability to generate SPARQL queries which can provide highly relevant results for user keyword queries. The Path Index based query transformation approach has also achieved high efficiency compared to the existing approach.**

***Keywords*— Semantic Web, Keyword query processing, SPARQL query generation, RDF Federations**

I. Introduction

Nowadays the World Wide Web (WWW) has become essential to everyone. People always tend to search the web to retrieve information about almost everything. Once enters a query, the underlying query processors must be able to gather results from available sources. What user is interested is, receiving relevant answers for their questions, efficiently.

Ability to understand the meaning of user query is important to provide relevant results. Once the user requirement is understood, it should be presented in a way which underlying data sources can understand and process. The relevancy of results provided by the data source depends on both the completeness of data stored in the source and how well the user query is understood by the data source.

The web contains huge amount of details about variety of topics. Most of them are stored as web documents. Web documents are capable of preserving complete details about topics rather than compacting them to traditional databases where only the details matches with the database schema are stored while skipping others despite their necessity for the completeness of information. However, as the amount of web documents are extremely increasing, requirement for effective storage mechanisms and efficient searching mechanisms were highly demanded. This paved way to the emergence of the concept of transforming web documents to web data.

Semantic Web[1] was introduced as a method of storing web data in such a manner which is understandable to a computer. . Resource Description Framework[2] (RDF) was presented as the standard format for storing and exchanging data. RDF preserves the interconnections among data elements and use graph structures for storage. Using graph structures for data storing, is crucial for web data as they contain huge amounts of relationships that relational databases are incapable of maintaining. These relationships are essential when recognizing the relevancy of data for a user query. Recently, many researchers and academic institutes have taken the initiative of exposing their data to the Web in RDF format. SPARQL[3] is the query language for RDF data. It is capable of both representing information needs along with relationships among elements and dive in RDF sources to extract information considering those relationships.

Relaxed models such as keyword queries are convenient for general users to query data sources as they do not have to consider the underlying complexity such as data structures and schema when composing queries.

Many researches have been carried out in keyword query search over tree [1], [2], [3], [4] and graph [5], [6] structured data. Basic idea behind keyword search is to identify matching data elements for keywords from the underlying data source and retrieve substructures which connect all those identified elements.

Structured queries are capable of retrieving more relevant results efficiently and accurately compared to keyword queries. However, composing structured queries require expertise knowledge which is lacking among general users. SPARQL is capable of retrieving more relevant results from web data. Therefore, bridging the gap between user friendly keyword queries and SPARQL allows general users to retrieve highly relevant results without having knowledge about underlying complexities. Transforming keyword queries to SPARQL is still a novel topic which has not gained much attention among semantic research attention. However, it could be identified as one of the key points in

Manuscript received on 23 Nov 2015. Recommended by Prof. K. P. Hewagamage on 16 June 2016.
This paper is an extended version of the paper “An Approach for Transforming Keyword-Based Queries to SPARQL on RDF Data Source Federations” presented at the ICTer 2015.

Thilini Cooray holds a B.Sc. (Honours) in Computer Science from the University of Colombo School of Computing, Sri Lanka. (*thilinicooray.ucsc@gmail.com*).
Prof. Gihan Wikramanayake is a Professor at the University of Colombo School of Computing. (*gnw@ucsc.cmb.ac.lk*).

[1] http://en.wikipedia.org/wiki/Semantic_Web
[2] http://www.w3.org/RDF
[3] http://www.w3.org/TR/rdf-SPARQL-query

doi: 10.4038/icter.v9i1.7168

familiarizing the importance of Semantic web to general public and sharing its privileges with them for fulfilling their information needs efficiently while enhancing the relevancy of results.

The process of translating keyword queries to SPARQL can be decomposed to following steps. 1) Mapping user keywords to data elements. 2) Identifying sub-graphs which can connect mapped data elements. 3) Generating queries based on the relationships in the sub-graphs. Most of the available approaches exploit graph traversal in real time for identifying suitable sub-graphs [3], [4], [6], [8], [9]. Only limited set of functions are carried out as preprocessing. Most of these approaches provide approximate results because traversing RDF graphs with millions or billions of data is very expensive and highly time consuming. Hence there is a requirement to seek for approaches which can reduce graph traversal in query generating time.

There are many contributors of the WWW who provides information on related topics individually. For an example, DBLP and ACM contain academic publication data individually. None of them have entire publication data. In contrast, Google Scholar connects sources such as DBLP and ACM to provide more complete set of results for the public. Therefore, general public is attracted more towards Google scholar for their publication related information needs. Most RDF sources are also maintained as individual dumps. In order to provide more complete results with high accuracy, it is important to combine those together. RDF federations have been presented as a solution for this problem. Yet existing RDF federations only accept SPARQL queries. Seeking for approaches which can direct user queries to RDF federations will enhance completeness and accuracy of provided results. This research focuses on transforming keyword queries to SPARQL on a RDF federation in order to allow general users to access Semantic Web and fulfill their information needs.

Following are the main contributions of this research:

- **Proposing an approach to map user keywords to data elements resolving vocabulary level heterogeneity -** An available ontology alignment mechanism is utilized for resolving vocabulary level heterogeneity. Results of the alignment mechanism are combined with a keyword index to map source wise matching elements for user keywords. This mechanism is capable of returning a set of keyword matching elements for each data source in the federation.

- **Building a Path Index capturing full paths accurately -** Path Index is an existing concept which reduces the cost of real time graph traversing for keyword query processing. The existing logic intends to store full paths from vertices to sink nodes as a preprocessing task. However, the breadth-first search based algorithm presented is unable to filter only full paths, causing unnecessary graph traversals at real time. Therefore, this research has proposed a depth-first search based approach which is capable of accurately capturing full paths.

- **Utilizing the stored templates in the Path Index to generate SPARQL queries without graph traversing in real time -** Path Index was previously used only for keyword mapping. This research proposes a way which Path Index can be utilized for SPARQL query generation. The proposed approach is capable of generating queries which can be directly executed on SPARQL query engines. Results of generated queries exhibit high precision and recall values.

The paper is organized as follows: Section 2 discusses related work; Section 3 gives an overview of the methodology, Section 4, 5 and 6 gives an in-depth explanation about each component of the suggested approach; finally Section 7 provides experimental evaluations and in Section 8, conclusions and future work are presented.

II. Related Work

Even though there is no existing approach for transforming keyword queries to SPARQL on RDF federations, research has taken place in addressing each step required for keyword to SPARQL transformation on RDF federations. They are as follows; *resolving vocabulary level heterogeneity*, *mapping user keywords to data source elements* and *identifying suitable sub-graphs connecting keyword elements*.

Heterogeneity resolution is a core function of federations according to Sheth and Larson [16]. There are different levels of heterogeneity such as vocabulary level heterogeneity and data level. However, vocabulary level heterogeneity must be resolved to process keyword based queries in a federation. SPARQL query rewriting is one approach for resolving vocabulary level heterogeneity [17], [18], [19]. The input query must be written in SPARQL using a specific vocabulary for applying this solution. This cannot be applied when input query is in keyword.

Ontology alignment is another method proposed for heterogeneity resolution among RDF data sources. Concept level, property level and instance level are the ontology alignment types according to Gunaratna et al. [20]. BLOOMS [13], Aroma [21] and RiMOM [14] are some concept level alignment approaches. Gunaratna et al. [20] have mentioned that very less amount of research have been focused on property level alignments. Alignment API [12] can be identified as a proper tool for property level alignments from different vocabularies. Since this research aims at heterogeneity resolution for a federation, both concept level and property level vocabulary resolutions are required.

Indexing is the common approach adopted by almost the entire keyword query processing approaches for matching data elements with keywords. However types of indices they have used are different. SearchWeb [8], Bidirectional Search [6] keeps indices for both vertices and edges. They consider the possibility of a user keyword occurs at an edge as well as at vertices. BLINKS [5] believes that keywords can only occur in vertices of the graph so index only vertices. Path Index [7] only store sink nodes in their index arguing that keywords only reside in sink nodes. When storing details, most approaches index only the label of the graph element. However SearchWeb [8] and Hermes [15] store some

doi: 10.4038/icter.v9i1.7168

additional information along with the label. Those details are used for efficient sub-graph identification on that approach.

Many approaches have been suggested for identifying suitable sub-graphs which can connect keyword elements. Basic tree search algorithms such as Breadth First Search [22] and Depth First Search [23] were first applied for substructure identification in tree structured data. Then several other algorithms [24], [25] were proposed by modifying those basic concepts. As RDF data sets mostly have a graph structure, graph exploration approaches were proposed such as Backward Search [9] and Bidirectional algorithm [6]. SearchWeb [8] has suggested a summarized graph which has reduced the size of graph which needs to be traversed at real time for finding suitable sub-graphs.

Real time graph traversing is highly time consuming. M-KS [26] uses a matrix to store the keyword relationships to eliminate graph traversal at real time. They only focus on binary relationships. G-KS [27] proposes a keyword relationship graph to find a suitable sub-graph to resolve the weakness of M-KS. Tran et al. [10] suggest a graph based model to sub-graph identification. Cappellari et al. [7] suggest a path index based approach, which stores edge sequences from source nodes to sink nodes of RDF graphs. As they have totally eliminated graph traversing at real time, efficiency of this approach is really high. However they require more storage space than other approaches.

Among keyword searching approaches, only three methods have been proposed for transforming keyword queries to SPARQL. SearchWeb [8] and Hermes [15] have proposed a method for converting keyword queries to SPARQL by identifying suitable sub-graph and converting it to conjunctive queries. They have only proposed that approach either to a single data source or a set of linked data sources. They have not considered the federation scenario. Unger et al. [28] have suggested a linguistic analysis based approach for transforming natural language queries to SPARQL. They have ignored the capabilities of Semantic Web in their solution.

III. Methodology

We propose an approach for transforming keyword queries to SPARQL on RDF data source federations within a set of defined limitations. The main objective of our research is to examine the feasibility of proposed approach for keyword to SPARQL transformation on federations as no existing approach has addressed this issue. We do not focus on examining the generalizability and scalability of this approach at this initial stage.

We use academic publication data for explaining and evaluating this approach at the initial phase even though this proposed approach can be used domain independently. Author name, published year and publication title are the initial set of keyword fields we are using for generating keyword queries. These fields were selected as those are the fields which are highly queried regarding academic publications even in digital libraries such as ACM and DBLP. We have defined a specified format of queries to this approach. All conditions of the user query should be represented according to the format <field name>:<field value>. Comma should be used if multiple conditions are presented. Field which needs results should be indicated using a question mark (?). For an example, accepted keyword query for “What are the publications of James published in 1995?” is “publication:?,author:James,year:1995”.

RDF data are stored in graph structures. Therefore cycles can occur. Inverse relationships can be identified as a main reason for causing cycles in RDF graphs. This research only aims at resolving cycles caused by inverse relationships on RDF graphs.

Heterogeneity resolving is a main focus in federations. Vocabulary level heterogeneity and instance level heterogeneity are the main levels of heterogeneity in RDF federations. Approaches for dealing with those heterogeneous situations are needed to be included in the architecture. Resolving vocabulary level heterogeneity in a federation is essential for retrieving complete results, because different vocabularies usually use different terminologies for similar concepts. Identifying the similarity among vocabularies paves the way for extracting relevant results from heterogeneous data sources. When consider instance level heterogeneity, it contains a separate level of complexity. Different data sets may store same value in different formats. For an example, an author name "A.Bernstein" may have stored in one data set as "A.Bernstein", while another data set as "Arnold Berstein". Same name can reside in several attribute fields as well. For an example, name "Levenshtein" can b e a name of an author as well as a part of an article titled such as "Levenshtein Distance". Heterogeneity caused by same entity identified by different names are not going to be resolved in this approach. Because the original format which the data are stored at each data store is required for generating SPARQL queries and retrieve answers. The ambiguity of same literal reside under different fields (author, title etc.) are resolved by introducing specific keyword fields to the user.

Architecture of our proposed approach is depicted in Fig. 1. It contains three main components namely query validator which validates the user inserted query, keyword to attribute mapper which identifies the elements in user query, their relationships and how they can be related to existing data elements in the federation. Final component is query converter where the SPARQL queries are generated based on the keyword queries entered by user. Each of these components is discussed in upcoming sections.

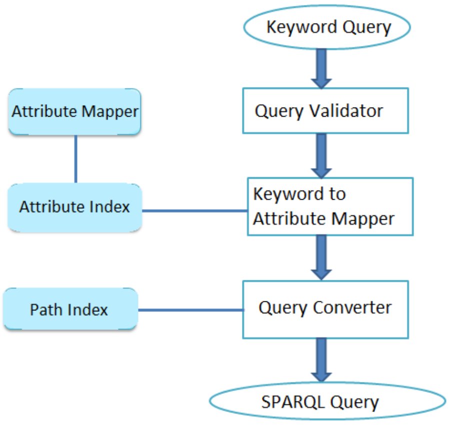


Fig. 1 Architecture of the proposed approach

doi: 10.4038/icter.v9i1.7168

IV. Query Validator

This component is used for validating the format of input keyword query. Keyword queries adhere to a defined format are only focused in this methodology. Keyword queries of proposed format are only expected by next steps in this approach as well. Therefore, validating the input query format is essential. A simple regular expression based lexical analyzer and a parser is used for this purpose. Once the user query is inserted, it is directed to the lexical analyzer for tokenizing and removing unnecessary white spaces. If any error occurred during this stage, the query will be rejected and an error is displayed. If the query was accepted by the lexical analyzer, a token steam is sent to the lexical parser. A regular expression which defines the accepted format of user query is included in the lexical parser. If the input query matches the grammar rules, it is a valid query. Invalid queries are notified as errors. Valid queries are stored as key-value pairs and sent to the next component.

V. Keyword to Attribute Mapper

Once the Query Validator sent a set of key - value pairs based on the input user query, Keywords to Attribute Mapper have several tasks to complete. They are as follows: identifying the fields which user query consists of, retrieving matching data source elements for user keywords from the federation, retrieving vocabulary dependent terms for user query's variable fields from heterogeneous vocabularies in the federation, clustering identified matching element sequences for each data source based on the vocabulary they belong to and finally ranking those clusters based on their capability to answer the user query.

There are several sub tasks aligned with the above list of tasks. In order to identify matching elements for user keywords, this research proposes maintaining pre-processed data as practiced by several other existing solutions [1], [6], [8], [10], [11]. With the pre-processed data, searching can be done faster. Two types of pre-processed data are proposed for this phase namely Attribute Mapper and Attribute Index.

A. Attribute Mapper for Heterogeneity Resolution in the Federation

Vocabulary level heterogeneity is a common characteristic of RDF data source federations. Even in the domain of academic publications, fields similar in meaning are represented in different labels. For an example, ACM RDF data source identifies publication title as "title" and SWE DBLP identifies publication title as "label". However, in order to identify that all those different labels means the same entity, a mapping is required among those vocabularies.

Ontology alignment is a highly researched field which can also be utilized for resolving vocabulary level heterogeneity only. On the other hand, WordNet[4] is a lexical database created for English. It has categorized different English words based on their similarities. Also, it keeps details about the origin of words. For example, if we consider the word, "author", we can retrieve that the parent class of "author" is "person". Also if we input "author" and "editor", we can retrieve the common word "person" which can be used to identify both of them. Hyponym and hypernym relations indicated in WordNet serve this purpose. These are two options for resolving vocabulary level heterogeneity. When considering the terms in vocabularies, it was identified that some terms which are defined as concepts on some vocabularies are defined as properties on other vocabularies in the federation. Therefore, both concept level and property level alignments are required in this approach. Alignment API [12] was selected for this task as it has capability of property and concept alignment. Its accuracy is better than other approaches [13], [14] and it also has the capability of integrating WordNet which provides added advantage.

Alignment API [12] is only capable of aligning two vocabularies at a time. Therefore, this research uses a semi-automated approach to resolve vocabulary level heterogeneity among all the vocabularies in the federation and construct the Attribute Mapper. First, the required concepts for the specified scope are manually identified from a single vocabulary. For an example, if DBLP is considered, "label" is the predicate used to identify publication title, "author" is the predicate used for indicating author list of a particular publication etc. Then all other vocabularies in the federation are aligned with DBLP source using Alignment API. For an example align vocabulary of ACM, vocabulary of CiteSeer with DBLP each at a time. Once all the output alignment details are received, only the alignment of all vocabularies gets started. All the entities (classes, attributes etc.) aligned with previously identified concepts of DBLP are clustered together along with their data source details. This helps to extract different terms used by heterogeneous sources to identify same entity in the federation.

B. Attribute Index for Mapping Keywords with Data Elements in the Federation

Matching data elements for user keywords must be identified as the first step of keyword to SPARQL conversion. Commonly used keyword index approach [5], [15] is decided to use for this phase with several modifications.

Definition 1: A *keyword index* is a keyword to element map which returns a set of matching elements to a keyword.

RDF is a graph structured data store. Data vocabulary elements are represented by vertices and edges of a graph.

Definition 2: A *RDF graph g* is a tuple (V, L, E) where

- V is a finite set of vertices as the union V_E U V_V with entity vertices V_E and value vertices V_V
- L is a finite set of edge labels as the union L_R U L_A with relation labels L_R and attribute labels L_A
- E is a finite set of edges of the form $e(v_1,v_2)$ with v_1,v_2 ϵ V and e ϵ E. Following types of edges can be defined:
    - e ϵ L_A (attribute edge) if and only if v_1ϵ V_E and v_2 ϵV_V
    - e ϵ L_R (relation edge) if and only if v1,$_{v2}$ϵ V_E
    - type is a special relation which indicates the membership of an entity to a particular class

[4] http://wordnet.princeton.edu/

doi: 10.4038/icter.v9i1.7168

Most of the available approaches have indexed both V and L in their keyword index, arguing that keywords can occur in V and L both. But Cappellari et al. [7] mentions that keywords can mostly reside on sinks. Sink is a node in RDF graph which does not have any outgoing edges from it. They have introduced a term called source to identify vertices which have no incoming edges. So they have only indexed sink vertices of a RDF graph. We also have decided to follow this approach and generate the index only for sink vertices. We are adopting the data structure used by Tran et al. [8] to store additional details about indexed elements. Additional details are the set of adjacent edges directed from one-edge distant vertices to this element, set of primary edges which are the adjacent edges to a source which this element belongs to, type of the source V_E to which this sink V_V belongs to and data source identifier. This additional information is required efficiently identifying suitable sub-graphs for query generation. Apache Lucene[5] document index was utilized for building our Attribute Index.

In most data sources, sink elements do not reside in one edge distance from its source vertex. For an example if we consider a publication, it can have many authors. In such cases, a collection approach is required to store multiple authors. So the adjacent edge to the element may not be the adjacent edge to the source. Adjacent edge to the source is the one which defines the relationship not the adjacent edge to the sink. Situations where these two are different have not been addressed by previous approaches. Therefore, this research proposes a primary edge to be stored as well as the adjacent edge. Primary edge is the adjacent edge of the source to which element connects. There can be either no other edges between primary edge and adjacent edge on the path or both adjacent and primary edges can be same or several edges can occur between primary and adjacent edge. But they are not subject to store in the data structure.

Proposed approach uses same keyword index to store all sink values from all the data sources in the federation. If same element resides in two or more datasets, several data structures for each source are created and stored under the same key. Only the sink element labels get indexed while all the information resides in data structure gets mapped to indexed element. Therefore, it takes a less amount of space to store this keyword index than other methods' indexes who index all the vertex data.

Fig. 2 shows a sample data graph fragment of DBLP. "James Peter" is a source as it does not have any outgoing edges. Paper-Peter95 is its source. Source does not have any incoming edges. The data structure which returns from the Attribute Index for the term "James Peter" is [James Peter (node label), ns#1(adjacent edge), author (primary edge), Book Chapter (type of source), DBLP (data source id)].

Once Attribute Mapper and Attribute index are ready, real time processing begins. Using the key-value pairs received from query validator, this component identifies what are the variable fields of the keyword query. Those variables are sent to Attribute Mapper and retrieve vocabulary dependent terms for them. Condition values of the keyword query are sent to Attribute Index and receive matches.

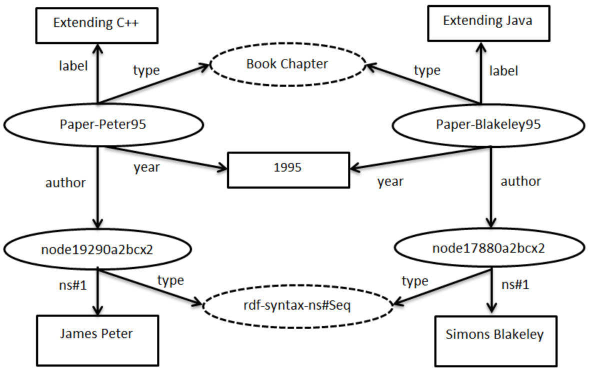


Fig. 2. Data graph fragment of DBLP

Primary edges of those index results are matched with attribute mapper to make sure we receive the values with the required attribute we want. Now cluster those results based on the data source identifier. Then we employ a data source ranking approach to decide which source is most capable of answering user query. In a non-distributed environment, it is most advisable to generate single query, process and send its results to users without keeping user waits until all queries for the entire federation are generated because users need efficiency. If a data source has matching elements for all the user keywords while another source only have matching elements for some of the user keywords, the former data source gets a high priority when query converting. Those ranked data sources are then sent to Query Converter component.

Following table shows sample tuples of DBLP and ACM for the query "publication:?,author:James,year:1995". Variable field of the query is "publication". That was sent to attribute mapper and "label" and "has-title" was retrieved for DBLP and ACM respectively. When consider condition values, DBLP has matching elements for both "James" and "1995" while ACM has matching elements only for "James". Therefore it is clearly understood that there is a more possibility of retrieving accurate results from DBLP them ACM for this query. Therefore DBLP gets higher priority.

TABLE I
OUTPUT TUPLES FROM KEYWORD TO ATTRIBUTE MAPPER

Source	**Keyword**	**Primary Edge**	**Adjacent Edge**	**Source Type**
DBLP	James Peter	author	ns#1	Book Chapter
	1995	year	year	Book Chapter
	Variable field : publication , vocabulary dependent value : label			
ACM	James McClelland	has-author	full-name	Article Reference
	Variable field : publication , vocabulary dependent value : has-title			

[5] http://lucene.apache.org/

doi: 10.4038/icter.v9i1.7168

VI. Query Converter

This component is used to identify the most suitable sub-graph which can connect identified keyword elements from the previous component and generate SPARQL queries based on identified sub-graphs. SPARQL queries are generated by identifying the format (template) of the sub-graphs which consist of the answers for the query. Vertices of the target sub graph were found by Keyword to Attribute Mapper. However, edges which connect those vertices were not provided. Hence Query converter first has to seek for a method for identifying relationships among sent elements by previous component. Many approaches [6], [8], [9] try to find suitable sub-graphs by graph traversing at real time, which is highly time consuming. A path store [7] is utilized in this research for sub-graph identification because it has totally eliminated real time graph traversal.

Objective of Path Index is to keep records of how vertices (classes, property values and literals of RDF graphs) are connected to each other prior to actual query processing starts. Paths are defined as the route from given element to another in RDF graph. Efficiency of query processing improves by utilizing those pre-processed data. Therefore, the cost required for real time graph traversal reduces heavily. Those path data are stored in a relational database.

Sources and Sinks are the main elements required for defining paths. Full path is defined by a route from a source to a sink. Template of a path is retrieved by replacing vertices of a path by wild cards. Templates indicate the relationship among vertices. When considering the SPARQL query generation context, those templates are the components which we need to find for generating queries. When considering a SPARQL query, intermediate nodes of a path is always indicated by variables. Therefore, we can easily generate queries by utilizing suitable templates.

Database schema presented in Path Index mainly consists of four main relations; Node, Template, Path and PathNode. Path Index assumes that user queries targets only sinks [7]. Therefore, only data about sinks are stored in Node table. Path table keeps data about all the unique paths from sources to sinks. Each tuple consists of path is, template id, length of the path and id of the final node of the path. PathNode table keeps track of which node resides in which position of a path. Index Organized Tables concept is used for indexing these tables.

Database schema presented in Path Index was adopted for this research due to its simplicity and usefulness for generating queries. Cappellari et al. [7] has presented a breadth first search based approach for exploring the RDF graph when populating tables with data in the database. This approach stores intermediate paths as well as full paths in Path table. When utilizing Path Index for SPARQL query generation, it was decided that storing full-paths are only required. Since the main intention of using this index is to identify relationships among keyword elements, sources can be considered as most promising connecting elements which can reach many other elements quickly once matching elements are found from sinks. Therefore sources can behave as local connecting points when we are trying to find the best sub-graph to connect keyword elements.

A. Full-path Identification

Cappellari et al. [7] have presented a breadth-first search based algorithm for capturing full-paths in a given RDF dataset. However that algorithm stores full paths as well as partial paths which requires huge unnecessary space. Therefore a depth-first search based algorithm was proposes in this methodology which can identify full-paths accurately avoiding unnecessary space wastage caused by the original algorithm. Proposed algorithm explores the RDF graph in Depth First manner. Sources available for each dataset are identified and depth first traversal starts from each source. It searches the entire RDF graph until it meets all the sinks which can be reached by the current source. Algorithm locally creates an n-ary tree considering current source as the root. All the triples whose subject is root are considered as the branches of the tree. If the object of each triple has become a subject in another triple, those branches grow accordingly. Sink nodes never become a subject of a triple so occur as leaves of the tree. Proposed algorithm for generating path tree for each source is shown below.

Algorithm 1 DFS based graph exploration for full-path capturing

Input : RDF triple dataset *DATA*, path tree *TREE*, parent node *P*, matching triple set *TRIPLES*

```
for each triple t in TRIPLES do
    add t to node P on TREE
    var newtriples = get triples from DATA (subject = t.object)
    if (newtriples.count > 0) then
        DFSbasedgraphtraversal (DATA, TREE,t.object,newtriples)
    end if
end for
```

Once a source-tree is generated, a method is required to identify all the full paths in the tree because those are needed to be stored in Path Index. Since tree nodes have only a single parent, full-paths can be captured by recursively traversing from each leaf to root. Proposed algorithm is shown below.

Algorithm 2 Complete algorithm for Path Index building

Input : RDF triple dataset *DATA*, path tree *TREE*, source list *SOURCES*

```
for each source s in SOURCES do
    var triples = get triples from DATA (subject = s)
    var TREE = generate tree (root = s)
    Algorithm 1 (DATA,TREE,s,triples)
    var pathlist
    var leaves = TREE.leaves
    for each leaf l in leaves do
        var leafnode = leaf
        var path
        while (leafnode $\neq$ TREE.root) do
            add leafnode to path
            leafnode = leafnode.parent
        end while
        add path to pathlist
    end for
    store pathlist in PathIndex
end for
```

B. *Resolving Cycles Caused by Inverse Properties*

Inverse relationships are a main reason for occurring cycles in RDF graphs. Suppose there is a triple is a RDF graph whose subject is *S*, predicate is *P*, object is *O*. If there exists another triple in the same RDF graph whose subject is *O*, predicate is *P'* and object is *S*, *P* and *P'* has an inverse relationship.

Both triples connected by an inverse relationship contains same amount of information. Therefore removing one does not cause any information loss in the RDF graph. A popularity based approach is used for filtering the most appropriate triple. Popularity is measured by the number of unique predicates each subject has. Higher the number of unique predicates, higher the amount of information it has access to. Proposed algorithm for inverse property based cycle resolution is shown in Algorithm 3.

Algorithm 3 Resolution for inverse property based cycles in graphs

```
Input : RDF triple dataset DATA
Output : Cycle resolved dataset DATA
Initialisation : Inverse statement list inverselist
for each triple t in DATA do
    if (inverselist.notcontain(t)) then
        var inversetriples = get triples (subject=t.object,object=t.subject)
        for each inverse i in inversetriples do
            var subjectpopularity = get unique predicate count (i.subject)
            var objectpopularity = get unique predicate count (i.object)
            if (subjectpopularity > objectpopularity) then
                add t to inverselist
                remove t from DATA
            else
                add i to inverselist
                remove i from DATA
            end if
        end for
    end if
end for
```

C. *Path Index Based Sub-graph Identification and SPARQL Query Generation*

Once Path Index generated, SPARQL query generation should be done in real time. Once mappings are retrieved, Path Index is queried to retrieve paths whose final node is the value of the mapping received from Attribute Index. There can be several sub-graphs in a data source which can connect those keyword elements. But they all use the same schema (vocabulary) Consider a situation in DBLP where 3 sub-graphs exist which connects author James, year 1995 with 3 publications. Those publications become answers as publication was the variable in user query. If all the vertices of each sub-graph are replaced with wild cards, the result graph is totally similar. That graph is the sub-graph which is needed to traverse to get answers. That is the sub-graph which we should convert to SPARQL syntax for generating the query. This sub-graph with wildcards is known as template graph. . For an example, template of full-path “PaperPerter95-year-1995” is “#-year-#”. When converting a sub-graph to SPARQL syntax, we have to replace the vertices by variables.

Several different templates can be received as matching paths when same keywords repeat under different concepts in the vocabulary. For an example, consider person "James". He can be a program committee member of one conference in 1995 while being an author of publications. Since templates were extracted only considering full-paths whose sink nodes are keywords without focusing on their relationship with variable field, templates matching for both scenarios will occur. If results were generated for both these template graphs, overall relevancy of results will become low. Therefore a filtering process is required. Additional information stored in index documents comes to use at this situation. Filtering process considers templates which matches with primary and adjacent edges keyword elements and ignore others. If there are several results on this approach, shortest template will be selected as the suitable sub-graph as lesser the length of the path, faster the query processor can reach it and output results. Tran et al. [8] has also mentioned path length as a common matrix used by many graph traversal approaches to rank selected sub-graphs. Lesser the path length, there is a high probability that it would reduce the overall size of the sub-graph it resides in.

Next this proposed methodology looks for extracting suitable templates for variable fields in the user query. Shortest template which contains the vocabulary specific properties of the variable field is selected as the template. A sub-graph which can connect all the keywords is required for generating a SPARQL query. Now paths for each field have been retrieved, finding connecting elements is required to generate a sub-graph using these paths. Details stored in Path Index are used for finding connecting elements. Sources of the data sets were identified while building Path Index. Sources are operating as centre nodes to connect all the property values of the source together. Sources mostly represent the main focus of the vocabulary. For example, if DBLP and ACM vocabularies are considered, publication is their main focus. All the attributes related to publication are defined as properties. Therefore sources can be identified as a potential element for connecting the paths for generating a possible sub-graph. SPARQL queries are generated based on this argument. Following is a sample query generated for the example keyword query “publication:?, author:James, year:1995”..

```
SELECT ?z
WHERE {
    ?x type Book Chapter .
    ?x label ?z .
    ?x year 1995.
    ?x author ?y.  ?y ns#1 “James Peter”. }
```

If there are several matching elements for a single keyword (Ex : Many different people with "James" as a part of their name) or many different sub-graphs match for the query, FILTER option of SPARQL is used in generated queries. Following is an example is a scenario where there are multiple authors named "James" available.

```
SELECT ?z
WHERE {
    ?x type Book Chapter .
    ?x label ?z .
    ?x year 1995.
    ?x author ?y.
    ?y ns#1 ?a .
     FILTER(regex(str(?a),"James")).
 }
```

VII. Results

The Proposed approach was implemented using Java with the support of Apache Jena Semantic Web library and Oracle 11g DBMS. Once a keyword query is inserted, it identifies the data sources in the federation which can answer the keyword query and transforms input query to vocabulary dependent SPARQL queries to match with the data sources in the federation.

Evaluation setup is as follows. A federation of 3 publication data sets were created. DBLP[6] RDF dataset with 37446 triples, ACM[7] RDF dataset with 63980 triples and Semantic Web Dog Food[8] (SDF) dataset with 37105 triples were used as test data. Portions from DBLP and ACM data sets were used because SDF data set was not large enough as them and publication details between 1986 and 2005 were only assessed due to the availability of data on all 3 data sets. These three sources were selected based on following aspects.

Publishers of these data sets have exploited three different ontologies (SWEtoDBLP, aktors and SWRC ontologies respectively) for publishing these data. Therefore schema level heterogeneity is clearly showcased among the selected data sources. RDF data are stored in graph structures. Therefore problems in graph data handling also arise when dealing with RDF data. Cycles are one such problem. Here we selected two data sets with cycles. ACM data set has cycles caused by the subject of a triple has become its own object. SWRC ontology is an integrated ontology of several ontologies. Therefore it has inverse relations. These inverse properties have introduced cycles in SDF data set. These data sets are used to experiment the proposed approach's ability to deal with common cyclic scenarios of RDF data. DBLP data set does not consist of cycles. However, specialty with this data set is that it consists of blank nodes. Blank nodes are anonymous nodes in a RDF graph. These can be used to group sub-properties of an instance. Likewise, data sets which exhibit different characteristics which covers most of the common characteristics of RDF graphs are used for the experiment in order to show the generalizability of this approach for RDF federations.

Experiments were conducted on a machine with AMD V120 processor of 2.2 GHz and 4GB RAM. A test keyword query set of 10 queries were created by considering all the possible combinations of the three keyword fields (author, year, publication) we selected for academic publication data. Table II shows test query set.

TABLE III
TEST QUERY SET

	Keyword query
Q1	publication : consistency , author : ?
Q2	publication : distributed , year : ?
Q3	author : sylvia , year : ?
Q4	author : andrew , publication : ?
Q5	year : 1986 , author : ?
Q6	year : 1988 , publication : ?
Q7	publication : multimedia , author : daniel , year : ?
Q8	publication : concurrent , year : 1990 , author : ?
Q9	author : david , year : 2004 , publication : ?
Q10	publication : protocol , author : ? , year : ?

A. *Quality Evaluation of the Federation*

The first evaluation criterion was to evaluate whether our proposed approach actually carries out its intended task of correctly translating user keyword queries to SPARQL. Measurements were taken by considering the relevancy of the retrieved results by executing the generated queries. Since main intention of this approach is to give general users more relevant and accurate results using the privileges of Semantic web, quality of results were measured using following measurements. Quality of the generated SPARQL queries were evaluated by measuring precision, recall and F-measure of the results received for above mentioned test query set. F-measure is a balanced measurement used to capture the balance between precision and recall of each result set. This measure was used as some results can be high in precision but low in recall and vice versa. F-measure gives a balanced score in such situations. Gold standard results were obtained by running SQL queries on Path Index of the data sources and manual evaluation on the raw RDF data sets. Fig. 3 shows the results graphically.

Overall precision has an average of 0.98 and overall recall has an average of 0.9. Based on the precision results, proposed approach is capable of generating queries which can give more accurate results. However, a loss of recall was detected compared to precision. This was caused by the decision made about finding the connecting elements of sub-graphs. The source node was selected as a connecting element and its type was decided by the type which has maximum matching from the attribute index. Sometimes this type did not match with the actual source of the paths which were retrieved from the Path Index. It caused loss of results. Another point was that the capability of a data source in producing results were decided for a user query if that source have matching elements for all the keywords in the query. However there are situations which keyword element reside in the RDF graph, but there is no a sub-graph which can actually connect them all. In such situations generating a query is wastage of effort.

Then recall values were compared of each data source with the recall of the federation. This evaluation was done to identify whether there is a significant impact on the result by processing the query over a federation rather than on an individual data source. One of the objectives of transforming a keyword query to SPARQL on a federation was to give a more complete result set to the user. If a single source is capable of giving the same result set, there is no requirement of a federation. Recall is the measurement which indicates the contribution of each source over gold standard result set.

[6] http://lsdis.cs.uga.edu/semdis/swetodblp/
[7] http://datahub.io/dataset/rkb-explorer-acm
[8] http://data.semanticweb.org/

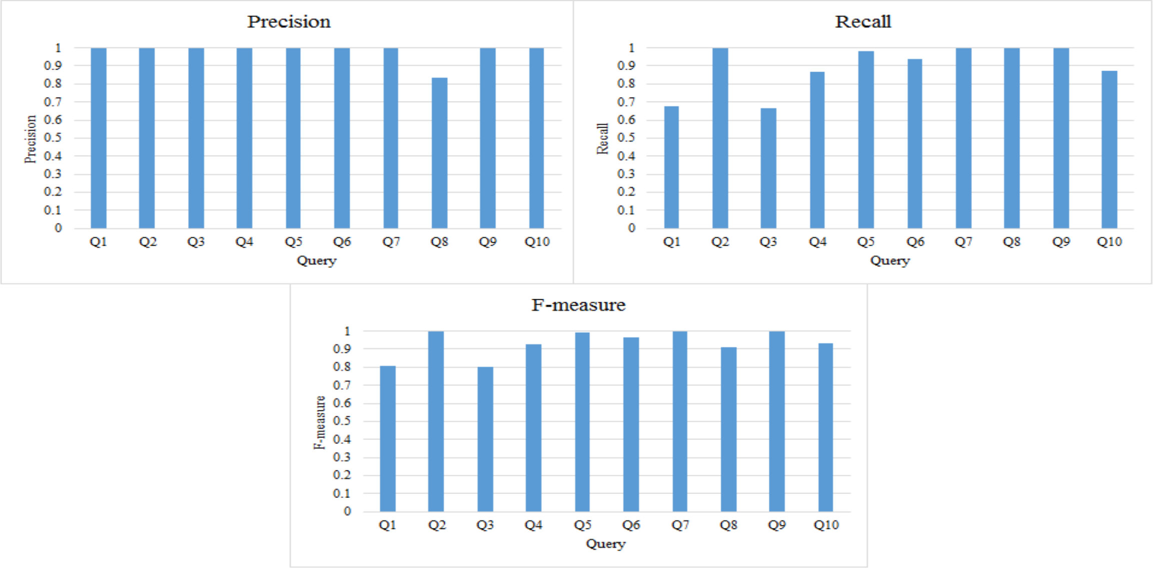


Fig. 3. Quality evaluation of the federation

Fig. 4 shows that federation gives more complete results over individual sources. Therefore it can be shown that federation approach is capable of producing more complete results than any of the individual sources.

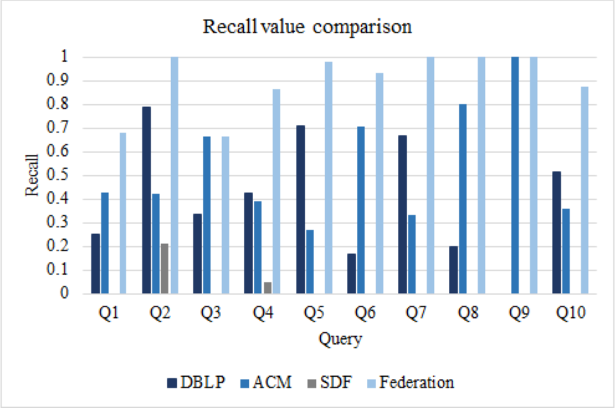


Fig. 4. Recall comparison of the federation

B. *Performance Comparison for the Proposed Path Index Based Approach*

Performance evaluation was carried out after receiving satisfying results for the quality of the proposed approach. Prior to this research, all the presented approaches for this task [8], [28] have either used online graph traversal approach or natural language processing support for query translation. Path index [7] was first suggested for keyword searching. Its functionality was exploited for generating SPARQL queries from keyword queries. An index based approach was adopted related to this method used by Tran et al. [8] Both these approaches and SearchWeb have utilized a keyword index for mapping keywords with data source elements. However proposed method used a path index based approach for identifying substructures which can connect keyword elements while they followed a real time graph traversal mechanism.

The graph exploration and top-k query calculation approach presented in [8], performs better compared to other available methods for finding sub-graphs such as backward search [9], bidirectional search [6], breadth first search and depth first search. Tran et al. [8] has shown in their evaluation that query generating time has achieved a comparable decrement by using their graph exploration mechanism. Natural language based approach suggested in [28], is totally deviated from the proposed approach. The intention of using a path index based approach for query generation was to experiment whether it can achieve a performance gain in query generation time by pushing graph traversing to the pre-processing stage. Therefore, query generation time of proposed approach against SearchWeb approach presented by Tran et al. [8] was evaluated.

Time taken to identify the most suitable query substructure which can be used for query generation as our matrix of performance was measured. Once it is identified, either proposed approach or their approach could have been used for translating the substructure to match with SPARQL query syntax. Time taken by SearchWeb approach to generate top-1 query was only considered since proposed approach only output a single query per data source. Query generation time only for ACM data source was compared at this section as SearchWeb method has only suggested for query generating for a single data source. In the next section, the performance comparison when it comes to federation scenario will be demonstrated. Fig. 5 shows the performance comparison.

Figure illustrates that proposed approach has a significant performance gain compared with SearchWeb suggested by Tran et al. [8] for the tested query set. Since search web has already been outperformed other related methods, we only evaluated against SearchWeb. Main reason behind the high query time for search performance is its real time graph traversing. Even though SearchWeb doesn’t do a full graph traversal in query generation time, it uses graph traversal. First, they create a graph summary by extracting the class vertices and entity vertices from the original RDF graph. This summary graph behaves as the schema. Once the matching elements are retrieved using the keyword index, SearchWeb embeds those matching elements to the summary graph while exploiting the "adjacent edge" property in the retrieved index records. Therefore, more the

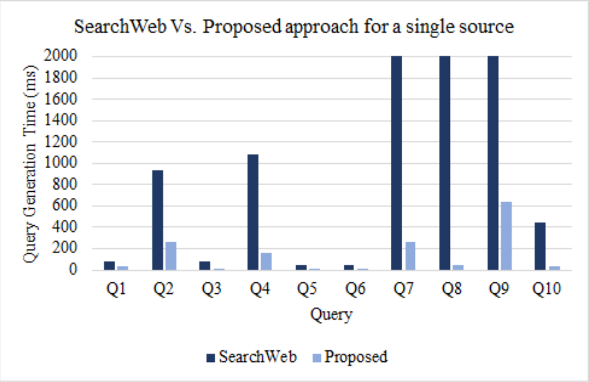


Fig. 5. SearchWeb Vs. Proposed approach for a single source

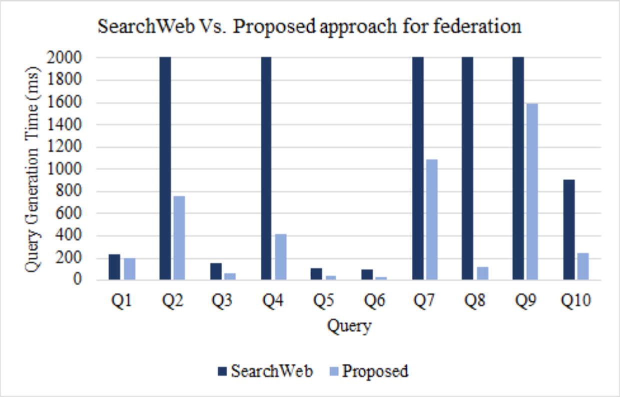


Fig. 6. Performance comparison for the federation

keyword matching elements, bigger the summary graph will be.

Keyword elements matching for "distributed" in Q2 is around 50 and "andrew" in Q4 is around 30. Once those 30 vertices are added to the summary graph, it gets bigger. SearchWeb traverse through all the possible starting from the shortest once generated the augmented summary graph by adding those matching elements. This is done to retrieve elements which can connect all keyword elements. Once a connecting element is found, SearchWeb considers all the path combinations among keyword elements even they have same adjacent edge to get the shortest path. Bigger the graph, higher the number of combinations will be. This leads to high query generation time.

Q7, Q8 and Q9 have two conditional keywords in the query. Therefore augmented graph of SearchWeb becomes bigger. Number of possible paths among elements also increases drastically. That is the reason for the huge query generating time.

In comparison, proposed approach shows a huge performance gain as there is no real time graph traversal in it. No matter how much matching elements are output from the keyword index, if they have same template, all of them are considered as a single element from template's point of view.

C. Performance Comparison for the Federation

Query generation times for the entire federation. Proposed approach is capable of generating SPARQL queries to all the sources in the federation in one go. However SearchWeb can only generate a SPARQL query for one source at a time. Therefore queries to all the sources in the federation were repeatedly generated and got the total time for the comparison. A huge performance gain was retrieved by this proposed approach this time as well. This showed that real time graph traversal for SPARQL query generation and keyword searching is highly inefficient when dealing with huge data sources. Proposed approach is the first research which utilized Path Index [7] for SPARQL query generation. Satisfying results were obtained on its performance. Fig. 6 shows performance comparison for the federation.

D. Proposed Federation Approach Vs. Digital Libraries

The main reason for emerging Semantic Web concept over Web of Data is ability of Semantic web to extract the meaning and relationships of data elements and exploiting them forgiving more meaningful and relevant results for user queries.

Once after comparing the performance of the proposed approach, this section focused on evaluating this aspect by running the same set of user queries on both digital libraries and proposed semantic web approach. Three digital libraries were selected for this evaluation. DBLP, ACM and Semantic Web Dog Food data dumps were used for sample federation in this research. Therefore DBLP digital library[9], ACM digital library [10] and Semantic web [11] online data store were selected as test digital libraries for this evaluation criterion. Google Scholar[4] was not used for this evaluation as it consists of publications from many sources, not only from above three.

Among these three digital libraries, DBLP and ACM only have advanced search capabilities. Semantic Web Dog Food website doesn't have an advanced search capability. Therefore, Semantic Web digital library was excluded from evaluation. When consider the query executable ability of digital libraries compared to proposed semantic web federated approach, proposed approach was capable of executing all the test queries. This means the proposed approach was capable of executing any combination of target query fields (publication title, author name, published year). In contrast, digital libraries were mostly capable of providing direct answers when only question needs publication titles as results.

For an example, consider Q4 and Q6. They ask for the paper titles authored by author named "Andrew" and paper titles published in year 1988. In Q9 user requests titles of publications authored by "David" in 2004. DBLP and ACM were able to answer those three queries correctly but Semantic web search was not advanced enough to directly provide answers to match with those conditions.

When it comes to queries like Q1, Q2, Q3, Q5, Q7, Q8 and Q10, those queries does not request publication titles. For an example, Q1 wants all the author names publishes papers titles with "consistency". Q3 wants a list of years

[9] http://dblp.uni-trier.de/

[10] http://dl.acm.org/

[11] http://data.semanticweb.org/

which author "Sylvia" has publications. Digital libraries were not capable of providing either an author name list or year list as answers for these queries. They provided a list of publications with word "consistency" for Q1. By manually extracting only their author names could be retrieved. Similarly for Q3, a publication list was received which were authored by "Sylvia". Their years needed to be manually extracted. Years were easier to be captured in DBLP while author list could be easily received with filtering in ACM. In comparison, proposed semantic web approach was capable of providing direct answers for all those queries as well. Fig. 7 shows comparison among F-measures of DBLP and ACM digital libraries with proposed federated approach.

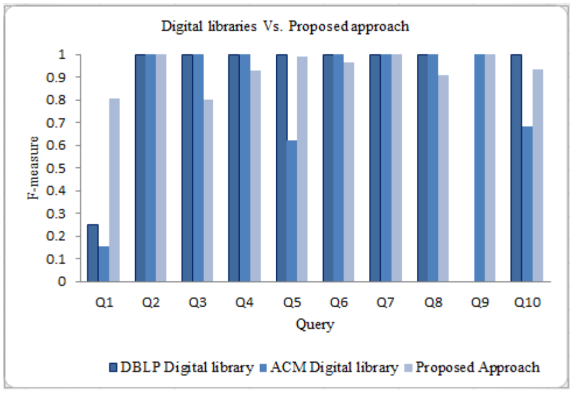


Fig. 7 Quality comparison of Digital libraries vs. Proposed Approach

This shows that our Semantic Web based federation approach is capable of giving highly relevant results for user queries than existing keyword matching digital libraries.

Google Scholar[12] follows a non- semantic web approach for combining all those results. However it also cannot answer queries like Q1, Q2 and Q5 directly. It also only focuses on filtering publications. If this proposed approach can be applied on all the publication sources, It could have performed better in answering all the publication related queries no matter whether it is about an author, publication, year or venue. This clearly exhibits the usefulness of semantic web and level of accuracy which can be gained from executing SPARQL queries over keyword based queries.

In queries like Q1, Q2 and Q5, user needs a list of authors or years. Semantic web approach was capable of giving that relevant information to user. However digital libraries were unable to output the relevant results. Users have to manually filter in order to retrieve the relevant result set. Therefore in relevancy of results, proposed semantic web approach stays in a high level compared to digital libraries including Google Scholar.

VIII. Discussion

A path index based keywords to SPARQL query transformation mechanism which aims at RDF data source federations is presented in this paper. A keyword index along with ontology alignment based vocabulary level heterogeneity resolution approach was suggested to identify matching elements for user keywords. A Path Index based approach was used for identifying suitable sub-graphs for connecting keyword elements. Real time graph traversal is one of the drawbacks of existing keyword query processing approaches as they extremely effect the performance as RDF data sources contains thousands to billions of nodes. This approach has totally eliminated real time graph traversal for query generation by generating an index for graph data in the pre-processing stage. It showed a significant performance gain over existing query generation approaches. Promising level of results were achieved from the quality evaluation of this approach and it was proved that federations are capable of giving more complete results for a user query than just querying from a single data source. This research also emphasized that Semantic Web related keyword query processing approaches can give more relevant results for user queries than traditional keyword matching.

This research can be further extended by including capabilities for handling more relaxed format queries on this approach. An approach which can accurately decide the connecting elements for the extracted paths from the Path Index in order to generate sub-graphs can be used to further enhance the accuracy of this approach. A mechanism which can decide whether there is actually a sub-graph existing for a given set of keywords before generating the SPARQL query is needed to be merged with this approach in the future. Also this research can be further extended to generate SPARQL queries for Linked Data sources by exploiting the Path Index. Map-reduce based mechanisms can be used to enhance this approach to function on distributed environments which will improve the scalability of this approach.

Acknowledgment

We would like to thank Sarasi Lalithasena, PhD candidate of Kno.e.sis Research Center, Wright State University, USA, for her expert advice and encouragement throughout this research.

References

[1] Hristidis L. G. V. and Papakonstantinou Y. (2003). Efficient ir-style keyword search over relational databases. *VLDB*, pp. 850–861.

[2] Hwang V. H. H. and Papakonstantinou Y. (2006). Objectrank: a system for authority based search on database. *SIGMOD Conference*, pp. 796–798.

[3] Liu W. M. F., Yu C. T., and Chowdhury A. (2006). Effective keyword search in relational databases. *SIGMOD Conference*, pp. 563–574.

[4] Kimelfeld B. and Sagiv Y. (2006). Finding and approximating top-k answers in keyword proximity search. *PODS*, pp. 173–182.

[5] Wang H. He, H., Yang J., and Yu P.S. (2007). BLINKS : Ranked Keyword Searches on Graphs. *Proceedings of the 2007 SIGMOD International Conference on Management of Data*, ACM.

[6] Kacholia S. C. S. S. R. D. V., Pandit S., and Karambelkar H., (2005). Bidirectional expansion for keyword search on graph databases. *VLDB*, pp. 505–516.

[7] Cappellari P., De Virgilio R., Maccioni A., and Roantree M. (2011). A Path-Oriented RDF Index for Keyword Search Query Processing. *DEXA*, pp. 366–380.

[8] Tran S. R. T., Wang H., and Cimiano P. (2009). Top-k exploration of query candidates for efficient keyword search on graph-shaped (rdf) data. *ICDE,IEEE*.

[9] Bhalotia C. N. S. C. G., Hulgeri A., and Sudarshan S. (2002). Keyword searching and browsing in databases using banks. *ICDE*, pp. 431–440.

[10] Tran T. and Zhang L. (2013). Keyword Query Routing. *IEEE Transactions on Knowledge and Data Engineering*, 1(2).

[12] https://scholar.google.com/

doi: 10.4038/icter.v9i1.7168

[11] Freitas A., Curry E., Oliveira J.G., and O'Riain S. (2011). a Distributional Structured Semantic Space for Querying Rdf Graph Data. *International Journal of SemanticComputing*, 05: 433–462.

[12] David F. S. J., Euzenat J., and dos Santos C. (2011). The Alignment API 4.0. *Semantic web journal*, 2(1): 3–10.

[13] Jain A. K. P., Hitzler P., and Yeh P. (2010). Ontology alignment for linked open data. *The Semantic Web ISWC 2010*, pp. 402–417, Springer Berlin Heidelberg.

[14] Li Y. J., Tang J. and Luo Q. (2009). RiMOM: A dynamic multistrategy ontology alignment framework. *Knowledge and Data Engineering, IEEE Transactions*, 21: 1218–1232.

[15] Tran T., Wang H., and Haase P. (2009). Hermes : Data Web search on a pay-as-you-go integration infrastructure. *Web Semantics: Science, Services and Agents on theWorld Wide Web*, 7(3):189–203.

[16] Sheth A.P. and Larson J.A. (1990). Federated Database Systems for Managing Distributed , Heterogeneous , and Autonomous Databases. *ACM Computing Surveys (CSUR)*, 22(3): 183–236.

[17] Makris K., Gioldasis N., and Bikakis N. (2010). Ontology Mapping and SPARQL Rewriting for Querying Federated RDF Data Sources (Short Paper). *On the Move toMeaningful Internet Systems, OTM 2010*, pp. 1108–1117, Springer Berlin Heidelberg.

[18] Correndo G., Salvadores M., Millard I., Glaser H., and Shadbolt N. (2010). SPARQL Query Rewriting for Implementing Data Integration over Linked Data. *Proceedings of the 2010 EDBT/ICDT Workshops*, p. 4, ACM.

[19] Makris K., Bikakis N., Gioldasis N., and Christodoulakis S. (2012). SPARQL-RW : Transparent Query Access over Mapped RDF Data Sources. *Proceedings of the 15th International Conference on Extending Database Technology*, no. c, pp. 610–613, ACM.

[20] Gunaratna S. K. and Sheth A. (2014). Alignment and dataset identification of linked data in Semantic Web. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4: 139–151.

[21] David F. J. and Briand H. (2006). Matching directories and OWL ontologies with AROMA. *Proceedings of the 15 th ACM international conference on Information and knowledge management*, pp. 830–831, ACM.

[22] Wikipedia, Breadth-first search," 2002. [online], http://en.wikipedia.org/wiki/Breadth-first_search (Accessed: 21 June 2014).

[23] Wikipedia, "Depth-first search," 2002. [online], http://en.wikipedia.org/wiki/Depth-first_search (Accessed: 21 June 2014).

[24] Florescu D. K. D. and Manolescu I. (2000). Integrating keyword search into xml query processing. *Computer Networks*, 33(1-6): 119–135.

[25] Guo C. B. L., Shao F. and Shanmugasundaram J. (2003). Xrank: Ranked keyword search over xml documents. *SIGMOD Conference*, pp. 16–27.

[26] Yu B., Li G., Sollins K.R., and Tung A.K.H. (2007). Effective Keyword-based Selection of Relational Databases. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 139–150, ACM.

[27] Vu Q.H., Ooi B.C., Papadias D., and Tung A.K.H. (2008). A Graph Method for Keyword based Selection of the top-K Databases. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 915–926, ACM.

[28] Unger C., Bühmann L., Lehmann J., Ngonga Ngomo A.C., Gerber D., and Cimiano P. (2012). Template-based Question Answering over RDF Data. *Proceedings of the 21st international conference on World Wide Web*, vol. ACM, pp. 639–648.

An Archived firefly Algorithm: A mathematical software to solve univariate nonlinear equations

M.K.A. Ariyaratne, T.G.I. Fernando, S. Weerakoon

Abstract— **In this article, we are presenting a software solution that proposes some modifications to the existing firefly algorithm. The modification known as the archived firefly algorithm [AFFA] exhibits the ability of finding almost all complex roots of a given nonlinear equation. The software implementation includes two main properties; an archive to collect the better fireflies and a flag to determine poor performance in firefly generations. The new modification is tested over genetic algorithms (GA), a phenomenal success in the field of nature inspired algorithms and also with a modified GA embedded with same properties that the AFFA has. A simple graphical user interface (GUI) is developed using MATLAB GUIDE to present the findings. Computer simulations show that the AFFA performs well in solving nonlinear equations with real as well as complex roots within a specified region.**

Keywords—**Firefly Algorithm, Nonlinear Equations, Archive, Real Roots, Complex Roots.**

I. Introduction

The computer scientists are looking forward to find various approaches that can contribute towards optimization. The future of optimization is now being conquered by modern meta-heuristic algorithms. Genetic algorithms, differential evolution, harmony search, firefly algorithm and cuckoo search are such meta-heuristic algorithms which have marked their success over many optimization tasks. Simplicity of the algorithm, less memory consumption and the minimal error of the approximations can be stated as the major reasons for their popularity. Natural optimization techniques are among them and are becoming popular due to the long lasting existence of such practices in the real world. Genetic algorithms; which have come to the stage around 1970 going along with the theory of evolution by Charles Darwin, can be identified as one of the first such algorithms. They still play an important role among nature inspired algorithms [1], [2]. Ant colony optimization algorithms which mimic the ants' strange communication behaviour are other popular algorithms which have been adopted for many real world optimization tasks [3], [4]. Bat inspired algorithm [5], cuckoo bird algorithm [6], firefly algorithm [7] are more recent nature inspired algorithms that represent different real world optimized phenomena. These algorithms can be classified into main categories as evolutionary algorithms and swarm intelligence algorithms. Evolutionary algorithms are population based algorithms which use mutation, recombination, and natural selection to reproduce better generations [8]. Genetic algorithms, differential evolution [9] and genetic programming [10] are examples for evolutionary algorithms. Swarm intelligence, by its name mimics the collective behaviour of different elements in the natural world. Particle swarm optimization [11], ant colony systems, firefly algorithms are some examples. One of the most advantageous properties of these algorithms is that most of them are of meta-heuristic type. Therefore these algorithms can be adopted to solve variety of optimization problems rather than heuristic algorithms. This paper presents the research carried out using a modern nature inspired algorithm; firefly algorithm to solve univariate nonlinear equations having real as well as complex roots. We also present the mathematical software that has been developed to accomplish the task.

II. Nonlinear Equations

Solving a nonlinear equation f(x)=0 is finding roots α, such that f(α)=0. Many numerical methods exist to solve such nonlinear equations, but they have some drawbacks such as the need for derivative information, strong dependence on the initial guess and the inability to give all the roots within an interval simultaneously. When the equation is having complex roots, the situation is even more difficult. Scientists have moved towards finding better algorithms to minimize these drawbacks. But we have to ensure the efficiency of the computational effort as well. The algorithms should maintain the speed, accuracy and low memory consumption. The research literature reveals that many improvements have taken place to tune these existing approaches as well as in finding new ways to solve nonlinear equations reducing the above mentioned drawbacks.

An improvement to the Newton's method is suggested by Weerakoon & Fernando in their joint paper: A Variant of Newton's Method with Accelerated Third-Order Convergence [12]. Their method involves changing the derivation of Newton's method. The derivation of the Newton's method involves approximating an indefinite integral of the derivative of the function by a rectangle. In their research Weerakoon & Fernando have modified it to be a trapezium so that the error of the approximation is reduced. They have proved that the order of convergence of the suggested modification is three. In fact, for some cases it is even higher than three. The main concern in this research was on accuracy of the approximation. However this method also contains the aforementioned drawbacks of numerical methods such as the need for derivative calculation.

Moving beyond numerical methods we could come across a handful of research done with the nature inspired algorithms to fulfill the task. Use of genetic algorithms to

Manuscript received on 24 Nov 2015. Recommended by Prof. K. P. Hewagamage on 16 June 2016.

This paper is an extended version of the paper "A Modified Firefly Algorithm to solve Univariate Nonlinear Equations with Complex Roots" presented at the ICTer 2015.

M.K.A. Ariyaratne and T.G.I. Fernando are from the Department of Computer Science, Faculty of Applied Sciences, University of Sri Jayewardenepura.(*anuradhaariyaratne@gmail.com,tgi.fernando@gmail.com*).

S. Weerakoon is a Professor at the Department of Mathematics, Faculty of Applied Sciences, University of Sri Jayewardenepura. (*sunethra.weerakoon@gmail.com*).

doi: 10.4038/icter.v9i1.7169

solve systems of nonlinear equations is addressed in some researches [13], [14], [15]. A hybrid algorithm implemented with genetic algorithms and particle swarm optimization also has been tested in a research [16]. Applicability of harmony search in solving system of nonlinear equations is also addressed [17].

Heuristics were also in use of solving systems of nonlinear equations. One research has introduced the use of continuous global optimization heuristic called "continuous GRASP" to solve a nonlinear system [18]. They are addressing the problem of finding all the roots of a system of equations assuming that all roots are real.

Most of these approaches have focused on solving systems of nonlinear equations rather than a single equation. In almost all these research, they have dealt only with the real roots. The problem of finding all roots in a reasonable range within a single run is also not addressed.

Our problem of interest is to solve a univariate nonlinear equation having real and/or complex roots. We need to find all the roots within a reasonable interval/range. Since we have interpreted the problem as an optimization problem, we can define the above problem as follows.

Let f be a function s.t. f: D→R where D⊂C,where C is the set of Complex Numbers. The problem is to find all x∈ D s.t. f(x)= 0, without requiring either the differentiability or the continuity of the function f. Thus we need to find x ∈D s.t. |f(x)|=0. However, since the function f(x) may have multiple roots, the optimization problem |f(x)|=0, also will have multiple optimal solutions. Our objective turns out to be finding all such optimal solutions.

III. ARCHIVED FIREFLY ALGORITHM

AFFA is the software solution proposed in this paper to solve univariate nonlinear equations with real as well as complex roots. Many nature inspired algorithms are good solution providers for various optimization tasks and their algorithms are very simple. That made it easier for the researchers to develop these algorithms for different optimization tasks. Firefly algorithm is a newly implemented such algorithm with the following assumptions about fireflies' behaviour [7].

1. Attraction of the fireflies to each other is gender independent.
2. Attractiveness is proportional to the brightness, for any two fireflies, the less bright one will be attracted and move towards the brighter one, this attraction decreases when distance increases, the brightest firefly will move randomly.
3. Brightness of a particular firefly is determined by its objective function.

The following pseudo code describes the original firefly algorithm and a simple 20 line MATLAB program can implement the algorithm solving a given mathematical optimization problem. MATLAB as a mathematical software provides an easy environment to implement these types of algorithms. It is also capable of creating user interfaces and enrich with plotting different functions and data so that the solutions can be graphically represented.

Algorithm 1: Original Firefly Algorithm

```
Begin;
Initialize algorithm parameters:
MaxGen: the maximum number of generations
$\gamma$: the light absorption coefficient
$r$: the distance between two fireflies
$D$: the domain space
Define the objective function $f(X)$,      where
$X = (x_1, \dots, x_d)^T$
Generate the initial population of fireflies,$X_i$   $(i = 1,2,\dots,n)$

Determine the light intensity $I_i$ of
        $i^{th}$ firefly $X_i$ via $f(X_i)$
while $t <$ MaxGen do
for $i = 1 : n$ (all n fireflies) do
for $j = 1 : n$ (all n fireflies) do
if $I_j > I_i$
            Move firefly i towards j byusing eq (1);
End if
Attractiveness varies with distance
  r via $e^{-\gamma r^2}$ using eq (2);
Evaluate new solutions and update
      light intensity;
end for
end for
Rank the fireflies and find the current best;
end while
Post process results and visualization;
End;
```

The initial population can be defined randomly with a set of feasible solutions for the problem. Then each firefly's light intensity is calculated using the problem specific objective function. Then each firefly in the population starts moving towards brighter fireflies according to the following equation.

$$x_i = x_i + \beta(x_j - x_i) + \alpha(rand - 0.5); \qquad (1)$$

where$\beta = \beta_0 e^{-\gamma r^2}$ (2)

β_0 is the attraction at $r = 0$;

The second term of the .1 is due to the attraction between two fireflies and the third term is a randomization term where α is the randomization factor drawn from the Gaussian or the uniform distribution.

In the modified version we have added a few more qualities to the original algorithm by introducing an archive and a flag. The new algorithm is named after its archiving property as archived firefly algorithm. The pseudo code of the modified algorithm is as follows.

Algorithm 2:Archived Firefly Algorithm

```
Begin;
Initialize algorithm parameters:
MaxGen: the maximum number of generations
$\gamma$: the light absorption coefficient
$r$: the distance between two fireflies
$D$: the domain space
Define the objective function $f(X)$,where
$X = (x_1, \dots\dots, x_d)^T$

Generate the initial population of fireflies,$X_i$
$(i = 1,2, \dots\dots, n)$
Determine the light intensity $I_i$ of
```

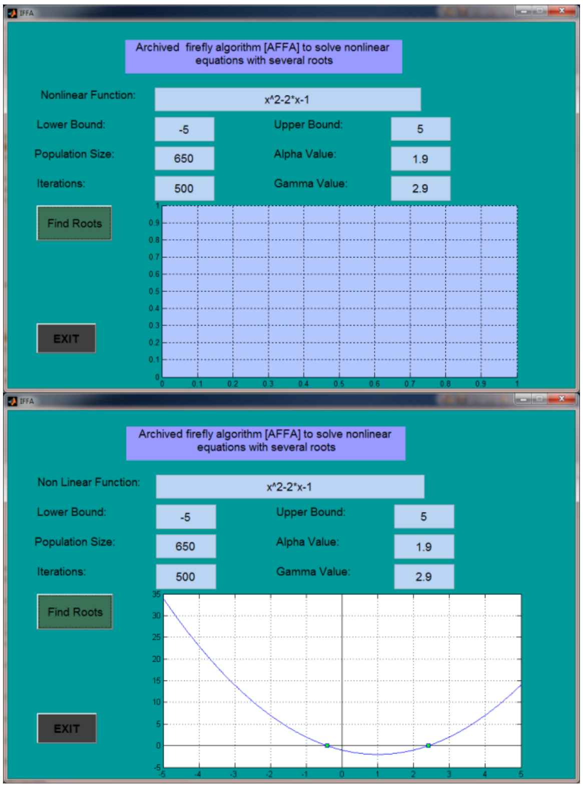
doi: 10.4038/icter.v9i1.7169

```
            $i^{th}$ firefly $X_i$ via $f(X_i)$
while $t < MaxGen$ do
flag = true;

while flag = true and $t < MaxGen$ do
for $i = 1 : n$ (all n fireflies) do
for $j = 1 : n$ (all n fireflies) do

if $I_j > I_i$
        Move firefly i towards j byusing eq (1);
End if
Attractiveness varies with distance
  r via $e^{-\gamma r^2}$ using eq (2);
Evaluate new solutions and update
      light intensity;
end for
end for
  Find the fireflies with the eligibility criteria
  $|f(X)| < 0.001$;
  Put them into the archive and replace the
      positions with random fireflies;

    If no fireflies matching with eligibility criteria
      flag = false;
    End if

    If flag = false

    count = random integer between 0 and n;
    Create random fireflies up to count and
              replace the population;
    End if
end while
end while
Post process results and visualization;
End;
```

IV. MATLAB APPLICATION

A MATLAB software is developed successfully to graphically represent the results of AFFA. To create graphical user interfaces, MATLAB GUIDE (graphical user interface design environment) is used [20]. MATLAB GUIDE provides tools for designing user interfaces for custom applications. Using the GUIDE layout editor, we have designed a simple user interface. GUIDE then automatically generate the MATLAB code for the user interface, and the user can modify the program to control the application.

The first step was taken to solve nonlinear equations with real roots only. The user has to provide the equation, the lower and the upper bounds to specify the interval to find roots and the values for problem specific parameters (such that population size, number of iterations, alpha and gamma values). Then the simulator will display the root approximations. Fig.1 shows the simple GUI created in the MATLAB environment which is implemented to display the results graphically.


Archived firefly algorithm [AFFA] to solve nonlinear equations with several roots

Nonlinear Function: x^2-2*x-1

Lower Bound: -5   Upper Bound: 5

Population Size: 650   Alpha Value: 1.9

Iterations: 500   Gamma Value: 2.9

Archived firefly algorithm [AFFA] to solve nonlinear equations with several roots

Non Linear Function: x^2-2*x-1

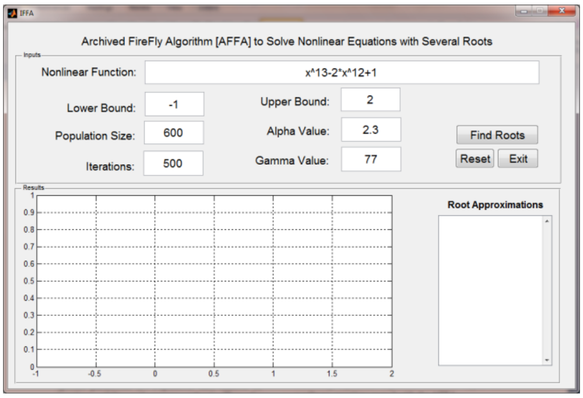
Lower Bound: -5   Upper Bound: 5

Population Size: 650   Alpha Value: 1.9

Iterations: 500   Gamma Value: 2.9

Fig. 1: A MATLAB GUI implemented to solve univariate nonlinear equations with real roots

When it comes to complex roots the environment is different. When the function is having real as well as complex roots, plotting the function and displaying the roots is sometimes not clear; instead we used real and imaginary axes to plot the roots. Apart from that we added a text box to display the root approximations. With all these modifications the new GUI to solve univariate nonlinear equations with real and/or complex roots is as follows.


Archived FireFly Algorithm [AFFA] to Solve Nonlinear Equations with Several Roots

Nonlinear Function: x^13-2*x^12+1

Lower Bound: -1   Upper Bound: 2

Population Size: 600   Alpha Value: 2.3

Iterations: 500   Gamma Value: 77

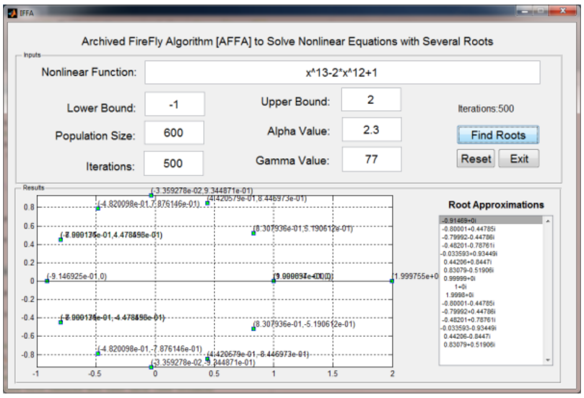
Fig. 2: A MATLAB GUI implemented to solve univariate nonlinear equations with real and/or complex roots

The new application displays the roots in an argand diagram and the root approximations are displayed in a list box. Because of the archiving property we can expect more than one approximation for an existing root. Apart from these, the application has a reset button to reset the inputs.

We have solved nonlinear functions having real as well as complex roots using the new application.

$$y = x^{13} - 2x^{12} + 1 \qquad (3)$$

The above polynomial has 13 roots; 3 real and 10 complex roots. We have used our application to find the root approximations with the accuracy of $10^{(-3)}$


IFFA

Archived FireFly Algorithm [AFFA] to Solve Nonlinear Equations with Several Roots

Inputs

Nonlinear Function: x^13-2*x^12+1

Lower Bound: -1

Upper Bound: 2

Population Size: 600

Alpha Value: 2.3

Iterations: 500

Gamma Value: 77

Iterations:500

Results

Root Approximations

```
-0.91469+0i
-0.80001+0.44785i
-0.79992-0.44786i
-0.48201-0.78761i
-0.033593+0.93449i
0.44206+0.8447i
0.83079-0.51906i
0.99999+0i
1+0i
1.9998+0i
-0.80001-0.44785i
-0.79992+0.44786i
-0.48201+0.78761i
-0.033593-0.93449i
0.44206-0.8447i
0.83079+0.51906i
```

Find Roots
Reset
Exit
(-3.359278e-02,9.344871e-01)
(4.420579e-01,8.446973e-01)
(-4.820098e-01,7.876146e-01)
(8.307936e-01,5.190612e-01)
(-8.000176e-01,4.478498e-01)
(-9.146925e-01,0)
(9.999934e-01,0)
(1.999755e+0
(-8.000176e-01,-4.478498e-01)
(8.307936e-01,-5.190612e-01)
(-4.820098e-01,-7.876146e-01)
(4.420579e-01,-8.446973e-01)
(-3.359278e-02,-9.344871e-01)
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.5
0
0.5
1
1.5
2

Fig. 3: Root approximations given by the application for the Eqn. 3.

V. NUMERICAL EXAMPLES

Several nonlinear functions are tested against the proposed MATLAB application of AFFA. Mainconcern is to identify different varieties of nonlinear equations where the modified algorithm can be applied. The following equations represent different types of nonlinear equations tested with the AFFA.

a. The following one dimensional trigonometric equation (adapted from Goldberg and Richardson, 1987) [2] has 51 real roots within the given interval and the AFFA approximated all 51 roots with an accuracy of10^{-3}. This equation proves the ability of AFFA in finding all the real roots of a nonlinear equation within a given interval.

$$y = si\,n^3(5\pi x) \qquad (4)$$
$$where\ x \in [-5, 5]$$

b. Usual numerical methods need the equation to be differentiable to apply those methods to find roots. The Weierstrass functions possesses the property of being continuous everywhere but differentiable nowhere [19]. Therefore numerical approaches are not suitable for such functions. But the AFFA as a mathematical software application for a modified nature inspired algorithm is capable of handling such functions successfully. Here we present an example Weierstrass function having 25 real roots within [-20, 20] interval.

$$W(x) = \sum_{i=1}^{3} \left(\frac{1}{2^i}\right) \sin(2^i\, x) \qquad (5)$$

c. Numerical methods also require the property of continuity of a nonlinear function to be solved. $y = (x)$ is a popular nonlinear function with discontinuities. Our solution does not need the continuity of the function and is capable of approximating all the roots in a large interval. (As an example we have used the interval [-40, 40].

$$y = \tan(x) \qquad (6)$$

d. There are some nonlinear functions where there is additional difficulty in calculating the roots. These are functions in which the desired root has a multiplicity greater than 1. Let α be a root of the function f(x), and assume that it could be factored into the form $f(x)=(x-\alpha)^m\ h(x)$, with some integer m ≥1 and some continuous function h(x) for which h(α)≠0. Then we say that α is a root of f(x) with multiplicity m. One of the main difficulties with the numerical calculation of multiple roots is that methods such as Newton's method and the secant method do not converge in general.
We have solved the following nonlinear equation and found that our approach is successful in finding root approximations for the multiple roots.

$$y = x^4 - 2x^2 + 1 \qquad (7)$$
$$where\ x \in [-2, 2]$$

e. The suggested application is capable of finding complex roots too. When the function is having both real as well as complex roots within a given range, the application calculates accurate approximations. The following equations were tested using the software.

TABLE I
ROOT APPROXIMATIONS GIVEN BY THE APPLICATION FOR THE EQN. 3.

	Equation	**Region**	**#of Roots**
1	$y = x^2 + 1$	[-1.5,1.5] X [-1.5, 1.5]	2 complex
2	$y = x^3 + 2x^2 + 3x + 4$	[-2, 0] X [-2, 0]	2 Complex, 1 Real
3	$y = x^5 - 3x^4 + 3x^3 - 2x^2 - 5$	[-1, 3] X [-1, 3]	4 Complex, 1 Real
4	$y = x^7 - x^6 + 2x^5 - 3x^4 + 3x^3 - 2x^2 - 5$	[-1, 2] X [-1, 2]	6 Complex, 1 Real
5	$y = x^{10} - 3x^9 + x^8 - 7x^7 + x^6 - x^4 + 2x^2 - 5$	[-1, 4] X [-1, 4]	8 Complex, 2 Real
6	$y = x^{12} - 6x^{11} + x^{10} - 5$	[-1, 6] X [-1, 6]	10 complex, 2 Real
7	$y = x^{13} - 2x^{12} + 1$	[-1, 2] X [-1, 2]	10 Complex, 3 Real

Here also we seek roots within a specified region. Region of the function is the area we seek for the roots. According to the notation we adopted, [-1, 2] X [-1, 2] region describes the area ABCD shown in Fig. 4.

doi: 10.4038/icter.v9i1.7169

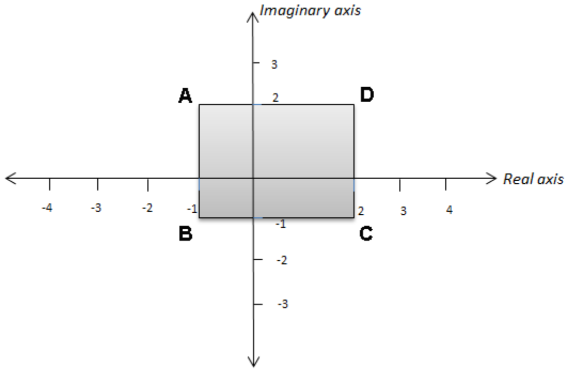


Fig. 4: [-1, 2] X [-1, 2], an example region used for the root finding

VI. RESULTS AND DISCUSSION

The quality of the developed software was compared with the quality of each of the three algorithms: the original firefly algorithm, the original genetic algorithm and also with a genetic algorithm embedded with the new features of AFFA. For the comparison purpose we have set the population size to 200 and the number of iterations per run to 200, as a reasonable scale for real roots and for both real as well as complex roots we allowed both of these values to be 600. We have run our application for 100 times (100 runs) and the average number of roots given by each algorithm, (maximum number of roots found / total roots)*100% are calculated. As an example in Table 2, a result 19, (49%) indicates the average number of roots found in 100 runs is 19 and the (maximum number of roots found in a run / total roots)*100% is 49%.

TABLE II
PERFORMANCE OF THE ALGORITHMS FOR REAL ROOTS OVER 100 RUNS

I. FUNCTION	II. A	Modified GA	FA	AFFA
$y = \sin^3(5\pi x)$	0, (0%)	51, (100%)	19, (49%)	51, (100%)
$y = \sum_{i=1}^{3}\left(\frac{1}{2^i}\right) sin(2^i x)$	1, (4%)	25, (100%)	13, (68%)	25, (100%)
$y = \tan(x)$	1, (4%)	25, (100%)	21, (96%)	25, (100%)
$y = x^4 - 2x^2 + 1$	2, (50%)	4, (100%)	4, (100%)	4, (100%)

The real roots situations were smoothly handled by both modified genetic algorithms and the proposed AFFA. The archiving property and diversifying the population during iterations is the reason for the good performance. When it comes to original algorithms, firefly algorithm performs better than GA.

TABLE III
PERFORMANCE OF THE ALGORITHMS FOR REAL AS WELL AS COMPLEX ROOTS OVER 100 RUNS

EQUATION	GA	Modified GA	FA	AFFA
Table 1: Equation 1	0, (0%)	2, (100%)	2, (100%)	2, (100%)
Table 1: Equation 2	1, (33%)	1, (33%)	3, (100%)	3, (100%)
Table 1: Equation 3	1, (20%)	1, (20%)	5, (100%)	5, (100%)
Table 1: Equation 4	1, (14%)	1, (14%)	7, (100%)	7, (100%)
Table 1: Equation 5	1, (10%)	1, (10%)	9, (90%)	10, (100%)
Table 1: Equation 6	1, (8%)	3, (25%)	11, (92%)	12, (100%)
Table 1: Equation 7	1, (8%)	2, (15%)	8, (92%)	13, (100%)

The modified genetic algorithm performed well for real roots but it fails in finding complex roots. Experiments done with larger population sizes (1200, 1500) were able to give complex roots to some extent. To solve for complex roots, modified GA needs a large chromosome population under the given parameter setting to provide approximations. To check its availability over complex roots, we have to do more research on its recombination and selection parameters. With the results obtained we claim that AFFA under same conditions performed well for the complex roots as well. It is capable of handling nonlinear equations having both real as well as complex roots. The suggested method can be further extended to solve a given system of nonlinear equations for real as well as complex roots.

VII. CONCLUSIONS

The implementation of the new software to find roots of a univariate nonlinear equation having real as well as complex roots is successful. MATLAB, mathematical package is used to implement the application and the graphical user interface design environment of MATLAB (GUIDE) is also applied successfully. The modification is successful within a reasonable interval/ region. The modification includes an archive to collect best fireflies during iterations and replacing their positions with random ones. Also a flag was used to identify poor iterations and to change the randomness of the existing population. The simulation results for finding roots of several numerical examples including an application of Weierstrass function and complex polynomials suggest that this new firefly algorithm with the archiving property is capable of finding almost all real as well as complex roots of a nonlinear equation simultaneously with the given accuracy of 10^{-3}. It works for the multiple roots functions too. Comparison with other similar nature inspired algorithms including the original firefly algorithm clearly shows that this modified firefly algorithm outperforms all of them. This evidence suggests that the proposed firefly algorithm is by far the best performer in solving nonlinear equations with several real as well as complex roots.

Although the implemented software works well we would like to suggest some modifications to improve it further. The algorithm lacks a good parameter optimization procedure. That is anyone who uses the software has to enter algorithm specific parameters by themselves. Since the users are not experts in the field of nature inspired algorithms we can't expect them to enter good parameter values. In this initial step we have set all parameters to some reasonable values where most of the equations can be dealt with. But research is open to find methods to do a goodparameter optimization for the algorithm for a given situation.

With the results obtained, we can conclude that the proposed firefly algorithm is capable of giving reasonably good approximations for the nonlinear functions:

a. with several roots.
b. with multiple roots.
c. which are continuous but not differentiable.
d. which have discontinuities within the interval.
e. which are having a pattern within the given range.
f. which have roots over a large interval.
g. having complex roots as well as real roots in a given interval.

Accuracy of the roots found by the modified FA is within 10^{-3} tolerance. For higher accuracies one can treat these solutions as initial guesses and try out a suitable numerical approach. The accuracy of the solutions is limited to 10^{-3} because our objective here is to find almost all the roots within the given region. The approximations provided here highly depend on the population size, number of iterations and also on the algorithm specific parameter values. It is essential to define the number of iterations and the population size properly according to the number of roots within the specified interval.

Differentiability and continuity of the nonlinear functions are inessential when applying nature inspired algorithms to obtain roots; thus they could be applied to the functions arising from various practical situations where it is impossible to apply formal numerical schemes. This can be considered as the biggest advantages of using nature inspired algorithms.

The basic firefly algorithm introduced by Yang is powerful, but the problem of finding all real as well as complex roots of a given nonlinear equation simultaneously has not been addressed before. Thus our approach of introducing an archive is undoubtedly advantageous. The software with a graphical user interface gives users a user-friendly environment to use the application. But still this approach needs higher number of iterations and alarge firefly population when we need higher accuracies in approximations. As an improvement, we can test parameter optimization techniques which are able to adapt with our algorithm. Apart from that, as a further improvement, one can check the ability of the given procedure in solving a given system of nonlinear equations.

References

[1] Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning (16th ed.). Reading, MA: Addison-Wesley Educational Publishers.

[2] Goldberg, D., & Richardson, J. (1987). Genetic algorithms with sharing for multimodal function optimization. Proceedings Of The Second International Conference On Genetic Algorithms On Genetic Algorithms And Their Application, 41-49. Retrieved from http://dl.acm.org/citation.cfm?id=42512.42519

[3] Dorigo, M., & Gambardella, L. (1997). Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Transactions On Evolutionary Computation, 1(1), 53-66. http://dx.doi.org/10.1109/4235.585892

[4] Dorigo, M. (1992). Optimization, Learning and Natural Algorithms, (Ph.D). Politecnico di Milano.

[5] Yang, X. (2010). A New Metaheuristic Bat-Inspired Algorithm. Nature Inspired Cooperative Strategies for Optimization (NICSO 2010) Studies in Computational Intelligence, 65-74. doi:10.1007/978-3-642-12538-6_6

[6] Yang, X., & Deb, S. (2009). Cuckoo Search via Levy flights. 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC). doi:10.1109/nabic.2009.5393690

[7] Yang, X. (2009). Firefly Algorithms for Multimodal Optimization. Stochastic Algorithms: Foundations and Applications Lecture Notes in Computer Science, 169-178. doi:10.1007/978-3-642-04944-6_14

[8] Back, T. (1996). Evolutionary algorithms in theory and practice. New York: Oxford University Press.

[9] Rainer, S., & P. K. (1997). Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. Journal of Global Optimization, 11, 341-359. http://dx.doi.org/10.1023/A:1008202821328

[10] Poli, R., Langdon, W. B., & McPhee, N. F. (2008). A Field Guide to Genetic Programming. Retrieved from http://www.gp-field-guide.org.uk

[11] Kennedy, J., &Eberhart, R. (1995). Particle swarm optimization. Proceedings Of ICNN'95 - International Conference On Neural Networks. http://dx.doi.org/10.1109/icnn.1995.488968

[12] Weerakoon, S., & Fernando, T. (2000). A variant of Newton's method with accelerated third-order convergence. Applied Mathematics Letters, 13(8), 87-93. http://dx.doi.org/10.1016/s0893-9659(00)00100-2

[13] El-Emary, I. M., & El-Kareem, M. A. (2008). Towards using genetic algorithm for solving nonlinear equation systems. World Applied Sciences Journal, 5(3), 282-289.

[14] Mastorakis, Nikos E. (2005). Solving non-linear equations via genetic algorithms. In Proceedings of the 6th WSEAS international conference on Evolutionary computing (EC'05), Ana Maria Madureira (Ed.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 24-28.

[15] Pourrajabian, A., Ebrahimi, R., Mirzaei, M., & Shams, M. (2013). Applying genetic algorithms for solving nonlinear algebraic equations. Applied Mathematics And Computation, 219(24), 11483-11494. http://dx.doi.org/10.1016/j.amc.2013.05.057

[16] Biswas, A., &Dasgupta, S. (2008). Finding Solution Of Simultaneous Non-linear Equations Using GA-PSO Hybrid Algorithm.

[17] Solving systems of nonlinear equations by harmony search. (2013). 13Th International Conference Computational And Mathematical Methods In Science And Engineering, IV, 1176-1186. Retrieved from http://hdl.handle.net/1822/27238

[18] Hirsch, M., Pardalos, P., &Resende, M. (2009). Solving systems of nonlinear equations with continuous GRASP. Nonlinear Analysis: Real World Applications, 10(4), 2000-2006. http://dx.doi.org/10.1016/j.nonrwa.2008.03.006

[19] Duncan, L. (n.d.). The Weierstrass Function. Retrieved from https://books.google.lk/books?id=pEbDNwAACAAJ

[20] MATLAB Help Files. GUI Building Basics", MathWorks. Url: http://in.mathworks.com/help/matlab/gui-building-basics.html

Employability and Related Context Prediction Framework for University Graduands: A Machine Learning Approach

Manushi P. Wijayapala, Lalith Premaratne, Imali T. Jayamanne

***Abstract*— In Sri Lanka, graduands' employability remains a national issue due to the increasing number of graduates produced by higher education institutions each year. Thus, predicting the employability of university graduands can mitigate this issue since graduands can identify what qualifications or skills they need to strengthen up in order to find a job of their desired field with a good salary, before they complete the degree.**

The main objective of the study is to discover the plausibility of applying machine learning approach efficiently and effectively towards predicting the employability and related context of university graduands in Sri Lanka by proposing an architectural framework which consists of four modules; employment status prediction, job salary prediction, job field prediction and job relevance prediction of graduands while also comparing performance of classification algorithms under each prediction module. Series of machine learning algorithms such as C4.5, Naïve Bayes and AODE have been experimented on the Graduand Employment Census - 2014 data. For the validation purposes, a wide range of evaluation measures was used to analyze the effectiveness of applying classification algorithms and class imbalance mitigation techniques on the dataset. The experimented results indicated that RandomForest has recorded the highest classification performance for 3 modules, achieving the selected best predictive models under hybrid approach having an area under the ROC curve interpretation as an 'Excellent' experiment, while a C4.5 Decision Tree model under Ensemble approach has been selected as the best model of the Salary Prediction module.

***Keywords*— Machine Learning, Employability Prediction, Data Mining, Supervised Learning**

I. Introduction

One of the main objectives of higher education is to prepare students to pursue different careers in a country. With many economies being reported as not producing adequate employment opportunities to absorb the growth in the working age population, a generation of productive young workers will have to face an uncertain future unless something is done to reverse this trend. Thus increasing the graduands' chances of obtaining decent jobs that match their education and training, equipping students in universities with the necessary competencies to enter the labour market, enhancing their capacities to meet specific workplace demands, improving the students' skills and qualifications to meet the employers' expectations are some of the essential tasks that need to be carried out in order to improve the employability of Sri Lanka [1].

In Sri Lanka, 'employability' has been a major topic among many parties in recent years. Especially, unemployable graduates and graduands are becoming a crucial issue that recent governments are facing. Conflicts between the parties involved in these matters are often experienced. Once it was difficult to find details about graduand unemployability, the census done by HETC with the guidance of the Ministry of Higher Education, proves to be a gold-mine and provided valuable insights into the main factors having a significant bearing on the employability of graduands. A systematic and scientific analysis using these data will result in a great solution for the issue of unemployment of graduands.

Machine learning (ML) has been recognized as a type of artificial intelligence which focuses on computer program development that can teach themselves to nurture without being explicitly programmed and change when exposed to new data [2]. In other words, the goal of ML is to invent or use the learning algorithms which will learn automatically without the human assistance or intervention.

The main aim of this research is to discover the plausibility of applying machine learning approach efficiently and effectively towards predicting the employability and related context of university graduands in Sri Lanka. Hence, objectives of this research can be devised as follows;

1. Propose a framework using ML based architecture to,
   a. Predict the employment status (Employed, Unemployed, Underemployed) of a university graduand at the time of official graduation
   b. Predict the salary range of an employable graduand (Very low, Low, Average, High)
   c. Predict the job relevance with the degree (Relevant field, Irrelevant field) of an employable graduand
   d. Predict the type of job field of an employable university graduand (Medicine field, Engineering field, Commerce field, Lecturing, Administrating field, Agriculture-Export field, and Support Staff field)

   while also coping with the constraints conflated in graduand employability data.
2. Identify the most important factors for the employment status of a university graduand, for the

Manuscript received on 18 April 2016. Recommended by Prof. G.N. Wikramanayake on 19 Dec 2016.

Manushi Wijayapala holds a B.Sc. (Hons) First class in Statistics with Computer Science from the University of Colombo, Sri Lanka and Bachelor of Information Technology (First class) degree from the University of Colombo School of Computing, Sri Lanka. (*manushimn@gmail.com*).

H.L. Premaratne is a Senior Lecturer at the University of Colombo School of Computing. (*hlp@ucsc.cmb.ac.lk)*

I.T. Jayamanne is a Lecturer at the Department of Statistics, University of Colombo. (*imali@stat.cmb.ac.lk*)

doi: 10.4038/icter.v9i2.7181

salary range of an employable graduand and for the type of job field of an employable graduand.

3. Compare and identify the most efficient and accurate classification algorithm/s to predict the employability of university graduands under each prediction module.

No successful prior research work has been considered fulfilling all the aforementioned research objectives. Even though researches have been carried out related to 'graduate' employability prediction [3, 4, 5], no research work has been found in literature related to predicting employability of university 'graduands'. Additionally, any kind of graduate or graduand employability prediction research has not been carried in the Sri Lankan context. Furthermore, none of the previous research work related to employability prediction of graduates have attempted to predict the job salary, job field and job relevance, which will be covered in this research. Moreover, we believe that this is the first occasion in employability prediction researches that considered the class imbalance problem and attempted this amount of class imbalance problem mitigation techniques to give more accurate results. Additionally, findings of this research can be used to reduce the overall unemployability of Sri Lanka. Even though this research focuses on graduand employability prediction, this can further extend with other sectors of the society to get a clear picture about unemployability.

This paper is organized as follows: Section II discusses related work; Section III gives an overview of the methodology; Section IV describes the experimental setup. Moreover Section V presents the results of the study and finally Section VI and Section VII present in-depth discussion together with conclusions and future work.

II. Related Work

The section exposes some of the previous research studies and results related to our research objectives.

Jantawan and Tsai [3] proposed a method to predict whether a graduate in Maejo University in Thailand will be employed, unemployed or undetermined. Thus authors try to build a graduate employment model using several classification algorithms in data mining and compare those algorithms to find the best algorithm to predict the employability in this university. The algorithms used by the authors for this comparison were Bayesian methods (AODE, WAODE, NaviveBayes, BayesNet and HNB) and tree methods (C4.5, BFTree, REPTree, NBTree, and ID3). Results showed that the WAODE algorithm, a type of Bayes algorithm has achieved the highest accuracy of 99.77%. However their framework is questionable since they have used two variables, 'work status of the graduate' and 'position of graduate' as the explanatory variables when they actually try to predict the employability status of the graduate. Because of using explanatory variables which are almost similar to the target variable, authors have gained an almost impossible accuracy of 99.8%.

Sapaat, Mustapha, Ahmad and Chamili [5], in their work focused on identifying features that influenced graduates' employability of Malaysian universities and tries to predict whether a graduate has been employed, remains unemployed or in an undecided situation in the first six months after the graduation based on actual data from the graduates. To accomplish it, authors have used data from the Tracer Study for the year 2009. The prediction has been executed through a series of classification algorithms of Bayes and decision tree methods. Results have shown that C4.5 decision tree algorithm gave the highest accuracy leading to the conclusion that a decision tree based classifier is more appropriate for the tracer data [5].

The study [4] presented a graduate employability model which uses different types of Bayesian methods to hunt the most important factors of graduate employability in Khon Kaen University, Thailand. In addition to that, authors try to compare the accuracy of all selected Bayesian algorithms. The researchers have used hold-out validation method to evaluate the models. The results have shown that the AODE and WAODE algorithms have gained the highest accuracy [4]. Furthermore, the experiment has shown that work province, the times which found the work and occupation type have a direct effect on employability.

All of the researches related to employability prediction [3, 4, 5] in literature, attempt to predict the employability status of university 'graduates'. However we could not find any research work related to predicting employability of university 'graduands'. A university graduate is a person who has already graduated whereas a university graduand is a person about to graduate after completing the degree. Furthermore, one of the limitations in their research work is they have only tried to predict the employability status of the graduate, i.e. whether a graduate is employed, unemployed or undetermined. These studies would have been more interesting if they had included modules which predict the salary range and the job field as well.

III. Methodology

In the previous section, we reviewed some existing related work and identified potential limitations in those approaches. This section outlines the proposed methodology to address the research questions by extensively describing the design aspects and the foundation of this study.

In this research, it is aimed to apply a machine learning (ML) based approach to build a framework of predictive models, which can correctly classify a university graduand into three classes first according to the employability status and then classify each employable graduand according to the job field, job salary and job relevance. The proposed framework is depicted in Figure 1. Model-1, Model-2, Model-3 and Model-4 which are depicted in this figure are the chosen best four models (best models are chosen according to the various evaluation measures that will be discussed in later sections) for four respective modules, which will be selected at the end of this study after mitigating all the conflated issues. It should be noted that second, third and fourth models will be applied only if the first model gives the output as 'employed'.

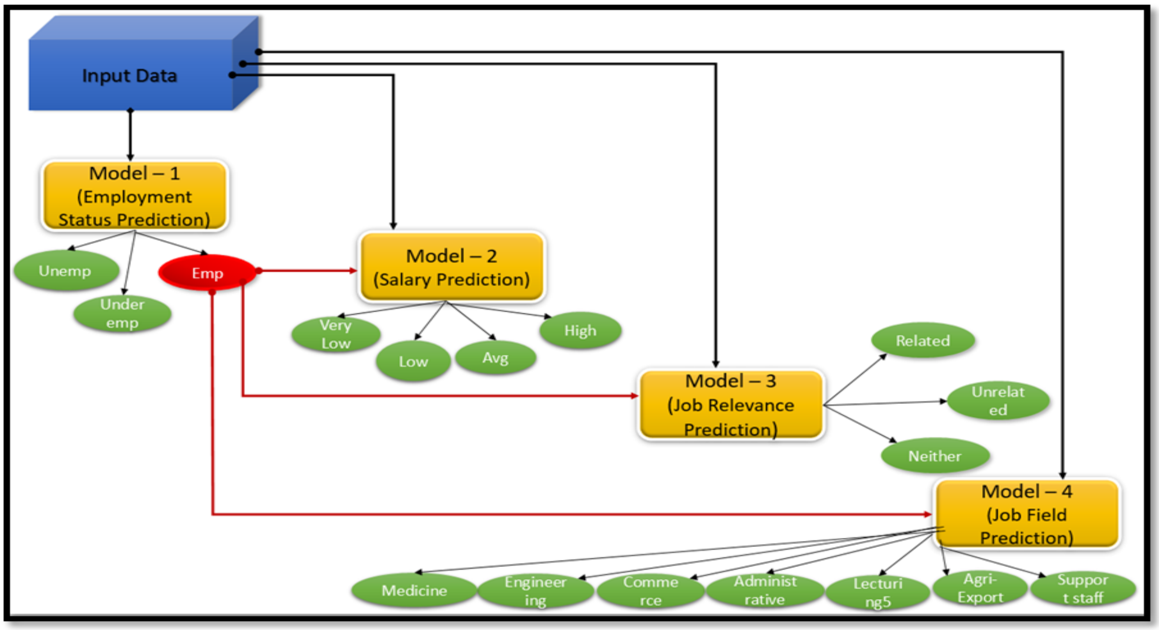


Fig. 1 Employability prediction framework proposed by the study

Reflecting the notion, the proposed comprehensive methodology for employability prediction of university graduands under four modules, explore decision tree algorithms, Bayesian algorithms and combination of multiple classifiers (ensemble methods) which are supervised learning models. Supervised learning would be the ideal solution to obtain a better classification, since there is a reasonable amount of annotated data already. In order to achieve a stable model, each classification algorithm was tried with a range of model parameters and compared them based on different evaluation metrics. Based on the constraints identified in university graduands' employment data and classification models (e.g.: class imbalance problem), several mitigation mechanisms were presented to overcome these limitations.

Figure 2 presents an architectural view of the proposed methodology. The proposed methodology basically consists of four main modules namely employment status prediction, salary prediction, job relevance prediction and job field prediction while the methodology of each of these four modules consist of four main phases namely pre-processing, feature selection, applying classifiers/ training models and selecting the best classifier as depicted in Figure 2. It should be noted that module 2, module 3 and module 4 depends on the output of module 1.

Original data which was taken from Graduand Employment Census – 2014, went through a series of pre-processing steps. This pre-processing stage consists of data cleaning (handling missing data, correcting inconsistent data and classifying detailed data into categories) and data

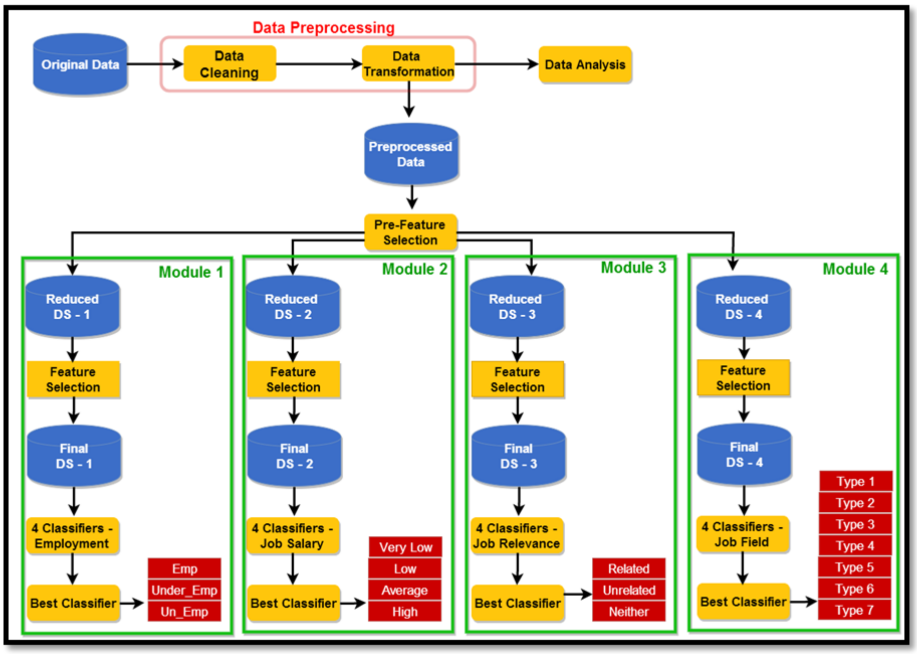


Fig. 2 Overall System Architecture

doi: 10.4038/icter.v9i2.7181

transformation (generalization and attribute construction). Most probably the original data set can have certain incompatibilities that will affect the performance of the final model. Therefore, a careful pre-processing would be vital to achieve better results in the next phase.

After pre-processing data, initial feature selection will be carried out using the domain knowledge, expert opinions and using previous literature. The reason for doing this initial feature selection was, the original data set contained additional variables which are not related to employability prediction at all, since employability prediction was not the only objective of the graduand employment census. For four different modules four sets of features will be selected from the original set of features. After doing the initial (manual) feature selection, automated feature selection will be carried out under each four modules separately as described in Section A.

After finalizing the initial steps, the next stage will be applying four different classifiers under each module to the final four data sets. Parameters of these classifier algorithms will be fine-tuned to suit the training models.

In order to overcome the limitations caused by a possible class imbalance phenomena, several mitigation techniques will be applied such as sampling, ensemble and hybrid approaches on these training models. Thereby the suitability of these approaches was measured on experimented algorithms and data. Employed different models were compared based on different evaluation metrics (accuracy, precision, recall, specificity, F-ratio, G-means and ROC AUC; area under the curve of receiver operating characteristic curve) to achieve a stable model. After the training, the model should be capable of predicting the employment status, job salary, job relevance and job field of the university graduands.

A. Feature Selection

Feature selection which is also known as subset selection, is the process of choosing a small subset of features that is sufficient enough to predict the targets easily and more accurately. The biggest of pros of applying feature selection techniques are being able to avoid over-fitting and being able to reduce the computational cost [6].

Even though decision trees have the ability to select features on their own, experiments with C4.5 decision trees have shown that adding a random binary attribute to standard datasets impacts classification performance, causing it to weaken (generally by 5% to 10% in the situations tested) [7]. The reason for this is, at some point in the trees that are learned, the unrelated attribute is invariably chosen to branch on, producing random errors when test data is processed. Even though one may think that how can this happen when decision trees are cleverly designed to pick the best attribute for splitting at each node, the reason is as proceed further down the tree, very less data is available to support making the selection decision [7]. Thus the feature selection step will be done before training the C4.5 classifier. However, in the case of RandomForest (RF), it is different since at each step, voting mechanism is used after randomly choosing the splitting attributes. Thus feature selection is not needed for the RF since it automatically chooses the best features.

Due to the negative effect of irrelevant attributes on most of the machine learning algorithms, it is common to do the learning after a feature selection step that attempts to eradicate all but the most relevant attributes. The finest way to select relevant attributes is by manually, based on the understanding of the learning problem and what the features actually mean [7]. This is the reason that initial (pre) feature selection was done as described in previous sections. However, automatic methods also can be useful. Reducing the dimensionality of the dataset by removing unsuitable attributes improves the classification performance of learning algorithms. Furthermore, it also speeds them up, although this may be compensated by the computation involved in feature selection. More importantly, reducing the number of features yields a more easily interpretable representation of the target concept, focusing only on the most relevant attributes.

Fundamentally there are two types of feature selection algorithms; Filter and Wrapper methods. Due to the expensive computational time taken when wrapper methods are applied, only filter methods will be used in this study for the feature selection. Two popular filter feature selection methods (Chi squared attribute evaluation and Gain ratio attribute evaluation) were used with the Ranker search method which ranks the attributes by their importance. After applying these feature selection methods on the training data, the results of these two methods will be analyzed and the common features which have given lowest ranks in both two methods will be removed.

B. Classification Algorithms explored

1) *C4.5 Algorithm:* Among decision tree algorithms, C4.5 is the most commonly used algorithm. C4.5 was originally proposed as a successor of ID3 algorithm [8]. It is also capable of handling pruning, missing values and numeric values. At each node of the tree, C4.5 chooses the attributes of the data that most effectively splits the training dataset into subsets of one class or the other. The splitting criteria are based on the normalized information gain.

2) *Naïve Bayesian (NB) Algorithm:* Naïve Bayes is simple probabilistic classifier, based on applying Bayes Theorem which provides a way to calculate posterior probability P(Ci|X) using prior probability of class, prior probability of data and likelihood of the data given the class [2]. Naïve Bayes uses a strong assumption of conditional independence where it assumes attributes are independent from each other. Naïve Bayes can be trained very efficiently [9].

3) *AODE Algorithm:* Averaged one-dependence estimator (AODE) is a probabilistic classification learning technique. AODE was developed to solve the attribute independence problem of the famous Naive Bayes algorithm. Even though it usually develops considerably more accurate and reliable classifiers than Naive Bayes, its computational complexity is relatively high. Similar to NB, AODE also does not use tuneable parameters and does not perform model selection. As a result of that, AODE also has low variance. It supports incremental learning where the classifier can be updated efficiently with information from new examples as they become available [7]. Instead of predicting the single class, it predicts class probabilities, allowing the user to recognize the confidence which each classification can be made. Moreover, AODE can directly handle some situations where some data are missing.

C. Ensemble Algorithms explored

The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm for improving generalizability and robustness of a single estimator [2]. Usually ensemble methods work well with unstable base classifiers like decision trees and on stable classifiers like Bayesian classifiers, these methods do not work very well [7].

1) *Bagging:* Bagging produces multiple versions of the same predictor and combines the numerical prediction of these versions using plurality vote to identify the prediction class [9]. Multiple versions of the same base algorithm will be applied on each replicated bootstrap, where each bootstrap is created based on random sampling with replacement technique.

2) *Boosting:* Boosting change the weights for incorrectly classified data with the previous model used. Boosting involves incrementally building an ensemble by using all data to train each learner, but instances that were misclassified by the previous learners are given more weight so that subsequent learners give more focus to them during training [10]. In this study, AdaBoost.M1 algorithm will be used for Boosting.

3) *Random Forests:* Random Forests operate by building a multitude of decision trees at training time and outputting the class that is the mode of the classes (for classification prediction) or mean of the classes (for regression prediction) of the individual trees. Random Forests correct the overfitting problem of decision trees. The algorithm for inducing a Random Forest was developed by Breiman and Cutler and "RandomForests" is their trademark [11]. When splitting a node while constructing the tree, the split which is chosen is no longer the best split amid all features [11]. Instead, the best split among a random subset of the features is picked as the split. Because of the randomness, the bias of the RF usually increases a little relative to the bias of a single non-random tree [11]. But because of averaging, its variance also drops, typically more than compensating for the increase in bias, thus, yielding an overall enhanced model [2].

D. Overcoming class imbalance problem

Chawla states that “A dataset is imbalanced if the classification categories are not approximately equally represented” [12]. The problem with imbalanced data is that in classification problems with such data, the minority class instances are more likely to be misclassified than the majority class instances, due to the design principles of most machine learning algorithms.

In a literature review [13] done by Longadge, Dongre, and Malik, they have stated that according to the existing literature all the methods which can be used to rectify data imbalance problem can be categorized in to three main approaches; data pre-processing approach, the algorithmic approach and feature selection approach. In data pre-processing technique, sampling is applied on data in which either new samples are added or existing samples are removed. Algorithmic approach includes the cost-sensitive method, recognition-based approaches, and ensemble based approaches. The aim of cost sensitive classification is to minimize the cost of misclassification that can be realized by choosing the class with the minimum conditional risk. Since the cost of each class was not known at the learning time, cost sensitive approach will not be used to mitigate imbalance problem in this study. Feature selection methods are also vital since the data imbalance problem is commonly accompanied by the problem of high dimensionality of the data set.

To mitigate the data imbalance problem, this study proposes 4 approaches as mentioned below.

1) *Undersampling:* This approach attempts to balance the distribution of class by randomly removing a majority class sample. The biggest drawback with this method is loss of valuable information.

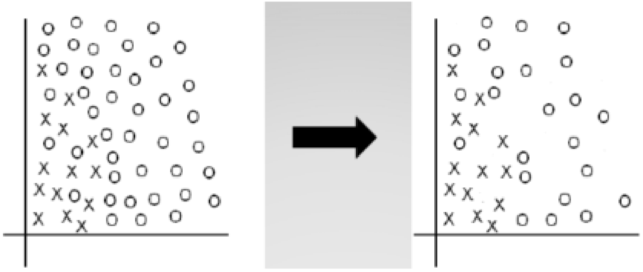


Fig 3 Overview of undersampling technique [13]

2) *Oversampling:* This approach attempts to balance the distribution of class by replicating minority class instances. The main drawback with this method is it can overfit the data.

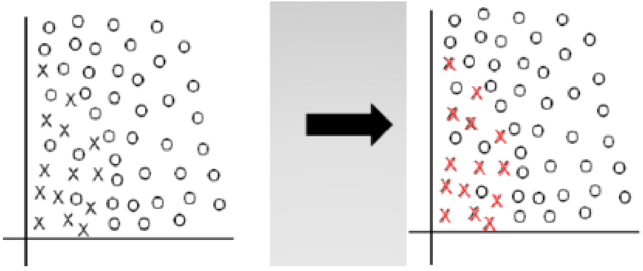


Fig 4 Overview of oversampling technique [13]

3) *Ensemble based approach:* Under this approach previously mentioned 3 ensemble algorithms will be used; Bagging, Boosting and RandomForest

4) *Hybrid approach:* This study proposes this approach which combines the aforementioned two methods; sampling technique and ensemble based technique. In this proposed hybrid approach, first, sampling technique will be applied to the training dataset and then generated new training dataset will be fed to the base algorithms along with the previously mentioned ensemble methods.

IV. Experimental Setup

A. Dataset

This study was carried out using a secondary dataset which included 15,726 records of detailed information about the state university graduands who were graduated in year 2014. This data was gathered for the 'Graduand Employment Census - 2014', under the HETC project with the guidance of the Ministry of Higher Education.

Originally, data set contained 82 variables. Table I shows chosen variables in this research after initial-feature selection.

doi: 10.4038/icter.v9i2.7181

TABLE I
INITIALLY SELECTED VARIABLE SET

No	Variable name	Variable Description
1	University	Name of the university
2	Gender	Gender of the graduand
3	Ethnicity	Ethnicity of the graduand
4	Faculty	Faculty the graduand studied in
5	Degree Type	Type of the degree
6	Stream	Stream of the degree specialized in
7	Medium	Medium of the studying degree
8	Class	Received class from the degree
9	English Proficiency (Written)	Written English skills of the graduand
10	English Proficiency (Oral)	Oral English skills of the graduand
11	O/L English results	O/L results for English subject
13	Browsing web	Whether graduand has ever browse web
15	Using office-packages	Whether graduand can use the office packages well
16	Writing Programs	Whether graduand can wrote computer programs
17	Extra activities	Whether graduand has done extra-curricular activities in university life
18	Extra activities : Detail	If response is yes, for the above variable, extra activities he has done
19	Vocational activities	Whether graduand has done vocational activities in university life
20	Vocational activities : Detail	If response is yes for the above variable, vocational activities he has done
21	Other education	Whether graduand has additional educational qualifications
22	Other education : Detail	If response is yes for the above variable, such qualifications he has
23	Lived Area	Type of area which the graduand has lived in most of his/her life
24	District	District which the graduand has lived in most of his/her life
25	GCE A/L	The type of school graduand have attended for GCE A/L
26	Parents' education	Highest level of education achieved by either graduands' father or mother
27	Expected salary	Expected salary of the graduand
28	Expected sector	Expected job sector of the graduands
29	Employment status	The employment status of the graduand
30	Employed sector	Employed job sector
31	Position hold	Position hold in the job
32	Economic sector	Economic sector of the job
33	Actual salary	Actual job salary
34	Job relevance	Whether job is related with the degree of the graduand

Three out of four target variables are highlighted in Table I in blue colour while the other target variable is a combination of two variables which are highlighted in green colour. Furthermore, variables which are numbered after 29 are only relevant for the graduands who have already employed at the time when this census was carried out. Moreover, the attributes which are coloured in yellow were concatenated into one single variable named computer literacy since all these three variables are binary.

Table II shows the final feature sets after applying all the pre-processing steps and feature selection algorithms for all four modules. However as mentioned in previous sections, for RF algorithm, the full datasets (module 1 – 22 features, other modules - 23 features each) will be used since RF itself can choose the relevant attributes very well.

TABLE II
SUMMARY OF FEATURES SELECTED IN ALL 4 MODULES

Rank	Module 1 : Employment Status Prediction	Module 2 : Job Salary Prediction	Module 3: Job Field Prediction	Module 4 : Job Relevance Prediction
1	Medium	Discipline	Discipline	Medium
2	Discipline	Faculty	Faculty	Discipline
3	Gender	Medium	Degree Type	Faculty
4	Faculty	Gender	Employed Sector	Stream
5	Stream	Expected Salary	Stream	Employed Sector
6	Degree Type	Stream	Medium	Degree Type
7	Lived Area	Degree Type	Vocational Training	Expected Salary
8	GCE O/L English	Employed Sector	Preferred Sector	English Proficiency(Oral)
9	Expected Salary	University	University	Lived Area
10	University	Lived Area	Computer Literacy	GCE O/L English
11	Preferred Sector	Computer Literacy	Professional Education	English Proficiency(Written)
12	Vocational Training	Vocational Training	Gender	University
13	English Proficiency (Oral)	Preferred Sector	Lived Area	School - GCE A/L
14	School - GCE A/L	GCE O/L English	Expected Salary	Parents Education
15	Parents Education		Class	Class
16	District			Computer Literacy
17				District

B. *Evaluation Procedure*

In this study, Nested Stratified K-Fold Cross Validation is used to evaluate the algorithms explored. In this approach, both the parameter optimization and the evaluation of the algorithm are done together. In order to understand this complex evaluation technique more clearly, first K-Fold Cross Validation is explained and then 'stratified' version of cross validation will be explained. 'Nested' stratified k-fold cross validation will be explained afterwards.

In the usual k-fold Cross Validation procedure, the data set is randomly split into k mutually exclusive subsets (the folds) of approximately equal size [7]. Then the model is learned using (k-1) folds, and the fold left out is used for testing. This process is then repeated k times, with each subsample is used exactly once as the validation data. Finally, the average value computed in the loop will be the performance measure reported by k-fold Cross Validation.

Stratified k-Fold is a variation of k-fold which returns stratified folds: each set contains approximately the same percentage of samples of each target class as the complete set. This approach can be computationally bit expensive, but without wasting much data, it is known to produce reliable results [7].

Nested variation of Stratified k-Fold Cross Validation comes into play when one needs to tune the parameters. As the name suggest, in nested variation, there is an outer cross validation as well as an inner cross validation as shown in Figure 5. Nested k-fold cross validation encapsulates one layer of cross-validation inside another one. The inner layer is used to try out different parameters and pick the ones that work best for the given distribution while the outer layer is used to evaluate the best parameters found in the inner layer [14].

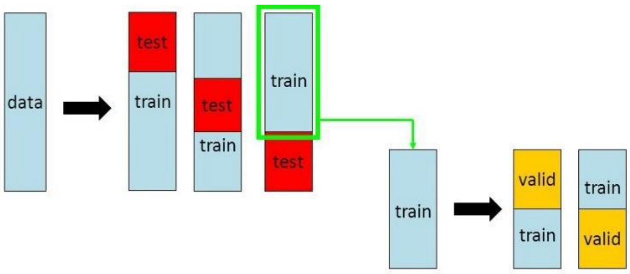


Fig 5 Inner and outer cross validations [14]

In nested variation, there is an outer cross validation as well as an inner cross validation. Nested k-fold cross validation encapsulates one layer of cross-validation inside another one. The inner layer is used to try out different parameters and pick the ones that work best for the given distribution while the outer layer is used to evaluate the best parameters found in the inner layer [14]. In short, the inner cross-validation is only used to find the "optimal" parameter settings by finding those settings that maximize estimated predictive performance in the inner cross-validation. Once those settings have been found, the model is rebuilt with those settings from the full training set (i.e. the particular training set of the outer cross-validation) and that single model is used for prediction.

This Nested Stratified k-fold Cross Validation method is quite powerful for detecting over fitting and estimating the generalization error conservatively [14]. Hence in this study, Nested Stratified 10-fold Cross Validation was used. The reason for choosing k=10 is that it is empirically proven that for majority of the datasets, 10-fold schema gives better training model with lesser possibility of having over-fitting scenarios [7].

C. Setting Up The Parameters

Prior to process of applying base classification algorithms, it is essential to find the optimal parameters of those algorithms that suits the training datasets. Here, the evaluation procedure (nested stratified cross validation) described in above section will be used to evaluate the parameter configurations. Since there are four modules, each consist of training with all aforementioned classification algorithms, the general procedure of setting up those parameters for a single module will be discussed, since the method of tuning parameters for each module is similar.

1) *Tuning C4.5 Decision Tree Algorithm:* For C4.5 algorithm two parameters out of the six parameters given in Table III were considered to tune, while keeping the values of other four parameters to the recommended values. The parameters tuned are confidence factor (C) and minimum number of instances per leaf (M). Root Mean Squared Error (RMSE) was used to tune the parameters in C4.5 (J48) algorithm. Smaller the error, better the classifier will be. We choose the optimal parameter pair from the all possible parameter pairs where the value of C goes from 0.05 to 1 by 20 steps and M value goes from 2 to 3 by 2 steps. i.e we chose the optimal values from 40 (C, M) pairs; (0.05, 2), (0.10, 2), … , (1, 2), (0.05, 3), (0.10, 3), … , (1, 3) by running nested cross validation.

TABLE III
PARAMETER CONFIGURATION FOR C4.5 ALGORITHM

Parameter description	Recommended default value
Confidence factor used for pruning (C)	0.25
Minimum number of instances per leaf (M)	2
Whether reduced-error pruning is used instead of C.4.5 pruning.	False
Whether to consider the subtree raising operation when pruning.	True
Whether pruning is performed.	False
Whether counts at leaves are smoothed based on Laplace.	False

2) *Tuning RandomForest Algorithm:* For this algorithm two parameters out of the three parameters given in Table IV were considered to tune, while keeping the values of other four parameters to the recommended values. The parameters tuned are the number of features to be used and the number of trees to be generated.

TABLE IV
PARAMETER CONFIGURATION FOR RANDOMFOREST

Parameter description	Default Value
The maximum depth of the trees	Unlimited
The number of attributes to be used in random selection	$\log_2$(numOfattributes)+1
The number of trees to be generated.	100

Here in number of trees parameter, the larger the better, but also the longer it will take to compute. In addition, note that results will stop getting significantly better beyond a critical number of trees. Thus tuning this parameter will make the computation time less without affecting the accuracy. Therefor we choose the values from 80 to 200 by 13 steps as the number of trees parameter values to be optimized (i.e, 80, 90, 100, …, 190, 200). The number of features value is the size of the random subsets of features to consider when splitting a node. The lower, the greater the reduction of variance, but also the greater the increase in bias. For classification tasks, empirical good value for this parameter is the square root of number of features used [11].

doi: 10.4038/icter.v9i2.7181

However Breiman, the co-developer of RandomForest algorithm, has said that it is better to try up the half and the twice of the square root of number of features also, to find the most optimal parameter [11]. Therefor we choose the values 5 (since there are 22 features for the first module and $\sqrt{22}=4.7$), 2 and 10 as the number of features parameter values to be optimized. Thus taking all the combination of aforementioned values for the number of trees and number of features, 13 x 3 pairs will be tested to find the most optimal pair.

3) *Tuning other algorithms:* For AODE, Naïve Bayesian, Bagging and Boosting algorithms there are no concrete parameters to tune. Yet for Bagging and Boosting algorithms, number of iterations to be performed have to be assigned. Therefore for Bagging, 50 iterations were assigned according to Breiman's recommendations [9] and for Boosting, 100 iterations were assigned. However, in Boosting it will stop at less than 100 iterations if the necessary goal is reached. Furthermore the size of each bag has to be specified, as a percentage of the training set size for Bagging algorithms. So 100 was specified for this option since the bag size needs to be the same as the size of training set.

V. Results

For each of the four modules, four approaches will be considered as depicted in Figure 6. First approach is the traditional approach of classification, which is, applying a single classification algorithm (NB, AODE and C4.5 Decision Tree) for the original dataset and then, according to the results of the first approach, second, third and fourth approaches will be considered respectively. Second approach is the use of multiple classifiers (Bagging, Boosting and RandomForest) in order to increase the accuracy as well as to overcome the data imbalance problem if exists, whereas the third approach is the use of sampling techniques (oversampling and undersampling techniques) for the training data set in order to overcome the data imbalance problem if the problem of data imbalance exists. Final and the newest approach proposed by this research is the use of hybrid method which combines the second and third approaches.

Results of model evaluation for first module (employment status prediction) are extensively described below and a summary of the results of the other 3 modules will be given in conclusion section.

Figure 7 shows the dispersion of target class (employment status). It signifies the fact that there is a higher probability of having class imbalance problem in training datasets (since employed: unemployed: underemployed class ratio is close to 6:3:1). In subsequent sections, how this problem has affected the prediction outcomes of each class will be analysed.

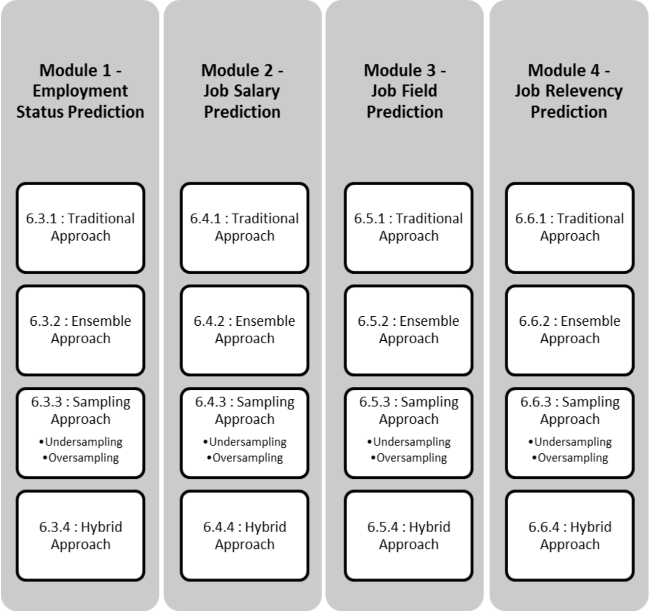

Module 1 - Employment Status Prediction
6.3.1 : Traditional Approach
6.3.2 : Ensemble Approach
6.3.3 : Sampling Approach
•Undersampling
•Oversampling
6.3.4 : Hybrid Approach
Module 2 - Job Salary Prediction
6.4.1 : Traditional Approach
6.4.2 : Ensemble Approach
6.4.3 : Sampling Approach
•Undersampling
•Oversampling
6.4.4 : Hybrid Approach
Module 3 - Job Field Prediction
6.5.1 : Traditional Approach
6.5.2 : Ensemble Approach
6.5.3 : Sampling Approach
•Undersampling
•Oversampling
6.5.4 : Hybrid Approach
Module 4 - Job Relevency Prediction
6.6.1 : Traditional Approach
6.6.2 : Ensemble Approach
6.6.3 : Sampling Approach
•Undersampling
•Oversampling
6.6.4 : Hybrid Approach

Fig 6 Overview of evaluation procedure

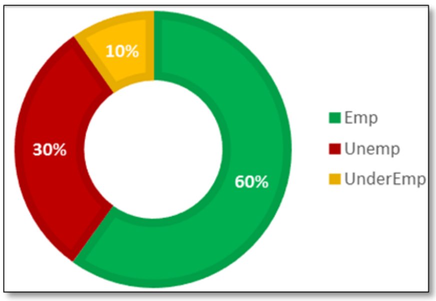


Fig 7 Dispersion of Employment status classes

A. *Traditional Approach – applying single algorithms*

The three diagrams shown in Figure 8(a),(b),(c) are the colour coded confusion matrices for each 3 classifier models. The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabelled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions (in the colour coded matrices, more blue implies higher number of correct classifications while less blue implies lesser number of correct classifications). Thus, from Figure 8 it can be seen that in all 3 classifiers, majority class (upper row of matrices) has performed really well compared to other two classes while the minority class (last row of matrices) has performed poorer than the other two classes, illustrating the symptoms of data imbalance phenomena.

1) *Individual Evaluation Measures:* When the class wise performance measures were analyzed by individual evaluation measures illustrated in Table V, it can be seen that the prediction outcomes of minority class ('UnderEmp' class) has given unsatisfactory results in all three classifier models. Yet, the evaluation measure, accuracy, which is insensitive to class imbalance phenomena, shows the best results on minority class. Evaluation measures which are sensitive to class imbalance phenomena, such as Precision and Recall gives reflective figures to notify that the class imbalance problem exists in the dataset.

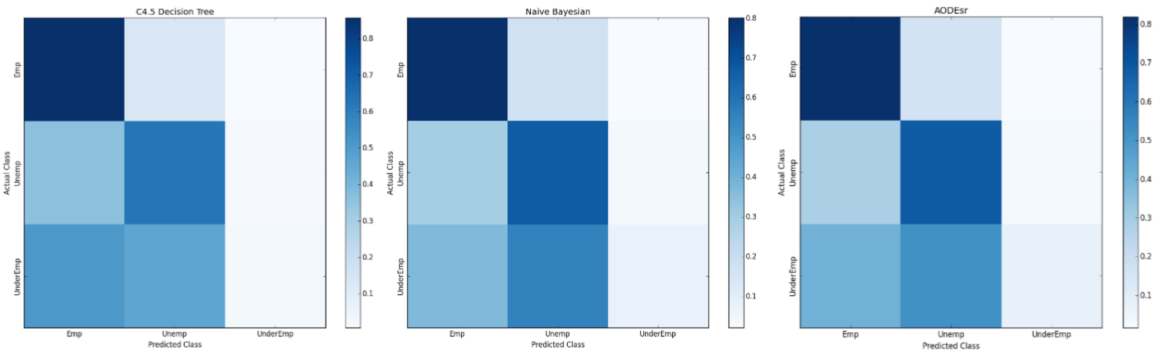


Fig 8(a),(b),(c) : Confusion Matrices for C4.5, NB and AODE (Traditional)

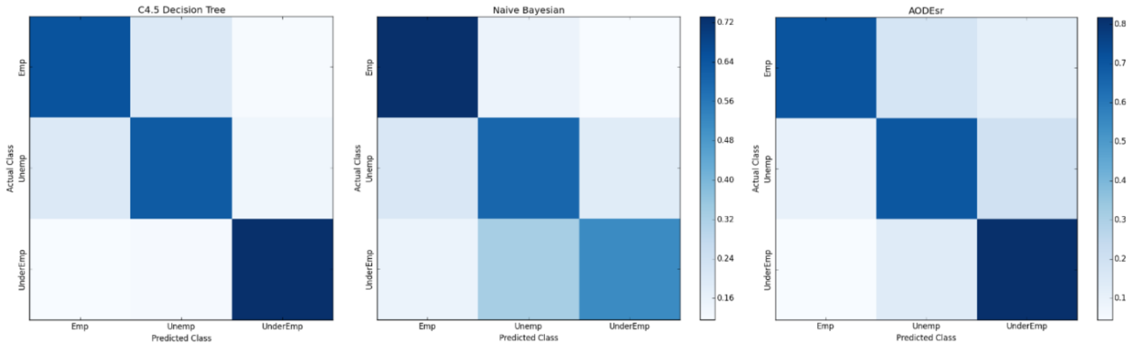


Fig 10(a),(b),(c) : Confusion Matrices for C4.5, NB and AODE (Oversampled)

TABLE V

PREDICTION PERFORMANCE ON INDIVIDUAL MEASURES
(E-EMPLOYED, N-UNEMPLOYED, U- UNDEREMPLOYED)

Class	Accuracy			Precision			Recall		
	C4.5	NB	AODE	C4.5	NB	AODE	C4.5	NB	AODE
E	0.7	0.7	0.8	0.76	0.79	0.80	0.86	0.80	0.82
N	0.8	0.7	0.7	0.60	0.56	0.59	0.63	0.67	0.69
U	0.9	0.9	0.9	**0.21**	**0.24**	**0.29**	**0.02**	**0.08**	**0.08**
All	**0.7**	**0.7**	**0.7**	**0.66**	**0.67**	**0.68**	**0.71**	**0.69**	**0.71**

2) *Combined and Graphical Evaluation Measures:* Since F-measure and G-means formulated based on the combination of Precision, Recall, and Specificity, impact from insensitive evaluation measures are abolished by these combined evaluation measures. Thus both F-measure and G-means of all three classifier models give unsatisfactory results on minority class. ROC AUC gives satisfactory results on minority class since it could not capture the impact from class imbalance problem.

TABLE VI

PREDICTION PERFORMANCE ON COMBINED MEASURES
(E-EMPLOYED, N-UNEMPLOYED, U- UNDEREMPLOYED)

Class	Accuracy			Precision			Recall		
	C4.5	NB	AODE	C4.5	NB	AODE	C4.5	NB	AODE
E	0.8	0.8	0.8	0.72	0.74	0.75	0.81	0.83	0.84
N	0.6	0.6	0.6	0.72	0.72	0.74	0.79	0.81	0.82
U	**0.1**	**0.1**	**0.1**	**0.15**	**0.27**	**0.27**	0.68	0.72	0.73
All	**0.7**	**0.7**	**0.7**	**0.71**	**0.72**	**0.73**	**0.79**	**0.81**	**0.83**

B. Ensemble Algorithm Approach

According to the results given by both Bagging and Boosting methods, it is evident that the original classification performance of the minority class have not improved as expected. Results are similar to the values retrieved from traditional approach. Figure 9 graphically summarize the effect of applying Bagging technique on the training dataset.

It is clear that applying Bagging technique directly on training dataset couldn't effectively increase the classification performance of the minority class significantly. Even though the F-measure and G-means have been slightly increased in C4.5 classifier and AODE classifier, that difference is not significant enough. Moreover Boosting shows similar results.

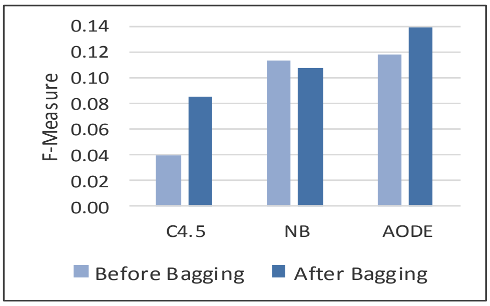


Fig 9 F-measure on minority class before and after Bagging

C. Sampling Approach

In oversampling, minority classes' instances will be replicated until original class ratio Emp: Unemp: UnderEmp reaches from 6: 3: 1 to 1: 1: 1. The 3 diagrams in Figure 10 show the confusion matrices for three classifier models separately. From the diagonals, it can be seen that in all 3 classifiers, even though majority class has performed really well compared to other two classes, the performance of other two classes are not significantly low compared to the majority class, indicating no symptoms of data imbalance phenomena anymore. Undersampling shows similar results.

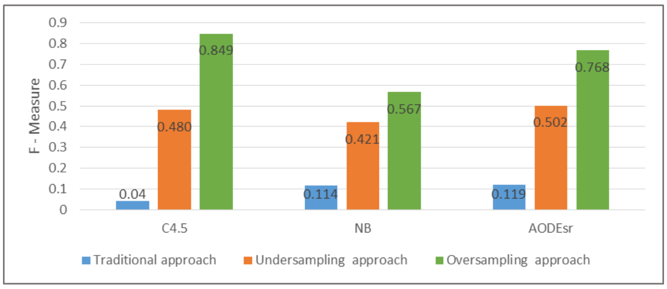


Fig 11 Performance comparison between 3 approaches (minority class)

Figure 11 attempts to summarize the F-measure results derived from traditional approach, undersampling approach and oversampling approach, with respect to classification performance of minority class. When the approaches given in this figure were compared, it's quite apparent that performance measures derived from oversampling have surpassed the performance results derived from undersampling in minority classes.

D. *Hybrid Approach*

From sampling approaches it is already shown that oversampling technique has surpassed the results of undersampling. Hence in this hybrid approach, oversampling was used as the sampling technique to apply on Ensemble approaches.

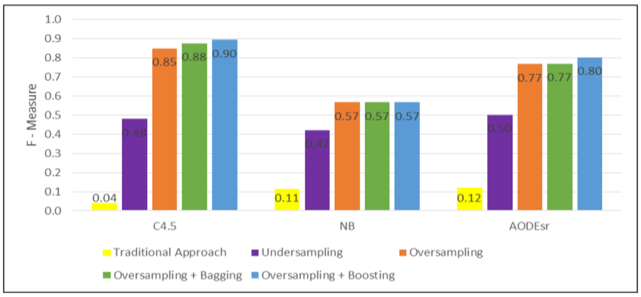


Fig 12 F-measure comparison on minority class in all 5 approaches

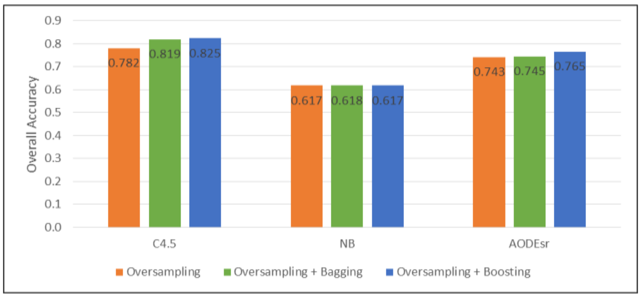


Fig 13 Overall accuracies among 3 approaches related to oversampling

When the Figure 12 is analyzed, it's quite apparent that the potential hybrid approach hasn't significantly augmented the results of oversampling. Although it has surpassed the evaluation measures initially acquired through the traditional approach significantly, Bagging or Boosting haven't add a significant benefit to original oversampling results for minority classes. Hence it can be concluded that the improvement on this minority class classification embedded in this hybrid approach merely contributed from the oversampling setting.

From Figure 13, it can be seen that both hybrid approaches (i.e. Oversampling+Bagging and Oversampling+Boosting) have slightly improved the overall accuracy of sole oversampling approach in both C4.5 classifier and AODE classifier, while for Naïve Bayesian classifier the accuracy have not been changed. The possible reasoning behind this might be that stable classifiers like NB do not respond well for bagging and boosting.

TABLE VII
PREDICTION PERFORMANCE ON INDIVIDUAL MEASURES (OVERALL)

Class	Bagging			Boosting			RF
	C4.5	NB	AODE	C4.5	NB	AODE	
Accuracy	0.82	0.62	0.75	0.83	0.62	0.77	**0.85**
AUC	0.94	0.80	0.90	0.94	0.75	0.88	**0.96**
G-means	0.86	0.71	0.81	0.87	0.71	0.82	**0.89**
F-measure	0.75	0.58	0.70	0.76	0.58	0.70	**0.79**

As previously shown, hybrid approach has recorded the best performance out of all four approaches and from Table VII, it can be seen that, RF ensemble method has shown the highest classification performance when considering all four measures under hybrid approach. Thus it can be concluded that for predicting the employment status using this dataset, RF with hybrid approach is the best technique. The order of other classifiers using the accuracy measure under the hybrid approach is C4.5+Boosting, C4.5+Bagging, AODE+Boosting, AODE+Bagging and lowest performance from NB.

VI. DISCUSSION

In first, third and fourth modules (employment status prediction, job field prediction, job relevance prediction), at the beginning (i.e. in our traditional approach), any of the classifier algorithms did not produce very good results with respect to the minority class due to the class-imbalance problem. Thus we experimented few mechanisms to improve this classification performance of minority class while also increasing the overall classification performance. For all these 3 modules, the order of classification performance with respect to different approaches is shown below in decreasing order.

1. Hybrid approach (Oversampling + Ensemble learning)
2. Oversampling
3. Undersampling
4. Ensemble learning
5. Traditional approach

Even though in these three modules, hybrid approach has shown the highest performance, hybrid approach hasn't significantly augmented the results of oversampling. Hence we can conclude that the improvement on this minority class classification embedded in this hybrid approach merely contributed from the oversampling setting and furthermore ensembling has not added a significant benefit to original oversampling results.

When considering the second module (job salary prediction), the results are bit different. Even though there was a class imbalance in this dataset as well, this has not affected the classification performance of minority class. Hence the imbalance data has not been a problem to the classification of minority class. Thus we did not carry out sampling approach and hybrid approach. But we carried out ensemble approach only to improve the overall classification performance, and we showed that applying ensemble method has slightly increased the performance of the traditional approach.

When we consider the traditional approach of three modules which had the class imbalance problem, in all three modules, C4.5 decision tree classifier has shown the poorest performance on minority class while AODE has shown the highest performance in minority class as well as in overall classification as well.

Furthermore when we compare the results under ensemble approach, in all three modules, ensemble approach has not been able to mitigate the class-imbalance problem. But this ensemble method has significantly increase the minority class performance in C4.5 classifier. Yet this increase has not been significant enough to rectify class-imbalance problem. Moreover in some of these 3 modules, ensemble approach has slightly decreased the minority class performance of AODE and NB classifiers.

Both undersampling approach and oversampling approach has significantly enriched the classification performance of minority class in all these 3 modules. But the difference is while enhancing the minority class performance, undersampling has decreased the majority class performance significantly while oversampling has been managed to increase or keep the traditional majority class performance as it is. Thus it can be concluded that in all three modules, classification performance of oversampling has surpassed the results of undersampling. It’s also observed that after applying any of these sampling techniques C4.5 classifier has achieved the highest percentage uplift on their evaluation measures with respect to minority class.

As mentioned before applying hybrid technique in all these 3 modules has slightly increased the performance of minority class when compared to oversampling technique in C4.5 classifier. Yet for AODE and especially for NB classifier applying bagging or boosting has not changed the classification performance at all in 2 of 3 modules. Moreover it is perceived that in C4.5 and AODE classifiers, minority class performance has been increased more in oversampling with boosting technique compared to oversampling with bagging technique. But when we consider the overall performance in NB and AODE classifiers oversampling with boosting has decreased the classification performance than the original oversampling approach. The most probable reason for these unchanged or decreased performance of NB and AODE classifiers after applying an ensembling method is ensemble methods usually works best at unstable classifiers (like decision trees). For stable classifiers like NB and AODE ensemble methods will not do a much.

When we consider the RandomForest method, which is a decision tree based ensemble method, the hybrid approach has been able to significantly increase the performance of in all these three modules.

VII. Conclusions and Future Work

Table VIII gives the summary of best models selected under each of the four modules with the different combined and graphical evaluation measures. A significant result which can be seen from the following table is in all 4 modules, RandomForest based approaches have been selected as the best model and apart from job salary prediction module, best models selected in all the modules have been able to take AUC values greater than 90% indicating that all of them are ‘Excellent’ experiments according to AUC interpretation.

From this research we have proved that applying a machine learning approach to predict employability is a plausible option, given that the constraints embedded in employability data (like class imbalance) is properly handled.

Another major objective of this research was identifying the important factors relevant to each of four modules and Table 2 shows the factors which are relevant to these modules separately, according to its importance.

TABLE VIII
Summary of best models for 4 modules

Module	**Employment Status Prediction**	**Job Salary Prediction**	**Job Field Prediction**	**Job Relevance Prediction**
Approach	Hybrid	Ensemble	Hybrid	Hybrid
Classifier	RF with oversampling	C4.5 after Bagging	RF with oversampling	RF with oversampling
ROC AUC	95.7%	80.6%	95.9%	98.2%
Accuracy	85.0%	58.4%	78.6%	93.0%
F-measure	78.7%	58.0%	78.7%	89.3%
G-Means	88.7%	69.7%	87.0%	93.9%
Kappa statistic	77.0%	42.0%	75.0%	88.0%
AUC interpretation	Excellent experiment	Very Good experiment	Excellent experiment	Excellent experiment

The tangible benefits derived from these four types of prediction models can be revealed by implementing these selected best four models in a web system so that current undergraduates can use this system to predict their future related to employability and enhance their skills before they complete the degree, until this system predicts their desired employment status, job field and job salary.

Even though we have been able to achieve very good results on first, third and fourth modules, classification performance is comparatively less in second module (i.e. salary prediction). Thus, it is better to consider other machine learning algorithms such as Neural Network methods, SVM, CART, Bayesian networks, etc. for the prediction of job salary in order to see whether a different algorithm could increase this performance.

Cost sensitivity learning was not carried out when trying to overcome the class imbalance problem since we did not know the costs of misclassification in each class at the learning time. However as a future work, if cost of misclassification for each class can be defined, cost sensitivity learning and cost curve may do even more better classification than the class imbalance mitigation techniques we applied. Furthermore in model evaluations we only compared the AUC of ROC curves only. However we explained that ROC AUC is not a measure which is sensitive to class imbalance problem. Even though we had G-means and F-measure which are sensitive to class imbalance phenomena, as a future work it is suggested to compare the results using AUC of PR (Precision-Recall) curve as well since it is one of the best measures to evaluate imbalanced data.

ACKNOWLEDGMENT

We would like to thank Dr. Asoka Ramanayake and the team at HETC for providing their graduand employment census 2014 dataset for this research work.

REFERENCES

[1] A. G. W. Nanayakkara. *Employment and Unemployment in Sri Lanka: Trends, Issues, and Options*. Department of Census and Statistics, 2004.

[2] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. “The elements of statistical learning: data mining, inference and prediction.” *The Mathematical Intelligencer*, vol. 27, pp. 83-85, 2005.

[3] B. Jantawan, and C. Tsai. “The Application of Data Mining to Build Classification Model for Predicting Graduate Employment.” *International Journal of Computer Science and Information Security*, vol. 11, pp.1-8, Oct. 2013.

[4] B. Jantawan, and C. Tsai. “A Classification Model on Graduate Employability Using Bayesian Approaches : A Comparison.” *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 6, pp. 4584–4588, 2014.

[5] M. A. Sapaat, A. Mustapha, J. Ahmad, and K. Chamili. “A Data Mining Approach to Construct Graduates Employability Model in Malaysia.” *International Journal on New Computer Architectures and Their Applications*, vol. 1, pp. 1111–1124, 2011.

[6] R. S. J. D. Baker, and K. Yacef. “The State of Educational Data Mining in 2009 : A Review and Future Visions.” *Journal of Educational Data Mining*, vol. 1, pp. 3-16, 2009.

[7] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical machine learning tools and techniques*, 2005.

[8] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali. “A comparative study of decision tree ID3 and C4.5.” *International Journal of Advanced Computer Science and Applications*, vol. 4, pp. 13–19, 2014.

[9] L. Breiman. “Bagging Predictors.” *International Conference on Machine Learning*, vol. 24, pp. 123–140, 1996.

[10] Y. Freund, R. E. Schapire, and M. Hill. “Experiments with a New Boosting Algorithm,” in *International Conference on Machine Learning*, pp. 148–156, 1996.

[11] L. Breiman. “Random Forests.” *International Conference on Machine Learning*, vol. 45, pp. 5–32, 2001.

[12] N. V. Chawla. “Data Mining for Imbalanced Datasets: An Overview.” *Data Mining and Knowledge Discovery Handbook*, pp. 853–867, 2005.

[13] R. Longadge, S S. Dongre, and L. Malik. “Class Imbalance Problem in Data Mining : Review” *International Journal of Computer Science and Network (IJCSN)*, vol. 2, 2013.

[14] C. F. Aliferis, A. Statnikov, and I. Tsamardinos. “Challenges in the Analysis of Mass-Throughput Data : A Technical Commentary from the Statistical Machine Learning Perspective.” *Cancer Informatics,*, vol. 2, 2006.

Motion Tracking by Sensors for Real-time Human Skeleton Animation

S.P. KasthuriArachchi, Chengkai Xiang, W.G.C.W. Kumara, Shih-Jung Wu, Timothy K. Shih

***Abstract*—Human Computer Interaction based research has emerged in the early 1980s with the advent of computer technology. Human Motion Capture is the process of recording the movement of people. Among many kinds of human motion capture devises, Microsoft Kinect sensor and inertial sensors are most popular nowadays. In this paper we propose an efficient motion tracking mechanism to construct real time human skeleton animation using inertial sensors. We compare the results of our proposed method with the Microsoft Kinect sensor over the complicated motion tracking and joint position. During the experiment we observed that our results are much steady than Microsoft Kinect results. Some motions like hand cross over or leg cross over, our method showed better results than Kinect because the Kinect may lose skeleton of the blocked parts. On the other hand, since we use radio frequency inertial sensors, our method has a larger working area than Kinect.**

***Keywords*—Human Motion Capture (HMC), inertial sensors, motion tacking, skeleton animation.**

I. Introduction

Human Computer Interaction (HCI) is an interaction interface between human and computers. With the growth of computing capability and the maturity of development of electronic devices, HCI has gained much attention in research and industrial fields and became extensively applied in sports, special effects, user interface, training, rehabilitation and computer animation for movies, and video games. Hadjidj et al. used wireless sensor networks for rehabilitation [1]. The data-driven system introduced by Kurakin et al. is capable of automatic hand gesture recognition in real-time using a commodity depth camera [2]. During the last decade, micro sensors have proven to be a good alternative to traditional optical motion capture system, because their low-cost and self-contained nature. Real-time inertial tracking of human motion requires attaching inertial sensors to the major segments of human body. Shiratori et al. presented the theory and practice of using body-mounted cameras to reconstruct the motion of a subject and show results in settings where capture would be difficult or impossible with traditional motion capture systems [3].

Slyper and Hodgins created a performance animation system that leverages the power of low-cost accelerometers [4]. Even though their setup based on upper body suit, another study contributed their research by introducing a novel framework for generating full-body animations controlled by only four 3D accelerometers that are attached to the extremities of a human actor [5]. Favre et al.has proposed a new calibration procedure adapted for the joint coordinate system (JCS), which required only inertial measurement units (IMUs) data[6]. Another study presented an Extended Kalman Filter for fusion of inertial and magnetic sensing that is used to estimate relative positions and orientations [7]. The study of Altun et al. provided different techniques of classifying human activities that are performed using body-worn miniature inertial and magnetic sensors [8].

In this paper, we introduce a wearable real-time human motion capture system using inertial sensors. A Kalman filter is applied to integrate the output of sensors data. Then we compare our position and tracking information with Microsoft Kinect sensor.

The rest of this paper is organized as follows. In Section II related work in inertial sensor human motion tracking and gait analysis is discussed. The proposed method is explained in section III. Section IVdiscusses results comparing with Kinect. Conclusion and future work is presented insection V.

II. Related Work

A. *Inertial Sensor Based Motion Capture*

Inertial sensors can offer an accurate and reliable method to study human motion, however the degree of accuracy and reliability is site and task specific [9].Most inertial systems use gyroscopes to measure rotational rates. Thus, inertial tracking system has attracted many interests [9-11]. Raptis et al. present a real-time gesture classification system for skeleton wireframe motion. Its key components include the design of an angular representation of the skeleton [12].

B. *Kinematics*

We used Euler angle to represent each segments rotation of a human skeleton. Human body is comprised of 14 segments, linked by 15 joints. We considered rotation information to determine the position and orientation of joints. For the relative transfer movement between father and child joint, we have used forward kinematics and inverse kinematics as shown in Figure 1. If used only forward kinematics, the model is fixed at the centre joint of the human model. Joint position of father can be calculated

Manuscript received on 04 Oct. 2016. Recommended by Prof. G.N. Wikramanayake on 29 Dec. 2016.

S.P. Kasthuri Arachchi is currently reading for the Ph.D. at the Department of Computer Science and Information Engineering, National Central University, Taiwan (R.O.C.).(*sandelik@gmail.com*).
Chengkai Xiang obtained his Masters from Department of Computer Science and Information Engineering, National Central University, Taiwan (R.O.C.).(*xiangchk08@gmail.com*).
W.G.C.W. Kumara is currently reading for the Ph.D. at the Department of Computer Science and Information Engineering, National Central University, Taiwan (R.O.C.). (*chinthakawk@gmail.com*).
Shih-Jung Wuis is an Associate Professor of the College of Global Development, Department of Innovative Information and Technology, Tamkang University, Lanyang Campus, Taiwan (R.O.C.). (*wushihjung@mail.tku.edu.tw*).
Timothy K. Shih is a Professor at the National Central University, Taiwan (R.O.C.). (*timothykshih@gmail.com*).

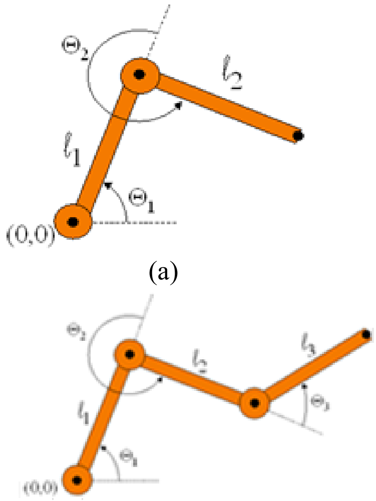
doi:10.4038/icter.v9i2.7180

using position of child joints and oriented vector in inverse kinematics.

The kinematics equations for the series chain of a robot are obtained using a rigid transformation [*Z*] to characterize the relative movement allowed at each joint and separate rigid transformation [*X*] to define the dimensions of each link [13]. For a serial open chain, the result is a sequence of rigid transformations alternating joint and link transformations from the base of the chain to its end link. A chain of *n* links connected in series has the kinematic equations,

$$[T]=[Z_1][X_1][Z_2][X_2]\ldots[Z_n][X_n] \quad (1)$$

where [*T*] is the transformation locating the end-link. Notice that the chain includes a 0th link consisting of the ground frame to which it is attached. These equations are called the forward kinematics equations of the serial chain [13].



(a)

(b)

Fig. 1Kinematics[14] (a) Forward kinematics, (b) inverse kinematics

C. Kalman Filter

The algorithm works under two steps and in the prediction step, the Kalman filter produces estimates of the current state variables, along with their uncertainties. Once the outcome of the next measurement is observed, these estimates are updated using a weighted average The objective of Kalman filter is to estimate the state of a linear system, by assuming the true state at time *k* is evolved from the state at (*k*-1) according to,

$$X_k = F_{kX_{k-1}} + B_{kU_k} + w_k \quad (2)$$

where, F_k is the state transition model which is applied to the previous state $X_{k-1}$and input values to the new state value .In our experiments, we have set this value as 1**.**

$B_k$is the control input model which is applied to the control vector and U_k, w_k is the process noise which is assumed to be drawn from a zero mean multivariate normal distribution with covariance Q_k;w_k~*N(0,*Q_k*)*[15]. At time *k* an observation (or measurement) Z_k of the true state X_k is made according to,

$$z_k = H_{kX_k} + V_k, \quad (3)$$

where,H_k is the observation model which maps the true state space into the observed space and V_k is the observation noise which is assumed to be zero mean Gaussian white noise with covariance R_k;V_k~*N(0,*R_k*)* [15]. The initial state, and the noise vectors at each step $\{x_0, w_1, ..., w_k, v_1, ..., v_k\}$ are all assumed to be mutually independent [15].

$$K_k = P_k H_k^T (H P_k H_k^T)^{-1} \quad (4)$$

$$P_k = (1 - K_k H_k) P_k \quad . \quad (5)$$

The Kalman filter produces an optimal state estimate by recursively updating the system state and the estimation error covariance, P_k. This estimation error covariance is used for calculating the optimal Kalman gain, K_k. That has been used in (4), estimating the further state with input data. *R* is a covariance which can be define by user. Finally updated state estimate P_k is again using updated K_k[15].

III. PROPOSED METHOD

A. System Overview

The system use sensors to capture motion information of human. First, denoising and processing is conducted to correct the errors and compensate for the disturbances and then display the real-time motion of human.

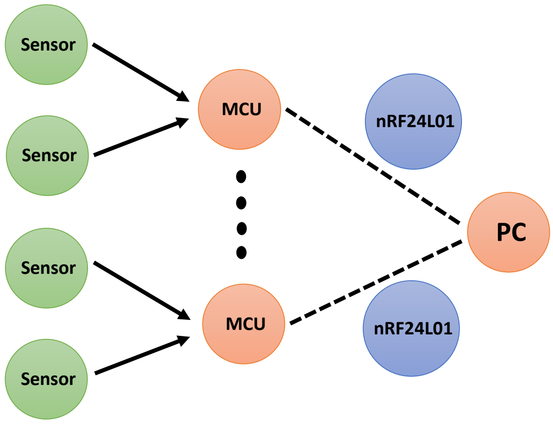


Fig. 2 Hardware design

As shown in the Hardware design diagram in Figure 2, inertial sensors connect Micro Centre Unit (MCU) using nRF24L01 wireless sensors. nRF24L01 is a highly integrated, ultra-low power 2Mbps RF transceiver IC for the 2.4GHz band. Transmitting of float data is not so easy by using nRF24L01 or Bluetooth as the communication of both 2 sensors are using byte. MCU is gathered the inertial sensor data, then use nRF24L01 for transmitting data to computer. For the first time, we wanted to use one sever of nRF24L01. But, since we had 5 clients the communication with those clients reduced the frequency of data. Even the gathering frequency is fast, it will be 5 times slower. Therefore, have to used 5 independent severs for 5 independent clients.

Three kinds of nodes; centre of gravity (COG), joint and segment used to construct virtual skeleton model. The COG is the centre point of the human model; at the same time it is also the root joint of human skeleton model. Segment is the component unit of the human skeleton model, which describes each part of the model. The COG is also the most important during estimation of human motion. As it is also a joint, but this joint is the centre of other joints, which is the coordinate origin of all the joints. When we have detected which leg is support and which leg is winging, we can calculate the position of the COG by using inverse kinematics. In this study we have constructed a human model containing 14 segments and 15 joints.

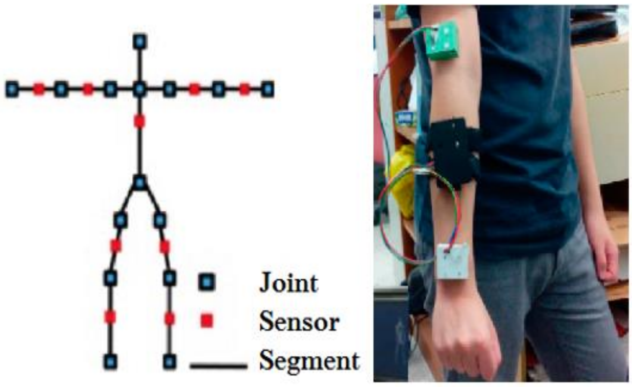


Fig. 3 Model construction and overview of the system

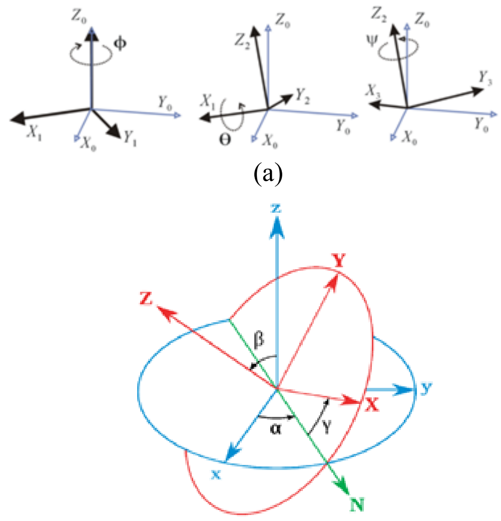
As shown in Figure 3, red points indicate sensors and blue points represent joints. We used 9 sensors attached on the major segments to record human motion. Also we have attached two sensors with the forearms and the upper arms???to record the motion of elbow and shoulder joints. One sensor is attached on the head to track the head's transfer. The chest has one sensor to describe the body's movement. For each leg, we have attached two sensors at thigh and calf to record the knee joints and ankle joints' motion.

B. Process of Motion

Euler angles represent three elemental rotations about the axes of the coordinate system. For instance, a first rotation about *z* by an angle *α*, a second about *x* by an angle *β,* and third again about *z*, by an angle *γ* (Figure 4 (a)). The axes of the original frame are denoted as *x*, *y*, *z* and the axes of the rotated frame are denoted as *X*, *Y*, *Z*. The line of nodes (*N*) is defined as the intersection of the *xy* and the *XY* coordinate planes. In other words, it is a line passing through the origin of both frames, and perpendicular to the *zZ*plane, on which both *z* and *Z* lie, shown as Figure 4 (b). The three Euler angles are defined as: *α* (or *Φ*) is the angle between the x axis and the *N* axis, *β* (or *Θ*) is the angle between the *z* axis and the *Z* axis, and *γ* (or *Ψ*) is the angle between the *N* axis and the *X* axis. This implies that: *α*, *β*, *γ* represent rotations around the *z* axis, *N* axis and *Z* axis respectively.

Most of the pedometers demonstrate an acceptable level of accuracy and reliability in step-count measurement [16]. There are plenty of studies which examined the accuracy, reliability, and validity of using pedometers [17-20]. Hasson et al. proposed first validation study of examining pedometer performance using a variable-speed condition [17].

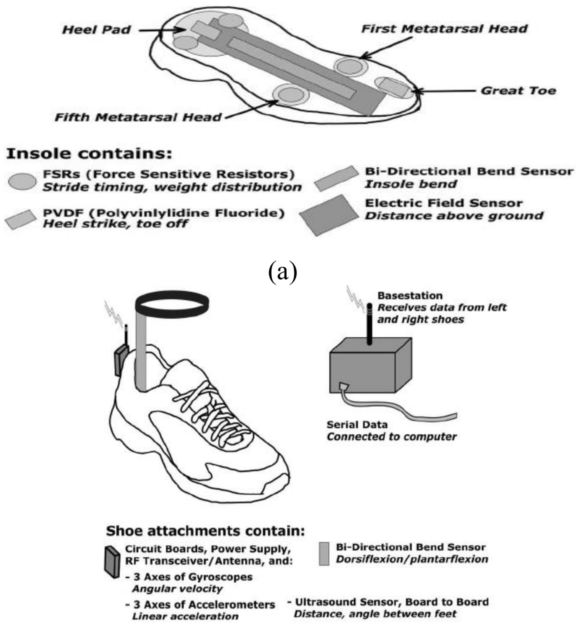
All accelerometers provide basic step counting and activity counts. Important gait parameters can be measured using accelerometer to evaluate one's risk of falling and mobility level [21]. Bamberg et al. proposed a gait analysis system using integrated wireless sensors [22]. Figure 5 describes the gait shoe system with labels indicating relevant anatomical markers. For the analysis of the kinematic motion of the foot, two dual axis accelerometers and three gyroscopes were placed at the back of the shoe, oriented such that the individual sensing axes were aligned along three perpendicular axes.



(a)

(b)

Fig. 4 Euler angles (a). A rotation represented by Euler angles (α, β, γ), (b). the Euler angle theory


Heel Pad
First Metatarsal Head
Great Toe
Fifth Metatarsal Head
Insole contains:
FSRs (Force Sensitive Resistors)
Stride timing, weight distribution
PVDF (Polyvinlylidine Fluoride)
Heel strike, toe off
Bi-Directional Bend Sensor
Insole bend
Electric Field Sensor
Distance above ground
Basestation
Receives data from left and right shoes
Serial Data
Connected to computer
Shoe attachments contain:
Circuit Boards, Power Supply, RF Transceiver/Antenna, and:
- 3 Axes of Gyroscopes
Angular velocity
- 3 Axes of Accelerometers
Linear acceleration
Bi-Directional Bend Sensor
Dorsiflexion/plantarflexion
- Ultrasound Sensor, Board to Board
Distance, angle between feet

(a)

(b)

Fig. 5 Schematic of the Gait Shoe system [22]

Figure 6 shows the relevant coordinate systems used for the analysis of the data. The global reference frame of the room and the second corresponds to the local body frame, in which the sensors are mounted and collect their measurements.

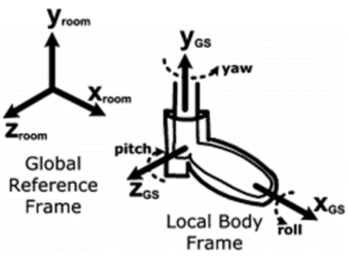


Fig. 6 Schematic of the Gait Shoe system [22]

C. Back Propagation Neural Network

First, low resolution sensor data of 8 bits were used which provide 0 ~ 255 range. Hence, a small change of sensor data can result a high angle change in the calculation. To improve sensor resolution, used 16 bit which provides 0 ~ 65,535 range. For more accurate results, we wanted to use Neural Network, to verifying our result with Kinect result. The above mentioned back propagation learning algorithm can be divided into two phases as propagation and weight update.

Algorithm 1:BP Neural Network

```
INPUT:            network weights (NW),
Proposed method's training pattern (TP)
                  Kinect's pattern (KP),
                  input neural network (INN)
OUTPUT:           build neural network (BNN)
BEGIN
  initialize NW
∀i∈TP
prediction = output (INN, TP)
actual     = output (INN, prediction)
error      = prediction - actual
∀NW from hidden to output layer
compute ΔWh
∀NW from input to hidden layer
compute ΔWi
update NW
  IF (TP classified correctly or satisfied network
performance)
  RETURN BNN
END
```

D. Human Model Construction

We take the right arm as an example to analyze the position and orientation of joints. The right arm is comprised of two body segments, i.e. right upper arm, right forearm. It is modeled as a kinematics chain of these two rigid segments linked by joints, i.e. right shoulder, right elbow and right wrist. The right shoulder is considered as the root joint in the right arm model. The skeleton model is established from the root joint downward to form a kinematics model with the joints obeying a Parent-Child relationship[23, 24]. We assume the length of each segment is fixed. The formula of obtaining joint position is,

$$P_{elbow} = P_{shoulder} + Q_{upperarm} \otimes V_{upperarm} \otimes Q^{-1}_{upperarm}. \qquad (6)$$

Here, $P_{shoulder}$ is the position of shoulder joint, $V_{upperarm}$ is the initial vector of the upper arm segment which should be known before calculation. $Q_{upperarm}$ is the orientation of the upper arm segment represented by quaternion which rotate the vector of the human body in the global coordinate system. ⊗ is the quaternion multiplication operator.

We can easily get the wrist joints position after we have calculated the elbow joints position using the following formula,

$$P_{wrist} = P_{elbow} + Q_{forearm} \otimes V_{forearm} \otimes Q^{-1}_{forearm}. \qquad (7)$$

As we have just got P_{elbow} in (6), we can easily get P_{wrist} by taking (7) into (6). Applying this in to the whole human model, we can get the relation and position of each joint. In the human lower body is comprised of seven body segments, i.e. pelvis, left and right femurs, left and right tibias, and left and right feet. It is modeled as a kinematics chain of these seven rigid segments linked by joints, i.e. hips, knees, ankles and toes. The pelvis is considered the root joint in the model. The skeleton model is established from the root joint downward to form a kinematics model with the joints obeying a parent-child relationship. These rigid body segments can be represented by vectors. We build two sets of the lower body segments, namely, S_L= {*Pelvis, LelfFemur, LeftTibia, LeftFoot*}andS_R= {*Pelvis, RightFemur, RightTibia, RightFoot*}, also two sets of the joints J_L= {*Pelvis, LeftHip, LeftKnee, LeftAnkle, LeftToe*}andJ_R= {*Pelvis, RightHip, RightKnee, RightAnkle, RightToe*}. In these sets, the preceding element is the parent. We take the right lower limb as an example to demonstrate how the position information is transmitted between lower body segments according to the segmental kinematics. From the proximal joint, e.g. root joint, to the distal joint, the position of the child joint can be calculated from its parent joints position using,

$$P_{J_R(k),t} = P_{J_R(k-1),t} + Q_{S_R(i),t} \otimes V_{S_R(i),0} \otimes Q^{-1}_{S_R(i),t} \qquad (8)$$

where i = 1,2,3,4 and k = 2,3,4,5. From the distal joint, e.g. right toe, to the proximal joint, the position of the parent joint can be calculated from its child joints position using,

$$P_{J_R(k-1),t} = P_{J_R(k),t} + Q_{S_R(i),t} \otimes V_{S_R(i),0} \otimes Q^{-1}_{S_R(i),t} \qquad (9)$$

where i = 4,3,2,1 and k = 5,4,3,2.

Range of Motion (ROM) is the angle that a joint may normally travel. ROM can help us get the maximum angle of joints motion, including ante flexion, posterior extension, abduction and adduction. Therefore, we used ROM to limit skeleton model joints motion angle, calibrating the sensor captured data

When taken right shoulder (*RS*) joint as an example, the range of motion of shoulder joint was 90° for ante flexion, 60° for posterior extension, 90° for abduction and 40° for adduction as depicted in Figure 7 (a). We can easily get $-60° < \theta_{RS_x} < -180°$ while $-180° < \theta_{RS_y} < 40°$. Since we considered as the length of each segment is fixed, assumed shoulder joints coordinate value is (0, 0, 0) and upper arm length is l_0. The elbow coordinate value can be calculated using,

doi:10.4038/icter.v9i2.7180

$$(x_{elbow}, y_{elbow}, z_{elbow}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_{RS_x} & -\sin\theta_{RS_x} \\ 0 & \sin\theta_{RS_x} & \cos\theta_{RS_x} \end{bmatrix} \begin{bmatrix} \cos\theta_{RS_y} & 0 & \sin\theta_{RS_y} \\ 0 & 1 & 0 \\ -\sin\theta_{RS_y} & 0 & \cos\theta_{RS_y} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ -l_0 \end{bmatrix}. \tag{10}$$

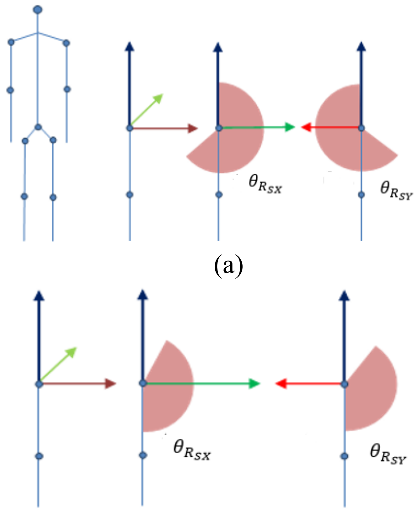
Using the values $\theta_{RSx}$and θ_{RSy} from 10, the motion range of elbow joint can be calculated. By comparing this elbow range with the result of 6, we could filter input sensor data. If the results of 6 is not in the range of elbow, we assume that it may not a human regular motion.

For the right-elbow (*RE*) joint shown in the Figure 7(b), it is not only deal with its own range of motion, since it is a kinematics chain. It also has to follow the transfer relation. So the angels can be described as following range:

$$\theta_{RS_x} < \theta_{RE_x} < \theta_{RS_x} + 150°,$$
$$\theta_{RS_y} < \theta_{RE_y} < \theta_{RS_y} + 150°.$$

After we get the elbows range of motion fore-arm length is l_1, the end values can be calculated using,

$$(x_{elbow}, y_{elbow}, z_{elbow}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\theta_{RS_x} & -sin\theta_{RS_x} \\ 0 & sin\theta_{RS_x} & cos\theta_{RS_x} \end{bmatrix} \begin{bmatrix} cos\theta_{RS_y} & 0 & sin\theta_{RS_y} \\ 0 & 1 & 0 \\ -sin\theta_{RS_y} & 0 & cos\theta_{RS_y} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ -l_0 \end{bmatrix} + \begin{bmatrix} x_{elbow} \\ y_{elbow} \\ z_{elbow} \end{bmatrix} \tag{11}$$



(a)

(b)

Fig. 7 The range of motion (a) right shoulder Joint, (b) right elbow joint

E. Extended Kalman Filter

In this research, by integrating the output of the gyroscope and the accelerometer, the Kalman filter provided a noisy and disturbed but drift-free measurement of orientation. We can get the Process Model and Measurement Model for the extended Kalman filter.

For a sensor in the state of rest, the linear acceleration is quite small. So the signals from an accelerometer can be regarded as the gravitational acceleration. But in most situations, the kinematics linear acceleration is usually in existence. Thus, the data from accelerometer and gyroscope are fused using EKF to calculate the gravitational acceleration. The state vector of the dynamical system is the gravitational acceleration expressed in the sensor coordinate frame, represented by g^s. The control vector of the system is the angular velocity from the gyroscope, denoted by ω^s. The measurement vector of the system is the acceleration data from the accelerometer, denoted by a^S. Define $g^S = [\, g_X^s, g_Y^s, g_Z^s \,]^T$, $\omega^S = [\, \omega_X^s, \omega_Y^s, \omega_Z^s \,]^T$, $a^S = [\, a_X^s, a_Y^s, a_Z^s \,]^T$. Then the state equation of the dynamical system is represented by,

$$\begin{pmatrix} g_x^s(t+T) \\ g_y^s(t+T) \\ g_z^s(t+T) \end{pmatrix} = \begin{pmatrix} g_x^s(t) + \left(\omega_y^s(t) g_z^s(t) - \omega_z^s(t) g_y^s(t) \right) T \\ g_y^s(t) + \left(\omega_z^s(t) g_x^s(t) - \omega_x^s(t) g_z^s(t) \right) T \\ g_z^s(t) + \left(\omega_x^s(t) g_y^s(t) - \omega_y^s(t) g_x^s(t) \right) T \end{pmatrix}, \tag{12}$$

where T denotes the sampling period of sensors. The measurement equation of the system is represented by,

$$\begin{pmatrix} a_x^s(t) \\ a_y^s(t) \\ a_z^s(t) \end{pmatrix} = \begin{pmatrix} g_x^s(t) + n_{gx}(t) \\ g_y^s(t) + n_{gy}(t) \\ g_z^s(t) + n_{gz}(t) \end{pmatrix} \tag{13}$$

where the $n = [n_{gx}, n_{gy}, n_{gz}]^T$ denotes the measurement noise.

The result of the human motion of human skeleton is described using Euler angle. So all data from sensors used to calculate angle in order to make the dynamic of human motion. The gyroscope can be calculated into the angle using integrating of sensor sampling time, and the accelerometer can be calculated a deflected angle by using trigonometric.

$$Angle(Axz) = sin^{-1}(x/gravity) \tag{14}$$
$$Angle(Axz) = tan^{-1}(x/z) \tag{15}$$

During the calculation, initially, we only considered equation 14. But, when the x is larger than gravity answer is not possible. And also there was an error when only used equation 15 when the z value is minimum. Hence we combined two formulas and run while wearing the sensors in order to get much more stable result for the angle. At that time, we noticed that there was a small difference between angles. Then assigned the minimum difference between two formulas and got the accelerometer value as 0.4477539 m/s^2.

Algorithm 2: Merge the Gyro and Accel data

```
INPUT:      acceleration (accel), gravity (gyro),
            distance alone x axis(x)
            distance alone z axis(z)
            total data(n)
OUTPUT:     acceleration angle ($Angle_{accel}$),
            gravitational angle ($Angle_g$)
BEGIN
$\forall n$
```

```
get accel[$n$] from sensor using median filter
get gyro[$n$] from sensor using high pass filter
$\forall a \in accel$
IF $a$> 0.4477539
$Angle_{accel}$= $\sin^{-1}(x/gyro)$
ELSE
$Angle_{accel}$= $\tan^{-1}(x/z)$
$\forall g \in gyro$
$g$ = Ext. Kal. Filter ($gyro, accel$)
$Angle_g$    = $\int g\, dt$
RETURN $Angle_{accel}$, $Angle_g$
END
```

Algorithm 3: Walking estimation

```
INPUT  :      legs status ($LS$), stands ($ST$), swing ($SW$)
OUTPUT:       Kinematic analysis ($Ka$)
BEGIN
$\forall LS$
IF both legs $ST$
$Ka$ = calculate (Forward Kinematics)
ELSE IF one leg $ST$ and the other $SW$
IF leg $stands$
$Ka$ = calculate (Forward Kinematics)
ELSE
$Ka$ = calculate (Inverse Kinematics)
RETURN $Ka$
END
```

F. *Gait Estimation*

Walking estimation is conducted in the following two steps: walking status detect and kinematics analysis as written in algorithm 3. Walking status detect is used to determine the support leg of which ankle joint is selected as the root joint of the lower body model during walking. Human walking is a cyclical motion. We can divide into two phase as shown in Figure 8 (a) stance phase (*ST*) and swing phase (*SW*).

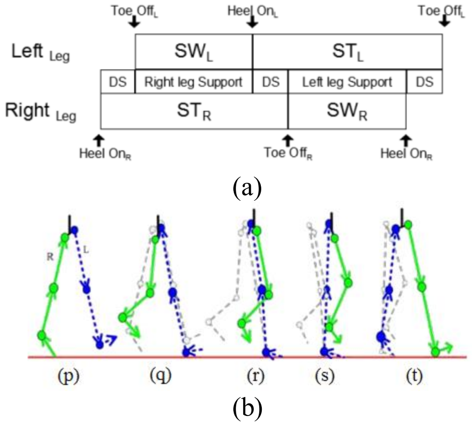


Fig. 8 (a) Leg status during walking, (b) walking simulation.

The stance phase (*ST*) starting by a heel on the ground is the portion of the cycle during which a foot is contacted with the ground. The swing phase (*SW*) starting with a toe off which is the portion of the cycle when the foot is not in contact with the ground. Something during a walking our both legs are standing or are in the double stance phase (*DS*). During the *DS* phase, we consider the leg which is just going into the ST as the support leg. The stick model of the walking gait shows in Figure 8 (b). (p): Left terminal swing, Right terminal stance, Right toe reference. (q), (r), (s): Left support leg, Left toe reference. (t): Left terminal stance, Right terminal swing, Left toe reference. Here grey dash lines represent the lower body movements

IV. EXPERIMENTAL RESULTS AND DISCUSSION

We performed the following experiments to evaluate the efficacy of our inertial sensor based human motion tracking system. We used OpenNI library since it is open source and we could easily get the coordinate value of each joint in human model. First, we compared the tracking result. Then we record 14 joints position data in our method and Kinect, include left-shoulder, left-elbow, left-wrist, right-shoulder, right-elbow, right-wrist, left-hip, left-knee, left-ankle, right-hip, right-knee, right-ankle, chest, and head.

A. *Comparing with Kinect Result*

While running our method and OpenNI project at the same time, recorded the posture result. Inertial sensors are mounted to the human body segments including upper-arms, forearms, chest, thighs and shanks. First we sit on chair and swing the arms. Then we stand and do much more postures. Figure 9 shows six screenshots of recorded results. The left picture of a particular screenshot represents the result of proposed method while the right picture represents Kinect OpenNI. During this experiment, we could notice that the proposed motion tracking system works well in position of human skeleton in real-time.

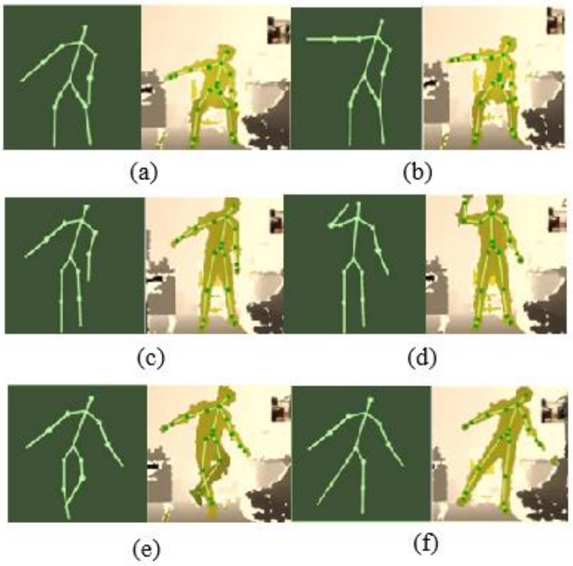


Fig. 9 Experimental results of both our method and Kinect OpenNI. Left: our method result; Right: Kinect OpenNI result. For (a) and (b) we sit on the chair, and for the left (c), (d), (e), (f) we just take much standing postures.

doi:10.4038/icter.v9i2.7180

B. Coordinate Diagram

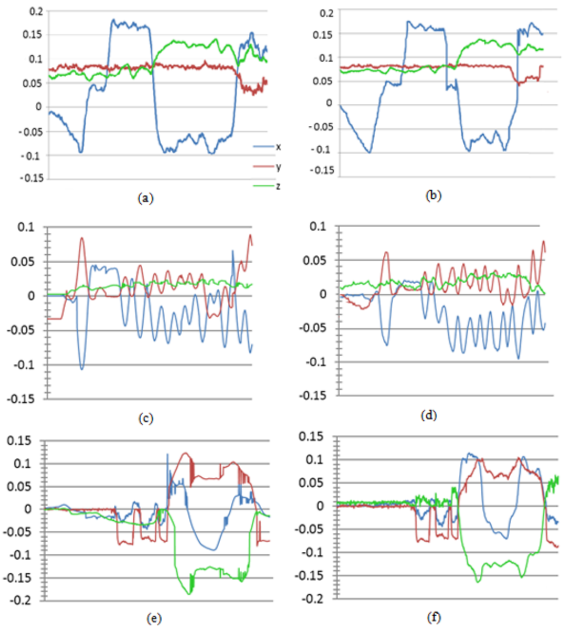
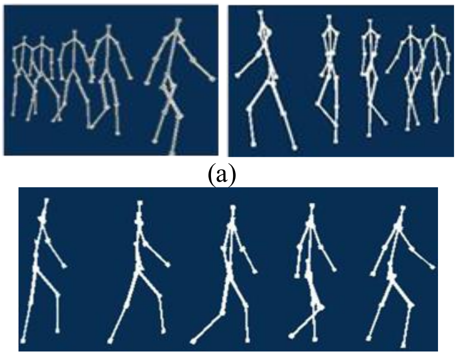


Fig. 10 Coordinate diagram result. (a) Chest result using our method. (b) using Kinect OpenNI. (c) takes the left wrist position result using our method, (d) using Kinect OpenNI. (e) the result of right ankle using our method (f) Kinect result



(a)

(b)

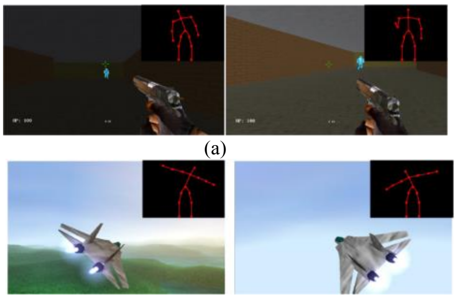
Fig. 12 Human walking gait record (a) View from right side (b) view from left side (c) full view.

We have recorded 13 skeleton joints coordinate value including shoulder joints, elbow joints, wrist joints, hip joints, knee joints, ankle joints and chest of both methods. Among them we have considered chest, left wrist and right ankle as examples. Figure 10 depicts the comparative results of the joints coordinate diagram result for our method and Kinect sensor. Motion analysis, which represent in y-axis in meters over time in seconds. Both 10(a) and 10(b) depict the result of chest joint's coordinate result. 10(a) is using our proposed method and 10(b) is using a little more convergent and the data is much smoother. 10(c) and 10(d) depict the location of the left wrist, while 10(e) and 10(f) are represent the right ankle diagram. Corresponding values of Figure 10(a), 10(b) and 10(c), 10(d) are as in Table 1 and Table 2 respectively.

C. Applications

HCI is wildly used in many parts of human life, like video games, virtual reality, 3D technology, Multi-Media, etc. We used our method to play some video games and use our method to control the Microsoft Power Point. Figure 11 (a) shows that play First Person Shooting (FPS) game using our method. Used chest joints angle to control the rotation of the player. When the chest joint leans to the right side, the player in the video game will leads to right side. Ankle joints movement controls the players step. Player will fire while we put right hand up. A flight simulator game has been tested using our method which has been shown in Figure 11 (b). We used upper body's posture to control planes fly state.

D. Walking Estimation



(a)

(b)

Fig. 11 Game control using our method (a). FPS game using our method (b) flight simulator game using our method

As we used two kinds of kinematics, our model is not fixed at the centre of mass. Model can walk as what we have done in the reality. We have recorded many walking frames as shown in Figure 12.

TABLE 1
MOTION ANALYSIS OVER TIME IN SECONDS AND JOINT POSITION VALUES CORRESPONDINGTO FIGURE10(a) AND 10(b)

	(a)			(b)		
Time	x	z	y	x	z	y
0	-0.0013	0.0076	0.0073	0	0.0075	0.0074
1	-0.0025	0.0075	0.0075	-0.005	0.0076	0.0078
2	-0.0075	0.0076	0.0075	-0.01	0.0077	0.0078
3	0.005	0.0125	0.0125	-0.0025	0.0076	0.0077
4	0.0025	0.0075	0.0076	0.005	0.0078	0.0078
5	0.01875	0.0075	0.0075	0.005	0.0076	0.0078
6	0.01875	0.01	0.00875	0.0185	0.0076	0.0079
7	0.01875	0.00875	0.00874	0.0175	0.0078	0.0078
8	-0.0025	0.01125	0.01	0.0175	0.0076	0.0077
9	-0.01	0.01375	0.00875	0.005	0.0077	0.0077
10	-0.0075	0.015	0.00876	-0.0085	0.0125	0.00775
11	-0.0055	0.01377	0.00875	-0.0045	0.015	0.0077
12	-0.01	0.01375	0.00878	-0.0045	0.014	0.0077
13	-0.0025	0.0125	0.00875	-0.0075	0.0125	0.0077
14	0	0.0158	0.0075	-0.005	0.0125	0.0077
15	0.015	0.01375	0.0025	0.0175	0.011	0.005
16	0.01125	0.01125	0.0026	0.0175	0.0125	0.006
17	0.0015	0.01	0.0125	0.0145	0.0126	0.0078
18	0.0015	0.01	0.0125	0.0145	0.0126	0.0078

TABLE 2
MOTION ANALYSIS OVER TIME IN SECONDS
AND JOINT POSITION VALUES CORRESPONDING TO FIGURE 10(c) AND 10(d)

	(c)			(d)		
Time	x	z	y	x	z	y
0	0	0	-0.02	0	0.024	0.001
1	0	0	-0.02	0.01	0.024	-0.001
2	0	0	0.01	0.01	0.025	-0.0125
3	0	0.0125	0.025	0	0.025	-0.0127
4	-0.05	0.0125	0.0125	0.0125	0.023	0.0125
5	0.025	0.0123	0.0125	-0.05	0.025	0.0628
6	0.0875	0.0125	0.0125	0.023	0.025	0.023
7	0.075	0.0128	0	0.025	0.0135	0.025
8	0	0.0128	0.025	-0.025	0.025	-0.025
9	-0.125	0.0125	0.05	0.025	0.025	0.01
10	0.025	0.025	0	-0.0125	0.0134	0.025
11	-0.025	0.025	0.0125	-0.075	0.024	0.039
12	-0.075	0.023	0.025	-0.075	0.025	0.05
13	-0.05	0.023	0.025	-0.073	0.03	0.0385
14	-0.025	0.024	0.022	-0.075	0.03	0.05
15	0.077	0.025	0.022	-0.075	0.0375	0.0385
16	0	0.025	-0.025	-0.073	0.0376	0.025
17	-0.05	0.025	-0.025	-0.0875	0.0375	0.05
18	0.025	0.0125	-0.0125	0	0.025	0.001
19	-0.125	0.0125	0.0125	0.0125	0.024	0.075
20	0.075	0.0124	0.075	0.0123	0.023	0.073

V. CONCLUSION AND FUTURE WORK

This research study implemented a HMC system using multiple inertial sensors, and introduced an efficient human tracking algorithm. Further by applying an accurate gait estimation algorithm, a real-time human skeleton animation capture system was born. The performance of system was fast and stably in an area of 20m×20m which is much larger than Kinect.

After performing a series of experiments such as human motion with complicated movements, it was shown that our method can do the same done by Kinect SDK with much less data shaking. Also some motions like hand cross over or leg cross over, our method produced better results than Kinect as Kinect may lose the skeleton of the blocked part. Furthermore, our method has a larger working area than Kinect as our method uses a radio frequency sensor which can provide a communication range of 10m or more. By applying kinematics, we could make the skeleton movement as in reality, such as walking estimation. Additionally, we have tested our method with video games and it worked well in both of these two games.

As possible extensions of this work, we would like to improve our system under hardware design, performance improvement and experiment. Currently, the method has only 20 fps, in future, we plan to use a faster micro centre unit to increase the number of fps. Since our proposed method has been used nRF24l01 for the communication between PC and MCU, which has 10 meter to 15meter plan to replace nRF24l01 with WIFI unit which can make an even larger area of working. Furthermore, by attaching more joints such ascrotch joint and back joint, much more detailed human motion could be presented. We plan to build a Back Propagation Neural Network by using Kinect results as the standard output and our method results as the input data, to get better capture results of human motion. In addition to Kinect, we would like to compare our method with other inertial sensor systems by Xsens technology or other companies.

REFERENCES

[1] Hadjidj, A., Souil, M., Bouabdallah, A., Challal, Y., & Owen, H. (2013). *Wireless sensor networks for rehabilitation applications: Challenges and opportunities*. Journal of Network and Computer Applications, 36(1), 1-15.

[2] Kurakin, A., Zhang, Z., &Liu, Z. (2012). *A real time system for dynamic hand gesture recognition with a depth sensor*. Signal Processing Conference (EUSIPCO).

[3] Shiratori, T., Park, H. S., Sigal, L., Sheikh, Y., & Hodgins, J. K. (2011). *Motion capture from body-mounted cameras*. ACM SIGGRAPH 2011 Papers on - SIGGRAPH '11.

[4] Slyper, R., &Hodgins, K. (2008). *Action capture with accelerometers*. ACM SIGGRAPH/Eurographics Symposium on Computer Animation.

[5] Tautges, J., Zinke, A., Krüger, B., Baumann, J., Weber, A., Helten, T., Eberhardt, B. (2011). *Motion reconstruction using sparse accelerometer data*. TOG ACM Trans. Graph. ACM Transactions on Graphics, 30(3), 1-12.

[6] Favre, J., Aissaoui, R., Jolles, B., Guise, J. D., & Aminian, K. (2009). *Functional calibration procedure for 3D knee joint angle description using inertial sensors*. Journal of Biomechanics, 42(14), 2330-2335.

[7] Schepers, H. M., Roetenberg, D., & Veltink, P. H. (2009). *Ambulatory human motion tracking by fusion of inertial and magnetic sensing with adaptive actuation*. Med Biol Eng Comput Medical & Biological Engineering & Computing, 48(1), 27-37.

[8] Altun, K., Barshan, B., & Tunçel, O. (2010). *Comparative study on classifying human activities with miniature inertial and magnetic sensors*. Pattern Recognition, 43(10), 3605-3620.

[9] Cuesta-Vargas, A. I., Galán-Mercant, A., & Williams, J. M. (2010). *The use of inertial sensors system for human motion analysis*. Physical Therapy Reviews, 15(6), 462-473.

[10] Zhou, H., Stone, T., Hu, H., & Harris, N. (2008). *Use of multiple wearable inertial sensors in upper limb motion tracking*. Medical Engineering & Physics, 30(1), 123-133.

[11] Zalewski, G. M., Marks, R., & Mao, X. (2008). *U.S. Patent No. US 7352359 B2*. Washington, DC: U.S. Patent and Trademark Office.

[12] Raptis, M., Kirovski, D., & Hoppe, H. (2011). *Real-time classification of dance gestures from skeleton animation*. ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '11.

[13] McCarthy, J. M. (1990). *An introduction to theoretical kinematics. Cambridge*, MA: MIT Press.

[14] Finkelstein, A. (2016). *Kinematics & Dynamics* (1st ed., pp. H&B 13.7). Princeton: Department of Computer Science, Princeton University. Retrievedfrom http://www.cs.princeton.edu/courses/archive/spr05/cos426/lectures/18-kinematics.pdf.

[15] Welch, G., & Bishop, G. (2008, July 28). The Kalman Filter. Retrieved July 20, 2015, from http://www.cs.unc.edu/~welch/kalman/

[16] Giannakidou, D. M., Kambas, A., Ageloussis, N., Fatouros, I., Christoforidis, C., Venetsanou, F., Taxildaris, K. (2011). *The validity of two Omron pedometers during treadmill walking is speed dependent*. European Journal of Applied Physiology Eur J Appl Physiol, 112(1), 49-57.

[17] Hasson, R. E., Haller, J., Pober, D. M., Staudenmayer, J., & Freedson, P. S. (2009). *Validity of the Omron HJ-112 Pedometer during Treadmill Walking*. Medicine & Science in Sports & Exercise, 41(4), 805-809.

[18] Ayabe M, Aoki J, Ishii K, Takayama K, Tanaka H (2008). *Pedometer accuracy during stair climbing and bench stepping exercises*. J Sports Sci Med 7:249–254.

[19] Doyle, J. A., Green, M. S., Corona, B. T., Simone, J., & Dennison, D. A. (2007). *Validation of an Electronic Pedometer in a Field-Based Setting*. Medicine & Science in Sports & Exercise, 39(Supplement).

[20] Holbrook, E. A., Barreira, T. V., & Kang, M. (2009). *Validity and Reliability of Omron Pedometers for Prescribed and Self-Paced Walking*. Medicine & Science in Sports & Exercise, 41(3), 670-674.

[21] Yang, C., & Hsu, Y. (2010). *A Review of Accelerometry-Based Wearable Motion Detectors for Physical Activity Monitoring. Sensors*, 10(8), 7772-7788.

[22] Bamberg, S., Benbasat, A., Scarborough, D., Krebs, D., & Paradiso, J. (2008). *Gait Analysis Using a Shoe-Integrated Wireless Sensor System*. IEEE Transactions on Information Technology in Biomedicine IEEE Trans. Inform. Technol. Biomed., 12(4), 413-423.

[23] Mihelj, M. (2006). *Inverse Kinematics of Human Arm based on Multisensor Data Integration*. J Intell Robot Syst., 47, 139-153.

[24] Downs, L. (Ed.). (2003, June 03). CS184: *Using Quaternions to Represent Rotation*. Retrieved December 23, 2016, from http://web.archive.org/web/20040214133703/http://www.cs.berkeley.edu/~laura/cs184/quat/quaternion.html.

List of Reviewers

Dr. L.N.C. De Silva
University of Colombo School of Computing

Dr. A.Y. Ekanayaka
University of Colombo School of Computing

Dr. H.E.M.H.B. Ekanayake
University of Colombo School of Computing

Dr. M. G. N. A. S. Fernando
University of Colombo School of Computing

Dr. M.D.J.S. Goonetillake
University of Colombo School of Computing

Dr. K.H. E. L. W. Hettiarachchi
University of Colombo School of Computing

Prof. N.D. Kodikara
University of Colombo School of Computing

Mr. R.S. Madanayake
University of Colombo School of Computing

Dr. A. Mahasinghe
Department of Mathematics, University of Colombo

Dr. S. Mahesan
Department of Computer Science, University of Jaffna

Dr. K.D. Sandaruwan
University of Colombo School of Computing

Mr. K.P.M.K. Silva
University of Colombo School of Computing

Dr. T.A. Weerasinghe
University of Colombo School of Computing

Mr. W.V. Welgama
University of Colombo School of Computing

Prof. G.N. Wikramanayake
University of Colombo School of Computing

ICTer COPYRIGHT FORM

To ensure uniformity of treatment among all contributors, this form shall be treated as the only copyright form for publications in "ICTer" and other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the ICTer and must accompany any such material in order to be published by the ICTer. Please read the form carefully and keep a copy for your reference.

TITLE OF PAPER/ARTICLE/REPORT/PRESENTATION/SPEECH (hereinafter, "the Work"):

COMPLETE LIST OF AUTHORS:

ICTer PUBLICATION TITLE (Journal):

Copyright Transfer

The undersigned hereby assigns to the International Journal on Advances in ICT for Emerging Regions, (the "ICTer") all rights under copyright that may exist in and to the above Work, and any revised or expanded derivative works submitted to the ICTer by the undersigned based on the Work. The undersigned hereby warrants that the Work is original and that he/she is the author of the Work; to the extent the Work incorporates text passages, figures, data or other materials from the works of others, the undersigned has obtained any necessary permission. **See reverse side for Retained Rights and other Terms and Conditions.**

Author Responsibilities

The ICTer distributes its publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements of the ICTer Author guidelines, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on ICTer's publishing policies may be found at http://www.icter.org. Authors are advised especially of the following.

1. It is the responsibility of the authors, not the ICter, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it.

2. Statements and opinions given in work published by the ICter are the expression of the authors.

General Terms

- The undersigned represents that he/she has the power and authority to make and execute this assignment.
- The undersigned agrees to indemnify and hold harmless the ICTer from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
- In the event the above work is not accepted and published by the ICTer or is withdrawn by the author(s) before acceptance by the ICTer, the foregoing copyright transfer shall become null and void and all materials embodying the Work submitted to the ICTer will be destroyed.
- For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.

(1)____________________________________ ___________________________

Author/Authorized Agent for Joint Authors **Date**

ICTer COPYRIGHT FORM *(continued)*

RETAINED RIGHTS/TERMS AND CONDITIONS

1. Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.

2. Authors/employers may reproduce or authorize others to reproduce the Work, materials extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the ICTer copyright notice are indicated, the copies are not used in any way that implies ICTer endorsement of a product or service of any employer, and the copies themselves are not offered for sale.

3. Authors/employers may make limited distribution of all or portions of the Work prior to publication if they inform the ICTer in advance of the nature and extent of such limited distribution.

4. In the case of a Work performed under any Government contract or grant, the ICTer recognizes that the said Government has permission to reproduce all or portions of the Work, and to authorize others to do so, for official Government purposes, only if the contract/grant so requires.

5. For all scenarios not covered by items 2, 3, and 4, authors/employers must request permission from the ICTer to reproduce or authorize the reproduction of the Work or materials extracted verbatim from the Work, including figures and tables.

6. Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use. The ICTer shall handle all such third-party requests.

INFORMATION FOR AUTHORS

ICTer Copyright Ownership

It is formal policy of the ICTer to own the copyrights to all copyrightable materials in its journal publications and to the individual contributions contained therein, in order to protect the interests of the ICTer, its authors and their employers, and, at the same time, to facilitate the appropriate re-use of this material by others. The ICTer distributes its journal publications throughout the world by means such as hard copy, and electronic media. It also abstracts and may translate its publications, and articles contained therein, for inclusion in various compendiums, collective works, databases and similar publications.

Author/Employer Rights

If you are employed and prepared the Work on a subject within the scope of your employment, the copyright in the Work belongs to your employer as a work-for-hire. In that case, the ICTer assumes that when you sign this Form, you are authorized to do so by your employer and that your employer has consented to the transfer of copyright, to the representation and warranty of publication rights, and to all other terms and conditions of this Form. If such authorization and consent has not been given to you, an authorized representative of your employer should sign this Form as the Author.

Reprint/Republication Policy

The ICTer requires that the consent of the first-named author and employer be sought as a condition to granting reprint or republication rights to others or for permitting use of a Work for promotion or marketing purposes.

PLEASE DIRECT ALL QUESTIONS ABOUT THIS FORM TO:

Chair-ICTer Steering Committee, University of Colombo School of Computing, No 35, Reid Avenue, Colombo 7, Sri Lanka
Tel: 94-112-581245, **Fax:** 94-112-591245 **Email:** icter@ucsc.cmb.ac.lk

International Journal on Advances in ICT for Emerging Regions

Volume 09 Number 1 and 2 2016

UCSC

1967 - 2017 50 years of computing at University of Colombo

www.ucsc.cmb.ac.lk

ICTer Journal
University of Colombo School of Computing
UCSC Building Complex
No 35, Reid Avenue, Colombo 07, Sri Lanka
journal.icter.org
icter@ucsc.cmb.ac.lk
info@icter.org
Tel: +94 11 258 1245 **Fax:** +94 11 258 7239