Online ISSN : 0975-4172
Print ISSN : 0975-4350
DOI : 10.17406/GJCST

Global Journal

of Computer Science and Technology: C

Software & Data Engineering

Crop Coverage Data

Review of Machine Learning

Highlights

Multimedia Data Mining

Evolution of Object-Oriented

Discovering Thoughts, Inventing Future

Volume 16 Issue 3 Version 0.1

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Volume 16 Issue 3 (Ver. 1.0)

Open Association of Research Society

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

*(A Delaware USA Incorporation with "Good Standing"; **Reg. Number: 0423089**)*

Sponsors: Open Association of Research Society
Open Scientific Standards

Publisher's Headquarters office

Global Journals® Headquarters
945th Concord Streets,
Framingham Massachusetts Pin: 01701,
United States of America
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Journals Incorporated
2nd, Lansdowne, Lansdowne Rd., Croydon-Surrey,
Pin: CR9 2ER, United Kingdom

Packaging & Continental Dispatching

Global Journals
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin: 452009, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investors@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Including by Air Parcel Charges):

For Authors:

22 USD (B/W) & 50 USD (Color)

Yearly Subscription (Personal & Institutional):

200 USD (B/W) & 250 USD (Color)

Integrated Editorial Board
(Computer Science, Engineering, Medical, Management, Natural Science, Social Science)

Dr. Bart Lambrecht
Director of Research in Accounting and FinanceProfessor of Finance
Lancaster University Management School
BA (Antwerp); MPhil, MA, PhD (Cambridge)

Dr. Carlos García Pont
Associate Professor of Marketing
IESE Business School, University of Navarra
Doctor of Philosophy (Management), Massachusetts Institute of Technology (MIT)
Master in Business Administration, IESE, University of Navarra
Degree in Industrial Engineering, Universitat Politècnica de Catalunya

Dr. Fotini Labropulu
Mathematics - Luther College
University of ReginaPh.D., M.Sc. in Mathematics
B.A. (Honors) in Mathematics
University of Windso

Dr. Lynn Lim
Reader in Business and Marketing
Roehampton University, London
BCom, PGDip, MBA (Distinction), PhD, FHEA

Dr. Mihaly Mezei
ASSOCIATE PROFESSOR
Department of Structural and Chemical Biology, Mount Sinai School of Medical Center
Ph.D., Etvs Lornd University
Postdoctoral Training,
New York University

Dr. Söhnke M. Bartram
Department of Accounting and FinanceLancaster University Management SchoolPh.D. (WHU Koblenz)
MBA/BBA (University of Saarbrücken)

Dr. Miguel Angel Ariño
Professor of Decision Sciences
IESE Business School
Barcelona, Spain (Universidad de Navarra)
CEIBS (China Europe International Business School).
Beijing, Shanghai and Shenzhen
Ph.D. in Mathematics
University of Barcelona
BA in Mathematics (Licenciatura)
University of Barcelona

Philip G. Moscoso
Technology and Operations Management
IESE Business School, University of Navarra
Ph.D in Industrial Engineering and Management, ETH Zurich
M.Sc. in Chemical Engineering, ETH Zurich

Dr. Sanjay Dixit, M.D.
Director, EP Laboratories, Philadelphia VA Medical Center
Cardiovascular Medicine - Cardiac Arrhythmia
Univ of Penn School of Medicine

Dr. Han-Xiang Deng
MD., Ph.D
Associate Professor and Research Department Division of Neuromuscular Medicine
Davee Department of Neurology and Clinical NeuroscienceNorthwestern University Feinberg School of Medicine

Contents of the Issue

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Crop Coverage data Classification using Support Vector Machine

By Tarun Rao, N. Rajasekhar & N C Naveen
Dayananda Sagar College of Engineering

Abstract- A statistical tool which can be used in various applications ranging from medical science to agricultural science is support vector machines. The proposed methodology used is support vector machine and it isused to classify a raster map. The dataset used herein is of Gujarat state agriculture map. The proposed approach is used to classify raster map into groups based on crop coverage of various crops. One group represents rice crop coverageand the othermillets crop coverage and yet another that of cotton crop coverage.Various statistical parameters are used to measure the efficacy of the proposed methodology employed.

Keywords: mining, SVM, supervised classification.

GJCST-C Classification: C.1.2

CROPCOVERAGEDATACLASSIFICATIONUSINGSUPPORTVECTORMACHINE

Crop Coverage data Classification using Support Vector Machine

Tarun Rao [α], N. Rajasekhar [σ] & N C Naveen [ρ]

Abstract- A statistical tool which can be used in various applications ranging from medical science to agricultural science is support vector machines. The proposed methodology used is support vector machine and it isused to classify a raster map. The dataset used herein is of Gujarat state agriculture map. The proposed approach is used to classify raster map into groups based on crop coverage of various crops. One group represents rice crop coverageand the othermillets crop coverage and yet another that of cotton crop coverage.Various statistical parameters are used to measure the efficacy of the proposed methodology employed.

Keywords: mining, SVM, supervised classification.

I. Introduction

Crop mapping is widely used in agriculture andremote sensing science. Crop mapping using classification methodologies serves various applications in agricultural sciencelike gauging water and soil demand etc.. For such applications information on the spatial distribution of classification error is of particular interest [1].Recent progresses in Information Technology systems, lead to dynamic communication among people of every profession. Information technology systems have changed the way people meet and communicate. There is an increasing tendency of professionals and experts in the agriculture sector to communication best practices in the field of agriculture via the medium of internet. Farmers who use the medium of internet get benefited from the various forums used therein to communicate advanced crop yield technologies. Crop mapping can also facilitate the farmers in planning their crop management in advance and they do not see internet and modern technologies has a hurdle [2].

Data is everywhere, abundant, continuous, increasing and heterogeneous. Extracting meaningful information from that data is useful but very difficult: rich data but poor information is a common phenomenon in the world. Data mining (DM) refer to extracting or mining useful knowledge from large amounts of data. One of the various phases of data mining is classification.

Classification is the process in which available data items are categorized into two or more categories depending on the various criterions. Methodologies in which the class label is known apriori is called supervised classification and those in which class labels are not known apriori are called unsupervised classificationor clustering [3]. Supervised classification can be further categorized as parametric and non-parametric categories. Based on whether or not the approach is based on probability distribution or density functions [4].

A well-known statistical method that can be used to solve optimization problems is Support Vector machines (SVM). The proposed methodology used here is SVM. The data items can be represented as feature vectors in a hyper plane and a line passing through the hyper plane can be used to categorize the data items into two different categories. The line can be considered a naïve form of SVM [6] [7]. The An advantage of SVM as a classification method is that in has feature extraction method in-built in its architecture. SVM is better compared to other existing classification methodologies like Naïve bayes, Artificial neural network, decision tree based classification etc. depending on previous research[8][9].

SVM which is inherently linear in nature. However by using kernel function it can be extended to non-linear space as well. In either of the approaches SVM takes a lot of time to classify the data items. SVM approach is used to solve a multi-class classification problem in this research work Its finds a suitable line which is far off from all equidistant points in the hyper plane [10-16].

SVM has numerous applicationsas inland analysis[10], species mapping[11], medicine[9], error identification[12], text and speech analysis[5,13], signal analysis[14etc... SVM is used in this research to classify raster TIFF datasets. Subsequent section explains about Literature Survey on SVM. Later Proposed methodology is explained followed by result analysis. The final section deals with conclusion followed by references.

II. Literature Survey

a) Introduction to SVM

SVM isa promising methodology which is used in various applications. It solves both two class and multi-class problems[15][16]. Problems in which input data items need to categorized into two categories are called two class problems and the ones in which data items need to be categorized into multiple classes are called multi-class problems[17]. The multi-class classification problem can be solved using divide and

Author α σ ρ: Dayananda Sagar College of Engineering, India. His current research interests include Data mining.
e-mails: tarun636@gmail.com, rajasekhar531@gmail.com, ncnaveen@gmail.com

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016

conquer approach. In this approach the problem can be divided into many two-class problems and in the future the results can be merged to arrive at the final solution to the problem.

b) SVM has two major features

Margin maximization: The classification boundary functions of SVMs maximize the margins, which leads to maximizing generalization performance [18].

SVM can be categorized as linear and non-linear SVM as in Fig 1. In linear SVM the hyper plane categorized under two different class labels by a line passing through the hyperplane[18][19][20].

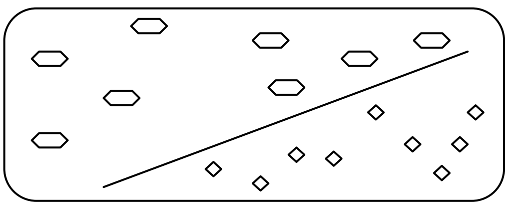


Fig.1: Linear SVM

The line representing the SVM can be denoted by (1)[21]:

$$m\theta i + c >= +1$$
$$m\theta i + c <= -1 \quad (1)$$

Data items can be represented by (2)[22]:

$$f(x) = \text{sign}(mc + b) \quad (2)$$

Where sign() represents sign function, m denotes slope and θ **happens to be the angle. Sign** function is denoted by:

$$\text{sign}(c) = \begin{cases} 1 & \text{if } c > 0 \\ 0 & \text{if } c = 0 \\ -1 & \text{if } c < 0 \end{cases} \quad (3)$$

Numerous lines might be able to split the planeas different categories but the one that maximizes the distance between itself and the data items in the two categories is known as the support vector as denoted in Figure 2.

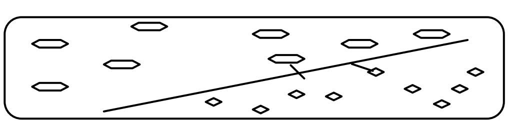


Fig. 2: Distance amid data items in a feature space

The above distance cam be denoted as:

$$M = \frac{(\theta^{+} - \theta^{-}).m}{|m|} = \frac{2}{|m|} \quad (4)$$

$$h(m) = \frac{1}{2} m^t m \quad (5)$$

subject to $y_i(m\theta_i + b) >= 1, \forall i$

The solution can be denoted with the help of a Lagrange multiplier α_i as:

$$m = \sum \alpha_i y_i \theta_i \quad (6)$$

$b = y_k - m^t x_k$ for any x_k such that Lagrange multiplier $\alpha_k \# 0$

Classifier representation[16]:

$$f(\theta) = \sum \alpha_i y_i \theta_i x + b \quad (7)$$

Systematic nonlinear classification via kernel tricks: SVMs effectively handle non-linear classifications using kernel tricks.

To improve the efficiency of the solution the input data item space can be mapped to a higher dimensional feature space denoted by [18]:

$$K(\theta i, \theta j) = f(\theta_i)^t . f(\theta_j) \quad (8)$$

The above representation is also known as a kernel function and can be denoted as [23]:

Linear Kernel function $= \theta_i^t \theta_j$

Polynomial kernel function $= (1 + \theta_i^t \theta_j)^p$

Gaussian kernel $= \exp(-\frac{|\theta i - \theta j|^2}{2\sigma^2})$

Sigmoid kernel $= \tanh(\omega_0 \theta i \theta j + \omega_1)$ (9)

c) Multi-class SVM

Multi class SVM can be categorized as one-versus-all, one-versus-one, and k-class SVM's[18].

i. *One-versus-all support vector machines*

In this approach SVM classifiers are constructed which separate one class from remaining patterns[18].

ii. *One-versus-one support vector machines*

In this approach k different SVM classifiers are constructed for every pair of classes [18].

iii. *k-Class support vector machines*

In this approachK binary classifiersare built concurrently [18].

III. Proposed Methodology

a) Datasets used

A TIFF data set is used in this research and SVM is used to classify the said data set[24].The TIFF data set is a Gujarat map which has crop coverage data across the state for rice, cotton and millet.

b) Proposed Approach

The TIFF dataset is initially pre-processed. [25]. Later Region Of Interest (ROI) is created from the image. In the next stage training set samples are selected from the ROI. Each of these training set samples correspond to a particular crop coverage in Gujarat map data set used. Three crop coverage's are used for performing the said classification. They are rice, cotton and millet crop coverage's. After the training data sample are collected the SVM classification methodology is used as explained[26]:

Begin
Step 1: Extract features from the data sets
Step 2: Select feature vectors and form the input data set
Step 2: Start dividing the input data set into two sets of data corresponding to two different categories
Step 3: If a data item is not assigned any of theregions mentioned then add it to set of support vectors V
Step 4: end loop
End

Finally the built model is validated against the test data set. Herein the test data set under consideration is the crop coverage area that is not covered as part of the selected training data set sample. One of the key steps involved in the classification process is feature extraction as mentioned below:

Energy (E): It facilitates in computing homogeneousness in the data set and is denoted by:

$$E = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} (p(i,j))^2 \tag{9}$$

Contrast(C): Contrast helps identify local data set variation and is denoted as:

$$C = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} (i-j)^2 p(i,j) \tag{10}$$

Inverse difference moment (IDM): Local texture alterations can be located using:

$$IDM = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \frac{1}{1+(j-2)^2} p(i,j) \tag{11}$$

Entropy (S): The data set complexity can be computed by:

$$S = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} p(i,j) \log(p(i,j)) \tag{12}$$

Where μ_k and mxn are the mean and size of the blockB_k

Spatial Frequency (SF): Frequency changes in the data set can be computed using:

$$SF = (RF)^2 + (CF)^2$$

Where

$$RF = \sqrt{\frac{1}{m\times n} \sum_{i=1}^{m} \sum_{j=2}^{n} [I(i,j) - I(i,j-1)]^2} \quad \text{and}$$

$$CF = \sqrt{\frac{1}{m\times n} \sum_{i=1}^{m} \sum_{j=2}^{n} [I(i,j) - I(i-1,j)]^2}$$

Variance (V): Level of focus in a data set can be computed using:

$$V = \frac{1}{m\times n} \sum_{i=1}^{m} \sum_{j=1}^{n} (I(i,j) - \mu)^2 \tag{16}$$

Where μ is the mean value of the block image and $m \times n$ is the image size Energy of Gradient (EOG): Measure of focus can also be computed using:

$$EOG = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} (f_i^2 + f_j^2) \tag{17}$$

Where, $f_i = f(i+1, j) - f(i, j)$
$f_j = f(i, j+1) - f(i, j)$

IV. Result Analysis

a) *Environment Setting*

Agricultural map of Gujarat was used as a dataset to perform the said classification. A region of interest (ROI) was extracted from the map that acted as a training data and it was validated against the complete data segment pertaining to a particular crop in the map. Environment in which the research was undertaken is shown in Table 1[27].

Table.1 : Environment Setting

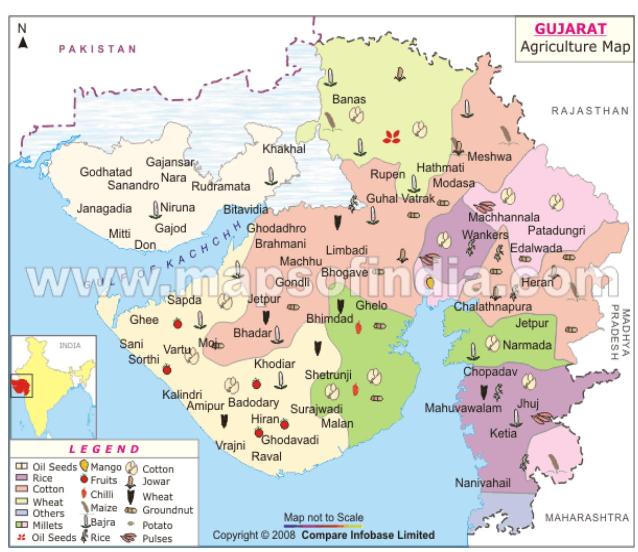
Item	Capacity
CPU	Intel CPU @2 GHz processor
Memory/OS	4GB /WIN 7
Applications	Monteverdi

b) *Result Analysis*

The ratio of correctly classified and uncorrectly classified data items can be represented using confusion matrix view as mentioned in Table 2. It helps measure the efficacy of the performed classification. Classification results is given in Figure 4.

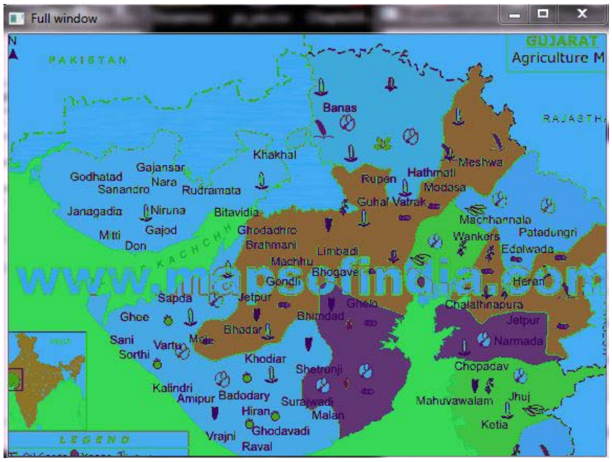
Table. 2 : Confusion Matrix

	Classification result	
	No Event	Event
No Event	True Negative(TN)	False Positive(FP)
Event	FalseNegative(FN)	True Positive(TP)



(a)

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016



(b)

(c)

Fig. 3 : (a) ROI from the TIFF data set. (b) Classified image with various crop coverage in the state of Gujarat displayed in various colors(Rice-Brown, Millets-Violet, Cotton-Brown). (c) Edge Feature extracted image of the crop data set

Accuracy and kappa statistics are used to measure the efficacy of the classification methodology used.Thes parameters are denoted by equations (18) and (19)[28][29][30]:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{(\text{TP}+\text{FN}+\text{FP}+\text{TN})} \times 100 \qquad (18)$$

Kappa statistics=Sensitivity + Specificity - 1 (19)

Confusion matrix in research is mentioned in Table 3.

Table. 3 : Confusion Matrix

Prediction	Reference		
	Rice	Millets	Cotton
Rice	14	0	0
Millets	0	16	0
Cotton	0	0	11

Accuracy and kappa statistics obtained while classifying TIFF data set are mentioned in Table 4.

Table. 4 : Performance measures for TIFF dataset

Data set type	Accuracy	Kappa Statistics
Raster TIFF datasets	100	100

V. Conclusion

SVM classification methodology is used to classify the Gujarat map TIFF data set.Accuracy and kappa statistics parameters are used to measure the efficacy of the said method and the values obtained for the said evaluation parameters prove beyond doubt that the method used classifies the data set with better accuracy.

References Références Referencias

1. Kenichi Tatsumi, Yosuke Yamashiki, Miguel Angel Canales Torres, Cayo Leonidas Ramos Taipe, Crop classification of upland fields using Random forest of time-series Landsat 7 ETM+ data, Computers and Electronics in Agriculture, Volume 115, July 2015, Pages 171-179, ISSN 0168-1699,http://dx.doi.org/10. 1016/j.compag.2015.05.001.
2. Henrik Skov Midtiby, Björn Åstrand, Ole Jørgensen, Rasmus Nyholm Jørgensen, Upper limit for context–based crop classification in robotic weeding applications, Biosystems Engineering, Available online 15 February 2016, ISSN 1537-5110, http://dx.doi.org/10.1016/j.biosystemseng.2016.01.012.
3. M. Pérez-Ortiz, J.M. Peña, P.A. Gutiérrez, J. Torres-Sánchez, C. Hervás-Martínez, F. López-Granados, A semi-supervised system for weed mapping in sunflower crops using unmanned aerial vehicles and a crop row detection method, Applied Soft Computing, Volume 37, December 2015, Pages 533-544, ISSN 1568-4946, http://dx.doi.org/10.1016 / j.asoc.2015.08.027.
4. Sybrand Jacobus Muller, Adriaan van Niekerk, An evaluation of supervised classifiers for indirectly detecting salt-affected areas at irrigation scheme level, International Journal of Applied Earth Observation and Geoinformation, Volume 49, July 2016, Pages 138-150, ISSN0303-2434,http://dx.doi.org/10.1016/j.jag 2016.0 2005.
5. Murat Olgun, Ahmet Okan Onarcan, Kemal Özkan, Şahinlşik, Okan Sezer, Kurtuluş Özgişi, Nazife Gözde Ayter, Zekiye Budak Başçiftçi, Murat Ardiç, Onur Koyuncu, Wheat grain classification by using dense SIFT features with SVM classifier, Computers and Electronics in Agriculture, Volume 122, March 2016, Pages 185-190, ISSN 0168-1699, http://dx.doi.org/10.1016/j.compag.2016.01.033.
6. P.J. García Nieto, E. García-Gonzalo, J.R. Alonso Fernández, C. Díaz Muñiz, A hybrid PSO optimized

SVM-based model for predicting a successful growth cycle of the Spirulinaplatensis from raceway experiments data, Journal of Computational and Applied Mathematics, Volume 291, 1 January 2016, Pages 293-303, ISSN 0377-0427, http://dx.doi.org/10.1016/j.cam.2015.01.009.

7. Najafi, B. Ghobadian, A. Moosavian, T. Yusaf, R. Mamat, M. Kettner, W.H. Azmi, SVM and ANFIS for prediction of performance and exhaust emissions of a SI engine with gasoline–ethanol blended fuels, Applied Thermal Engineering, Volume 95, 25 February 2016, Pages 186-203, ISSN 1359-4311, http://dx.doi.org/10.1016/j.applthermaleng.2015.11.009.
8. Yang Long, Fan Zhu, Ling Shao, Recognising occluded multi-view actions using local nearest neighbour embedding, Computer Vision and Image Understanding, Volume 144, March 2016, Pages 36-45, ISSN 1077-3142, http://dx.doi.org/10.1016 / j.cviu.2015.06.003.
9. Sunil S. Morade, Suprava Patnaik, Comparison of classifiers for lip reading with CUAVE and TULIPS database, Optik - International Journal for Light and Electron Optics, Volume 126, Issue 24, December 2015, Pages 5753-5761, ISSN 0030-4026, http://dx.doi.org/10.1016/j.ijleo.2015.08.192.
10. Yang Shao, Ross S. Lunetta, Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 70, June 2012, Pages 78-87, ISSN 0924-2716, http://dx.doi.org /10.1016/j. isprsjprs. 2012.04. 001.
11. Hongji Lin, Han Lin, Weibin Chen, Study on Recognition of Bird Species in Minjiang River Estuary Wetland, Procedia Environmental Sciences, Volume 10, Part C, 2011, Pages 2478-2483, ISSN 1878-0296, http://dx.doi.org/10.1016/j.proenv.2011.09.386.
12. Baoping Tang, Tao Song, Feng Li, Lei Deng, Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine, Renewable Energy, Volume 62, February 2014, Pages 1-9, ISSN 0960-1481, http://dx.doi.org/10.1016/j.renene.2013.06.025.
13. A.D. Dileep, C. Chandra Sekhar, Class-specific GMM based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines, Speech Communication, Volume 57, February 2014, Pages 126-143, ISSN 0167-6393, http://dx.doi.org/10.1016/ j.specom.2013.09.010.
14. Rao, T.; Rajasekhar, N.; Rajinikanth, T.V., "Drainage water level classification using support vector machines," Engineering (NUiCONE), 2013 Nirma University International Conference on, vol., no., pp.1, 6, 28-30 Nov. 2013doi: 10.1109/NUiCONE.2013.6780068.
15. Lam Hong Lee, Rajprasad Rajkumar, Lai Hung Lo, Chin Heng Wan, Dino Isa, Oil and gas pipeline failure prediction system using long range ultrasonic transducers and Euclidean-Support Vector Machines classification approach, Expert Systems with Applications, Volume 40,Issue 6,May 2013, Pages 1925-1934,ISSN 0957-4174, http:// dx.doi.org/10.1016 /j.eswa.2012.10.006.
16. Elias Zintzaras, Axel Kowald, Forest classification trees and forest support vector machines algorithms: Demonstration using microarray data, Computers in Biology and Medicine, Volume 40, Issue5, May 2010, Pages 519-524, ISSN0010-4825.
17. Omar Abdel Wahab, Azzam Mourad, Hadi Otrok, Jamal Bentahar, CEAP: SVM-based intelligent detection model for clustered vehicular ad hoc networks, Expert Systems with Applications, Volume 50, 15 May 2016, Pages 40-54, ISSN 0957-4174, http://dx.doi.org/10.1016/j.eswa.2015.12.006.
18. Sarah M. Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, Christopher Leckie, High-Dimensional and Large-Scale Anomaly Detection using a Linear One-Class SVM with Deep Learning, Pattern Recognition, Available online 14 April 2016, ISSN 0031-3203, http://dx.doi.org/10.1016/j.patcog.2016.03.028.
19. Yi Yan, Wei Kai Zheng, JianBao, Ran Liu, An enhanced M-ary SVM algorithm for multi-category classification and its application, Neurocomputing, Volume 187, 26 April 2016, Pages 119-125, ISSN 0925-2312, http://dx.doi.org/10.1016/j.neucom.2015.08.101.
20. Azadeh Sadat Mozafari, Mansour Jamzad, A SVM-based model-transferring method for heterogeneous domain adaptation, Pattern Recognition, Volume 56, August 2016, Pages 142-158, ISSN 0031-3203, http://dx.doi.org/10.1016/j. patcog.2016.03.009.
21. Muhammad Tahir, Asifullah Khan, Protein subcellular localization of fluorescence microscopy images: Employing new statistical and Texton based image features and SVM based ensemble classification, Information Sciences, Volume 345, 1 June 2016, Pages 65-80, ISSN 0020-0255, http://dx.doi.org/10.1016/j.ins.2016.01.064.
22. Yan Leng, Chengli Sun, Xinyan Xu, Qi Yuan, Shuning Xing, Honglin Wan, Jingjing Wang, Dengwang Li, Employing unlabeled data to improve the classification performance of SVM, and its application in audio event classification, Knowledge-Based Systems, Volume 98, 15 April 2016, Pages 117-129, ISSN 0950-7051, http://dx.doi.org/10.1016/j.knosys.2016.01.029.

23. Chao Zhou, Kunlong Yin, Ying Cao, Bayes Ahmed, Application of time series analysis and PSO–SVM model in predicting the Bazimen landslide in the Three Gorges Reservoir, China, Engineering Geology, Volume 204, 8 April 2016, Pages 108-120, ISSN 0013-7952, http://dx.doi.org/10.1016/j.enggeo.2016.02.009. http://www.infobase.co.in.
24. D.Lu& Q. Weng (2007):A survey of image classification methods and techniques for improving classification performance, International Journal of Remote Sensing, 28:5,823-870,http://dx.doi.org/10.1080/01431160600746456.
25. S. N. Jeyanthi, Efficient Classification Algorithms using SVMs for Large Datasets, A Project Report Submitted in partial fulfillment of the requirements for the Degree of Master of Technology in Computational Science, Supercomputer Education and Research Center, IISC, BANGALORE, INDIA, June 2007.
26. Jennifer A. Taylor, Alicia V. Lacovara, Gordon S. Smith, Ravi Pandian, Mark Lehto, Near-miss narratives from the fire service: A Bayesian analysis, Accident Analysis & Prevention, Volume 62, January 2014, Pages 119-129, ISSN 0001-4575, http://dx.doi.org/10.1016/ j.aap.2013.09.012.
27. David J. Rogers, Jonathan E. Suk, Jan C. Semenza, Using global maps to predict the risk of dengue in Europe, ActaTropica, Volume 129, January 2014, Pages 1-14, ISSN 0001-706X.
28. Rafael Pino-Mejías, María Dolores Cubiles-de-la-Vega, María Anaya-Romero, Antonio Pascual-Acosta, Antonio Jordán-López, Nicolás Bellinfante-Crocci, Predicting the potential habitat of oaks with data mining models and the R system, Environmental Modelling & Software, Volume 25 ,Issue 7, July 2010, Pages 826-836, ISSN 1364-8152, http://dx.doi.org/10.1016/j.envsoft.2010.01.004.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Bottom-Up Update Mechanism for Re-Structured Complete Binary Trees

By Mevlut Bulut

University of Alabama, United States

Abstract- This paper introduces a bottom-up update mechanism together with a non-recursive initial update procedure that reduces the required extra memory space and computational overhead. A new type of tree is defined based on a different geometrical interpretation of Complete Binary Trees. The new approach paves the way for a special and practical initialization of the tree, which is a prerequisite for an implementation of unilateral update operation. The details of this special initialization and the full update procedures are given for Complete Binary Trees. In addition, a comparison is on is made between the introduced update method and the bilateral update methods in terms of different performance related metrics.

Keywords: *data structure, complete binary tree, CBT, sCBT, unilateral update, bottom-up update, replacement selection.*

GJCST-C Classification : *I.1.2, I.2.2*

BOTTOMUPUPDATEMECHANISMFORRESTRUCTUREDCOMPLETEBINARYTREES

Strictly as per the compliance and regulations of:

Bottom-Up Update Mechanism for Re-Structured Complete Binary Trees

Mevlut Bulut

Abstract- This paper introduces a bottom-up update mechanism together with a non-recursive initial update procedure that reduces the required extra memory space and computational overhead. A new type of tree is defined based on a different geometrical interpretation of Complete Binary Trees. The new approach paves the way for a special and practical initialization of the tree, which is a prerequisite for an implementation of unilateral update operation. The details of this special initialization and the full update procedures are given for Complete Binary Trees. In addition, a comparison is made between the introduced update method and the bilateral update methods in terms of different performance related metrics.

Keywords: data structure, complete binary tree, CBT, sCBT, unilateral update, bottom-up update, replacement selection.

I. Introduction

At the center of the modern programming paradigm rises the art of obtaining the maximum performance out of a given computer system with limited resources, e.g. computational power, memory or I/O operation capabilities. In designing comparison based algorithms such as searching and sorting, in order to circumvent these limitations, tree formation was suggested a long time ago[1] and it has been widely used since then. The main idea of forming a tree or treating a given array as a tree is to minimize the number of comparisons as close to the theoretical minimum as possible. Although there are many different techniques for the formation (or branching), setup (usage of nodes and node hierarchy), traversing (top-down, bottom-up; preorder, in order, etc.), and initialization of trees (recursive and iterative) new attempts are still being made to improve the efficiencies of these algorithms by optimizing the usage of the limited resources.

As explained in the next section, a new definition for the root node together with a new geometric interpretation of tree formation is proposed. Although the introduced novelties do not change the number of comparisons for the basic tree operations, it brings considerable reduction in required memory space, computational overhead, and number of accessed memory locations.For all the graphical descriptions, only Complete Binary Tree (CBT) structures will be used throughout the article, however the introduced concepts can be applied to other types of trees as well.

The bottom-up update mechanism can simply be described as a unilateral traversing of the nodes from a leaf host to the root. Unlike the bilateral update mechanism, which is based upon comparing two sister node contents followed by the registration of the winner in the parent node, the unilateral update mechanism requires that the overall winner of the previously done consecutive comparisons should be compared to the content of the parent node. If the parent node content is not the winner of this comparison, then the consecutive parent nodes are checked until a parent node content wins, at which point the winner item and the parent node content are swapped. The iteration of this procedure goes on until the root node is reached, where the global winner is registered.

This article introduces a modified bottom-up update mechanism which differs from the previously suggested unilateral implementations[2] in terms of the required auxiliary memory space, the initial update technique, and the overhead reduction during the update operations thanks to the elimination of the redundant nodes from CBTs. As a result, the overall implementation of a bottom-up update operation gets simpler, lighter, and faster.

II. Geometric Definition

Analogous to real trees, the definition of an abstract tree with a stem is suggested (Figure-1).The zeroth node is placed at the end of the stem and utilized as the root of the tree. A CBT with such a structure can be called a stemmed CBT (like most of the trees in the real world). Any Stemmed CBT (sCBT) can be decomposed into smaller sCBTs. In this regard, the smallest sCBT shell encompass two nodes, one of which characterizes the body of the tree and the other one is the root. This definition leads to a new way to compose and decompose a given tree. Figure-2depicts how two minimal sCBTs are combined together. One can decompose a given sCBT along a path from a leaf node to the root. In cases, the s CBT is utilized for replacement selection[3] or priority queue applications [4] then the logic dictates the path of the overall winner to be chosen as the decomposition path. The decomposition will be outlined in the 'initial update' section.

Author: *615 Idlewild Cir Apt 25 Birmingham, Alabama, USA.*
e-mail: mevlutbulut@yahoo.com

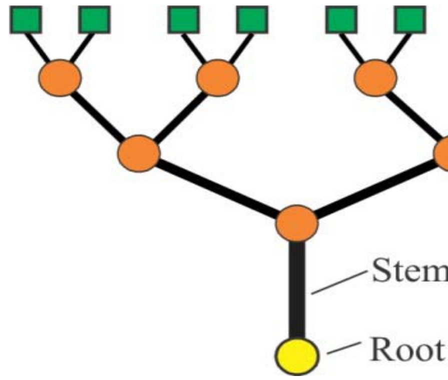


Figure 1: Proposed abstract tree

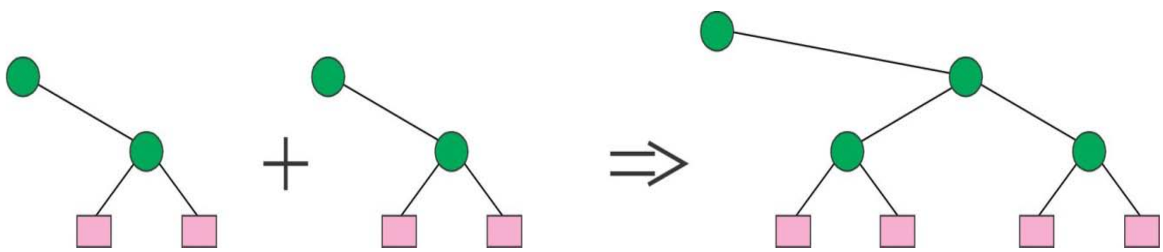


Figure 2 : Two minimal sCBTs are combined through modifying one bond and reforming another, without introducing any new node. Here, note that two regular CBTs cannot be combined without adding a new node.

III. Unilateral Update Versus Bilateral Update

A CBT setup with loser elements rather than winner elements was first suggested by[5] with a coined name 'loser tree', as opposed to 'winner tree', based upon the fact that each and every key appearing in an internal node is a loser exactly once, champion being the only exception. Although they are all losers exactly once, they are the winners of all comparisons up to their current levels. This property is not so different from the case of so called 'winner tree' setup. The logic is the same: both of them promote the winner towards the root. Therefore, there is no point for calling one of them a 'loser tree' and the other one a 'winner tree'. The difference between these two tree setups is that their geometries are different. The difference is dictated by the geometry not by the selection procedure. Therefore, 'winner tree' and 'loser tree' naming convention is abandoned here, instead CBT and sCBT are used to imply the two different geometries and the corresponding bilateral and unilateral update mechanisms respectively.

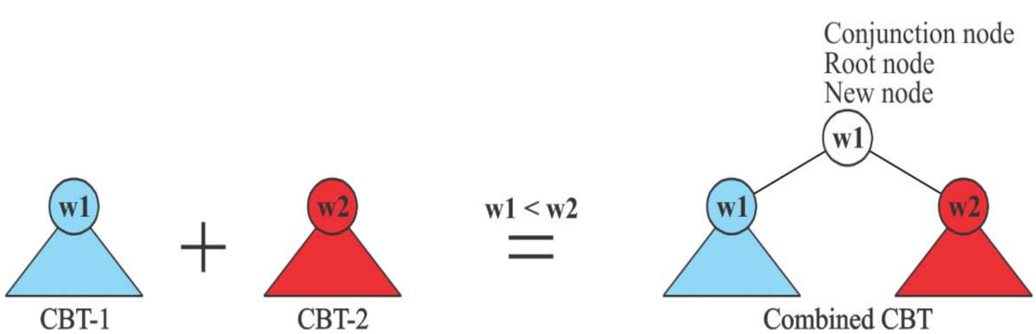


Figure 3 : The winners of two CBTs are compared and the winner (in this case the smaller) is written into the conjunction node serving as the root of the combined CBT. During this operation, three nodes are accessed and the root node should be introduced as a new node.

The comparison operation can be regarded as a procedure to compose two sub-trees. Figure-3 and Figure-4 show how a comparison between the winners of two sub-trees is implemented and how the winner is promoted in CBT and sCBT cases respectively. Note that the procedure of combining two CBTs is not possible without adding a new node, whereas in the sCBT case, there is no need for a new node.

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016

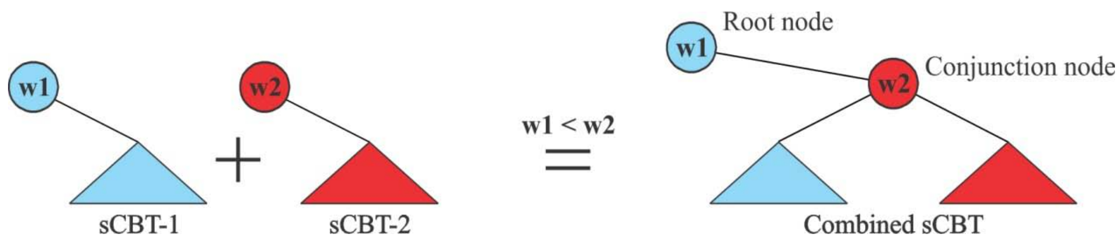


Figure 4 : When combining two sCBT swith the 'smaller wins' rule, we find the winner of the two keys hosted by the two roots then register the winner at the root of the combined sCBT, leaving the loser one in the conjunction node. During this operation, only two nodes are accessed. No extra node is required.

IV. Initial Update

An sCBT is said to be properly initialized only if every node along the winner path hosts the winner of the corresponding sub-sCBT (a node can be the root of either the left or the right block; whichever side hosts the content of the root constitutes the body of the sub-sCBT) and every sub-sCBT also exhibits this same property. Figure-5 depicts the way we can see a properly initialized sCBT. We regard the initialized sCBT as consisted of smaller sCBTs along the path of the winner key, from the winner leaf to the root. All the node contents that lose against the winner are the winners of their own sub- sCBTs.

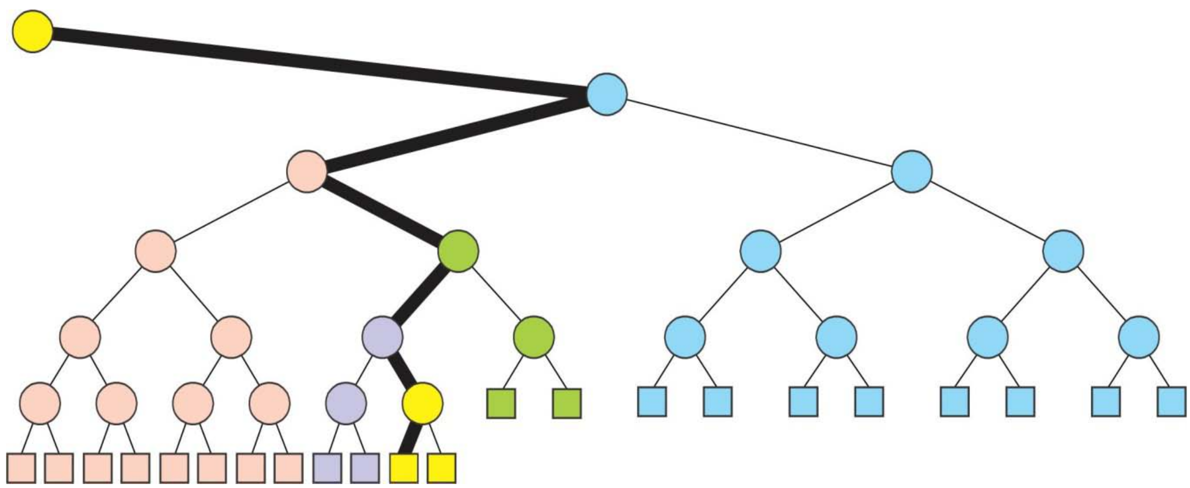


Figure 5 : An sCBT as comprised of smaller sCBTs along the winner path from the leaf node to the root. All the nodes along this path should host the winners of their own sub-sCBTs.

The new idea about initializing an sCBT is that the global tree can be thought of as a composition of already initialized smaller sCBTs. There are two different ways an initialized sCBT can be achieved:

1. Start with the maximum number (N/4) of minimal sCBTs at the lowest level of the tree; grow them independently while merging them as necessary.
2. Start with a minimal sCBT, enlarge it by adding two new leaves and update the obtained sCBT, and repeat this operation until the targeted sCBT size is reached.

In the first way, initialization starts with the non-interfering minimal sCBTs at the bottom of the sCBT and proceeds upward by growing and/or combining them until the whole tree size is reached. Following the initial update, the root (the zeroth node) contains the index of the winner element of the given key array and all the other nodes contain the indexes of the winner elements of their own sub-sCBTs. Figure-6 visualizes this method by the color coded update paths. Initializing an sCBT consisting of just two nodes requires only one comparison between the two leaves hanging from the only body node of this sCBT. After the comparison, the loser is stored in the lower node, while the winner is stored in the upper node. When all depth-1(below the root node, there is only one node) sCBTs are initialized, then the initialization of depth-2 sCBTs starts. To initialize a depth-2 sCBT, we start comparing the two new leaves that come into the picture when we grow the previously initialized depth-1 sCBT into a depth-2 sCBT. The loser of this comparison is stored into the first parent of these leaves and the winner is kept at hand to be compared to the content of the next parent node (which was the winner of the depth-1 sCBT). If it loses the comparison against the content of the next parent

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016

node, they are swapped and the one next parent node will host the winner leaf index of the whole depth-2 sCBT (green update paths inFigure-6). Then the procedure goes on to depth-3, depth-4, and so on until the whole tree is initialized.

In this way, all sub-sCBTs with the same depth can be handled in a sub- loop, allowing any depth specific variable to be calculated faster. One such variable is the index of the root node of a givensub-sCBT, which can be found by right shifting the index of the leftmost bottom node of that sub-sCBT until the least significant bit disappears. Here is a suggested C++ code to find the root index for a given leftmost node:

```
unsigned long level;
_Bit Scan Forward(&level, leftmost Node);
root= leftmostNode>> (level+1);
```

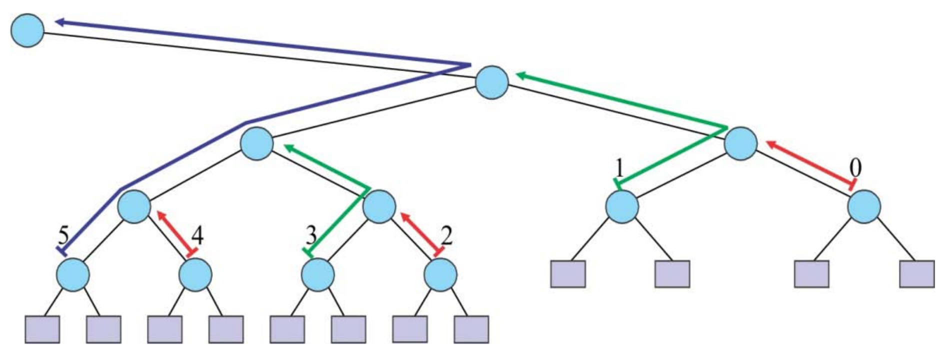


Figure 6 : A graphical depiction of two different ways to implement the initial update operation for a given sCBT. The first way is to initialize the constituent sCBTs from the smallest to the largest as indicated by the color coding in the figure, in the order of red, green, and blue. The second way is to start updating them from right to left as identified by the ascribed counting numbers from zero to five in the figure.

Figure-7 shows that the indexes of the root nodes of the same depth sCBTs form a sequential array when they are traversed from the end of the tree array towards its head (in this example the sequential array is 5; 4; 3). This gives an easy way of finding the root indexes during the initial update. The provided C++ code following the 'Redundant Tree Nodes' section uses the advantage of this first technique. As an example, Figure-7 depicts an sCBT with 12 lexical leaves. By following the sub-figures from a) to d), the initialization of this sCBT can be followed step by-step.

The second way for initial update requires the initialization of sub-sCBTs starting from rightmost depth -1 sCBT and growing/going to the left while initializing the next available size/initializable sCBT on the way. Figure-6 shows the sequence of these consecutive update paths by ordinal numbers from zero to five for the initialization of the depicted sCBT.

The advantage of the second technique is that all sub sCBTs can be processed in a single loop. Depending on the node hierarchy being used, there are some cases where this second technique becomes faster and easier to implement. However, for the simple node hierarchy used throughout this article, the implementation of the first technique proves to be more efficient.

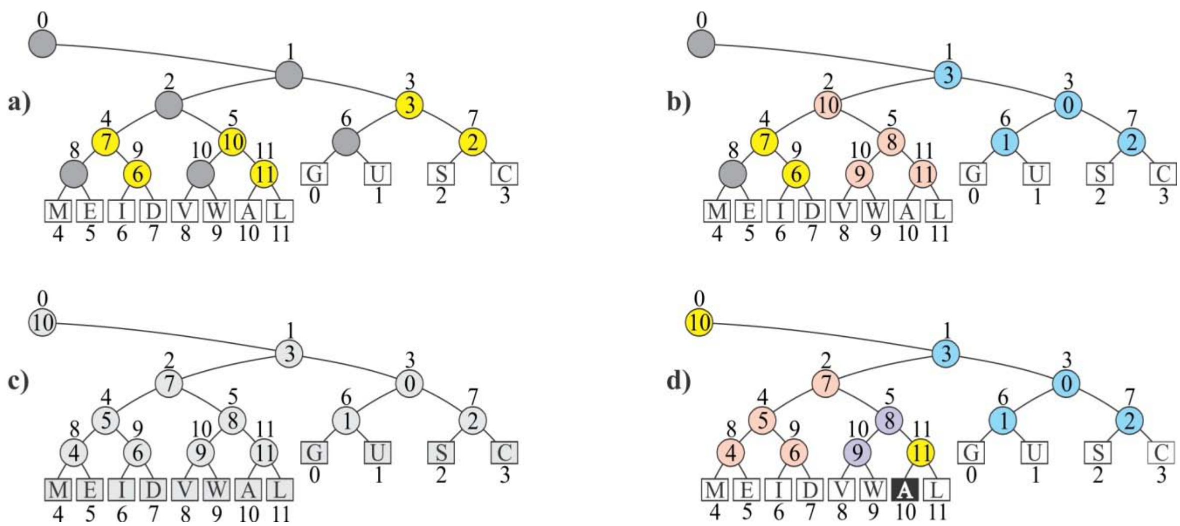


Figure 7: A lexical array of size 12 is used as the leaves of the sCBT in order to demonstrate the introduced initial update procedure using the first of the two suggested methods. a) Only the sub-trees with a depth of one are initialized, in b) the ones with a depth of two and in c) with a depth of four (which is the whole sCBT) are initialized. Here there is no sub-sCBT with a depth of three. In d) the decomposition of the sCBT along the winner path is visualized by using different colors for each sub-sCBT.

V. Redundant Tree Nodes

If a tree node is written but never read, then writing that node is considered redundant. In the case of sCBT and the proposed unilateral update mechanism combination, the bottom nodes, or in other words, the immediate parent nodes of the leaves are all redundant. This is because they host the loser keys not the winner ones. Thus, we can implement the sCBT and the proposed update mechanism by using only N/2 tree nodes. After comparing the sister leaves, we register only the winner to the grandparent node (we can think of the immediate parent node as a ghost node). Figure-8 displays a worked out example of such an sCBT using a lexical key array.

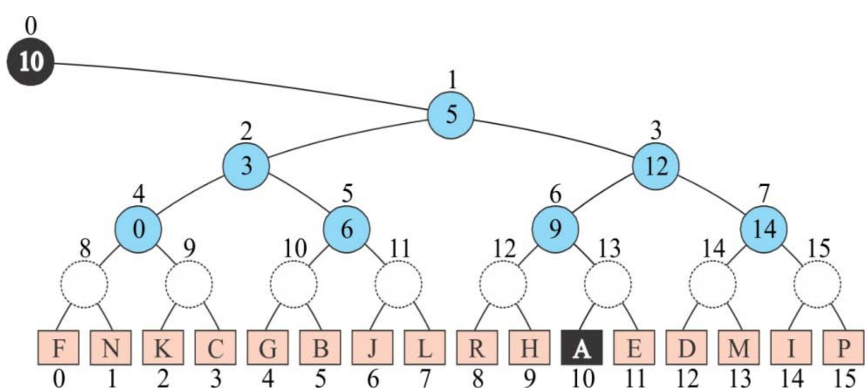


Figure 8 : Leaving out the redundant tree nodes. During the proposed unilateral update procedure, the lowest level tree nodes are not read at all, therefore there is no point of using them to write the indexes of the looser leaves. This reduces the number of required nodes to implement an sCBT to N/2.

VI. Update Mechanism

When an update is required after a new key is assigned to the winner leaf, a unilateral update procedure is implemented: First, the new key is compared to its sister, and then the winner of this comparison is kept at hand as the new winner candidate. Then this new candidate is compared to the hosted keys along the winner path. Wherever the key at hand loses the comparison, it is registered there and the previously registered key in that node is taken as the new winner candidate.This procedure goes on until the root node is reached, where the final winner is registered.

The following is a working C++ code for the proposed initial update and the proposed unilateral update methods. Initial update method follows the first technique explained in'Initial Update'section. Although the graphical examples up to this point all use 'even number of leaf nodes', the provided code takes care of odd cases by the additional lines marked with (**). If N is guaranteed to be even, then these lines can be safely removed from the code.

```
// int N; //the size of the keys array.
//float*Keys; // the given array containing the keys.
//int offset=N,*sCBT=new int[(N+1)>>1];//"+1" is necessary for odd N cases.
//sCBT:auxiliary integer array used for the formation of stemmed complete binary tree.
// int max ID= N-1;
Void Initial Update ()
{
Int h = N-1;   //h: host, immediate parent node for a pair of leaves.
If (N&1) {sCBT[h>>1]= h; h--; offset++;} // (**)

For (int jump = 2, UpNode = h>>1, Tail= maxID>>1; ;UpNode --)
   {
          Int w= 2*h - offset; if(Keys[w^1] < Keys[w]) w ^= 1;
          For (int n= h>>1; n >UpNode; n>>= 1)
if (Keys[sCBT [n]] < Keys[w]) swap(sCBT[n],w);
          sCBT [UpNode]= w;

          h-= jump;  if(h > Tail) continue;
          h<<= 1;
          if(UpNode> 1) jump<<=1; else{ if(UpNode ==0) break; if(h < Tail) h <<= 1;}
   }
}
//w: winner, it was the index of the previous winner key, when a new value is assigned//to the winner key, the sCBT
// should be updated accordingly. This update procedure  will provide the index of the new winner key.
voidUpdate_sCBT()
{
int w= *sCBT;
if((w^1)!=N) // (w^1)==N can happen only ifN is odd. (**)
          if(Keys[w^1] <Keys[w]) w ^= 1;//loser doesn't need to be registered anywhere.

for(int node= (w + offset)>>2; node> 0 ;node >>= 1)
{
          Int const guest= sCBT [node];//guest: index of the registered key in the node.
          if (Keys[guest] <Keys[w]){sCBT [node]= w; w= guest;}
}
*sCBT= w;
}
```

VII. Results and Discussion

The benefits of the introduced unilateral update mechanism compared to the bilateral update mechanism can be itemized as follows:

1. Every key index appears in the tree at most once. More precisely, half of the key indexes will appear in the tree only once, while the other half will not have any appearance in the reduced sCBT approach. If there is a necessity for a specific application, sCBT can also be formed using N nodes, in which case, the entire key indexes will appear in the tree once and only once. In the bilateral update, some leaves are registered as many as log N times while some others are not registered at all.
2. Except for the computation (or identification) of a leaf level sister, neither is there a need for any sister node computation nor a need for accessing its content.
3. Reduction in the number of read nodes by 50%.
4. During a unilateral update, the number of writes can be between 1 and log N depending on the results of the comparisons, whereas during a bilateral update, log N writes are necessary for every update operation. The number of writes in bilateral updates can be reduced by checking the previous guest

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016

index of a node in order to avoid re-storing the index which is already there. But this will bring extra overhead of log N integer index comparisons.

5. The required number of tree nodes is reduced by 50% in comparison to the required number of nodes for the bilateral update mechanism implemented on a CBT.

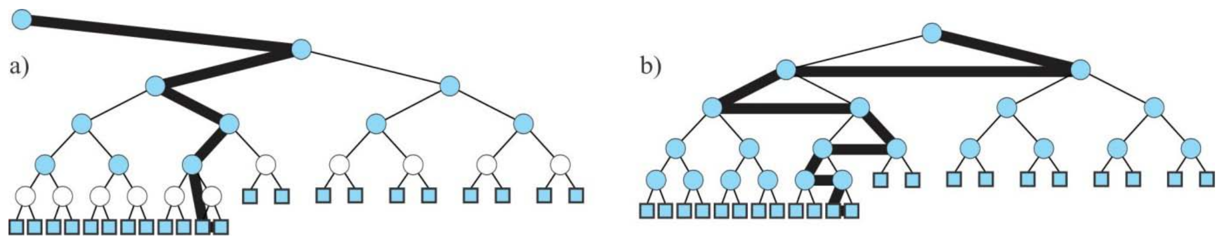


Figure 9 : Nodes visited during a) unilateral update on an sCBT, b) bilateral update on a CBT.

Table-1 shows the algebraic quantities for the two different update mechanisms in five different metrics, whileFigure-9 depicts the visited nodes and the update paths side by side for these two update mechanisms, in order to help visualize the differences.

Table 1 : Comparison between unilateral update and bilateral update for a full update operation on a complete binary tree comprising of N leaves.

Type of Update \ #of	Required Tree Nodes	Comparisons	Accessed Nodes	Sister Node Computations	writes	reads
Unilateral Update	N/2	Log N	Log N	0	1≤ ≥Log N	Log N
Bilateral Update	2N	Log N	2Log N	Log N	Log N	2Log N

In terms of initial update cost, there is not much difference between the unilateral and the bilateral update methods. Both of them require exactly N comparisons. However, the number of accessed nodes, writes, and reads are different. In the case of a bilateral update on a CBT, N nodes are accessed, N reads and N writes are implemented. On the other hand, the initial update of an sCBT accesses N/2 nodes, and implements N/2 reads and a minimum of N/2 writes (in the worst case scenario, number of writes can be equal to N if all the comparisons require the swapping of node content and the winner candidate at hand). Table-2 summarizes these quantities.

Table 2 : Comparison between unilateral and bilateral initial update operations on a complete binary tree comprising of N leaves.

Type OfInitial Update \ #of	Comparisons	Accessed Nodes	writes	reads
Unilateral Initial Update	N	N/2	N/2≤ ≥ N	N/2
Bilateral Initial Update	N	N	N	N

VIII. Numerical Comparisons

A test run for a given number of keys was repeated 10 times but only the averages were used for graphing. For the obtained numerical results, the maximum encountered error (standard deviation divided by average) was less than 3%. The computer used for the presented results was a Dell OptiPlex 790 with an Intel Core i5-2400 CPU @3.10 GHz and 8GB RAM. The operating system of the test computer was Windows 7 enterprise 64-bit edition. For coding, Visual C++ 2010 programming environment was used. The compilations were done with SSE2 and maximize-speed options enabled.

A uniform distribution (0.0<x<1.0) was used to generate random key values for the hold model[6]. CBTs were constructed using the given number of initial keys. Then a loop of N hold operations was performed for timing. Timing was achieved by counting the total number of CPU cycles between the beginning and the end of the computational block by using the CPU clock register. The accumulated number of CPU cycles was divided by number of given keys to get an average cost for one hold operation. The presented empirical results have been scaled to the scores of the implementations running on the same test system based on reference CBT that Marin used [6].

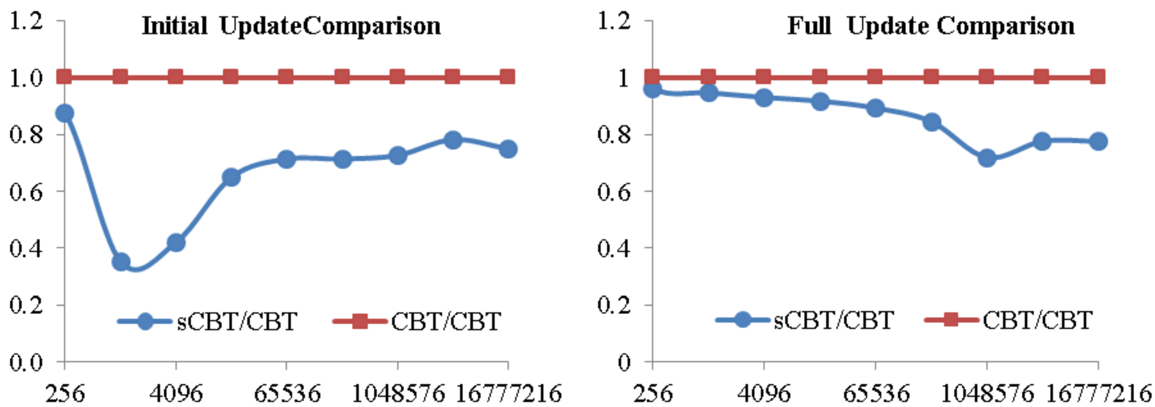


Figure 10 : Comparison of numerical performance results for the introduced unilateral update method and the reference bilateral update method. The left graph shows comparison results of unilateral and bilateral initial update methods while the right one shows the results of full update operations for the update mechanisms. The horizontal axis shows the number of keys, while the vertical axis shows the test scores scaled to the score of the reference structure (CBT) for the same test. The maximum number of keys used for the tests is 2^{24}.

[Fig. 10] presents the obtained results for the test system in two categories: Initial update comparisons and full update comparisons. In the case of initial update comparisons, introduced unilateral initial update performs at least 20% better than bilateral initial update except when the number of keys is very small. This should be because of the smaller footprint of the bilateral initial update code as can be seen in the fallowing lines compared to the code for unilateral initial update given earlier.

```
//intN; //the size of the keys array.
//float *Keys; // the given array containing the keys.

//int*CBT=new unsigned [2*N];
//CBT:auxiliary integer array used for the formation of complete binary tree.
voidInitialUpdate() //Initial Update CBT
{
for(int n=0; n < N; n++) {CBT[N+n]= n;}
for(intn=2*N-1; n > 1; n -= 2)
{if(Keys[CBT[n]] < Keys[CBT[n-1]]) CBT[n/2]= CBT[n];
else CBT[n/2]= CBT[n-1];}
}
```

Full update comparisons show that the superiority of unilateral update gets better as the number of keys increases and it stabilizes around 20%for cases the bulk of the data remains outside the cache memory.

IX. Conclusion

A new graphical formation of binary trees is introduced. As a result of this formation, binary trees can be decomposed or composed without adding or deleting any nodes regardless of their leaf and node hierarchies. This new formation leads to a unilateral bottom-up update mechanism that promises acceleration by reducing computational overhead, auxiliary memory field, and memory operations. When the suggested sCBT structure is used to produce the initial runs for external sorting [7], it will increase the average length of the runs, since larger size trees can be established in a given amount of cache memory thanks to the elimination of redundant tree nodes. The suggested unilateral update mechanism can be coupled with different leaf hierarchies such as Super CBT [8] and/ or with different node hierarchies such as hardware conscious trees[9].

References Références Referencias

1. Kirchhoff G. Ueber die auflösung der gleichungen auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. Ann Phys 1847; 148: 497-508.
2. Sahni S. Structures, algorithms, and applications in C++. 2nd ed. Summit, NJ, USA: Silicon Press, 2005.
3. Friend EH. Sorting on electronic computer systems. J ACM 1956; 3: 134-168.
4. Marín M, Cordero P. An empirical assessment of priority queues in event-driven molecular dynamics simulation. Comput Phys Commun 1995; 92: 214-224.
5. Knuth DE. The art of computer programming, 2nd ed. San Francisco, CA, USA: Addison-Wesley, 1998.

6. Marín M. An empirical comparison of priority queue algorithms. Technical Report. Oxford University, 1997.
7. Martinez Palau X, DominguezSal D, LarribaPey JL. Twoway replacement selection. Proceedings of the VLDB Endowment; September 2010; 3: pp. 871-881.
8. Bulut M. ReducedCBT and SuperCBT, two new and improved complete binary tree structures. Turk J Elec Eng&Comp Sci 2016; 24: 2150-2162
9. Kim JC, Chhugani NS, Sedlar E, Nguyen AD, Kaldewey T, Lee VW, Brandt SA, Dubey P. FAST: fast architecture sensitive tree search on modern CPUs and GPUs. In: 2010ACM SIGMOD/PODS Conference; 2010; Indianapolis, Indiana, USA.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

5M: Multi-Instance Multi-Cluster based Weakly Supervised MIL Model for Multimedia Data Mining

By Girisha GS & Dr. K. Udaya Kuma

BNM Institute of Technology

Abstract- The high pace rise in online as well as offline multimedia un annotated data and associated mining applications have demanded certain efficient mining algorithm. Multiple instance learning (MIL) has emerged as one of the most effective solutions for huge un annotated data mining. Still, it requires enhancement in instance selection to enable optimal mining and classification of huge multimedia data. Considering critical multimedia mining applications, such as medical data processing or content based information retrieval, the instance verification can be of great significance to optimize MIL. With this motivation, in this paper, Multi-Instance, Multi-Cluster based MIL scheme (MIMC-MIL) has been proposed to perform efficient multimedia data mining and classification with huge un annotated data with different features. The proposed system employs soft max approximation techniques with a novel loss factor and inter-instance distance based weight estimation scheme for instance probability substantiation in bags.

Keywords: multimedia data mining, multiple instance learning, multi-instance, multi -cluster based mining.

GJCST-C Classification : *H.2.4 H.2.8*

5MMULTIINSTANCEMULTICLUSTERBASEDWEAKLYSUPERVISEDMILMODELFORMULTIMEDIADATAMINING

Strictly as per the compliance and regulations of:

5M: Multi-Instance Multi-Cluster based Weakly Supervised MIL Model for Multimedia Data Mining

Girisha GS [α] & Dr. K. Udaya Kumar [σ]

Abstract- The high pace rise in online as well as offline multimedia un annotated data and associated mining applications have demanded certain efficient mining algorithm. Multiple instance learning (MIL) has emerged as one of the most effective solutions for huge un annotated data mining. Still, it requires enhancement in instance selection to enable optimal mining and classification of huge multimedia data. Considering critical multimedia mining applications, such as medical data processing or content based information retrieval, the instance verification can be of great significance to optimize MIL. With this motivation, in this paper, Multi-Instance, Multi-Cluster based MIL scheme (MIMC-MIL) has been proposed to perform efficient multimedia data mining and classification with huge un annotated data with different features. The proposed system employs soft max approximation techniques with a novel loss factor and inter-instance distance based weight estimation scheme for instance probability substantiation in bags. Unlike conventional clustering scheme, the proposed MIMC algorithm performs instance-level verification, class-level clustering and bag-level classification, simultaneously to perform mining with minimal possible complexity. The performance evaluation with SIVAL image datasets with 10 fold cross validation affirms that the proposed system performs better than existing clustering based approaches.

Keywords: multimedia data mining, multiple instance learning, multi-instance, multi-cluster based mining.

I. Introduction

The high pace emergence of information technologies and associated applications, the accumulation of data and its efficient mining and information retrieval has been increasing with a exponentially. Recently, Multimedia Data mining has emerged as one of the most sought technology. MDM can be stated as the process dealing with data processing based intended multimedia data or information retrieval. Multimedia data can be of various categories such as video, audio, image, animation, moving data sequences, etc. MDM exhibits various tasks such as prediction, or trend analysis based on association retrieval, clustering, and classification etc.

Author α: Research Scholar, Department of Information Science & Engineering, BNM Institute of Technology, Bangalore, India. e-mail: girisha_gs@yahoo.com,

Author σ: Principal, Adarsha Institute of Technology, Bengaluru, Karnataka, India. e-mail: udayakumarkrishnappa@gmail.com

The rising applications and utilities have motivated academia-industries to develop certain optimal technique for MDM.

Numerous approaches such as machine learning, artificial neural network, and association rule mining etc have been used for MDM. However; most of the existing approaches do fail to process large scale data sets. Moreover, it gets more complicate with the huge un annotated data. The emergence of MIL [1] has enabled better learning and classification efficiency than conventional supervised learning schemes. With the motivation to develop a robust and efficient MDM technique, in this paper an efficient MIL algorithm has been developed to classify un annotated multimedia data. In function, MIL classifies bags of instances, where bags represent the images and instances signify related features. In MIL, the labelling is performed on each bag and hence instance based labelling is not required. Such features significantly reduce the computational complexity and makes classification efficient.

MIL approach have exhibited appreciable effectiveness for major applications such as mining application, Classification [2], Vision based biomedical applications and His to pathological data analysis [1], Content Based Image Retrieval (CBIR) [3], Moving object detection [4], Image and Video processing [5][6], and numerous surveillance applications [7,8]. A number of MIL algorithms have been proposed such as APR [1], DD [9], EM-DD [10] that used a generative models to identify the concept region or the region of interest (ROI) by localizing all the true positive instances in the region space or feature space. In such schemes, the Single-Instance Learning (SIL) problems are generalized to the MIL problem. To achieve better performance recently few efforts were made that intend to explore the additional machine learning approach for classification. Some of these MIL algorithms are MI-SVM [2], MI-Kernel [11], MIO [12], Citation KNN [13] and MIL Boost-NOR [14]. Furthermore, MIL schemes such as DD-SVM [5], MILES [4], MILD B [15] and MILIS [16] have also used support vector machine (SVM) to perform classification. Considering significance of clustering scheme for MIL algorithm, in [17] a Multiple Instance Clustering Scheme was developed that primarily functions to learn the clusters formed by similar instances. However, this

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016

approach could consider only one cluster to perform classification and does not consider any negative bag during classification. In [18] multiple components were assessed to detect single object class. On contrary, in this paper we have developed multi-instance, multi-cluster based MIL model (MIMC-MIL) for MDM. In the proposed model, we have considered multiple instances in one cluster and multiple clusters in bag for effective classification accuracy. In [19, 20], few assumptions were incorporated to form multiple label MIL to perform multimedia data (image) classification. Our proposed MIMC-MIL model employs soft max approximation to estimate the probability of an instance in a bag to perform multimedia mining. The enhanced loss function and fair weight estimation based MIMC-MIL scheme has exhibited better performance than other existing systems. The remaining sections of the paper are presented as; Section II discusses the proposed MIMC-MIL algorithm and its implementation for ROI verification, clustering and classification. Section III presents results and analysis, which is then followed by conclusion in Section IV. References used in this paper are given at the last.

II. Our Contribution

In this paper, the general concept of bag and instance based weekly supervised MIL algorithm has been considered for multimedia mining. The generic functional definition of MIL states that even if a bag contains at least one positive instance, it can be labelled as positive bag. On contrary, the rise in highly critical data mining where accuracy plays significant role, such as medical data analysis and vision based decision process, such hypothesis often creates suspicion and question over functional accuracy and reliability. There are a number of multimedia mining applications where classification accuracy is of great significance and therefore to alleviate such ambiguity in conventional MIL approaches, the verification of the Region of Interest (ROI) also called concept region in bags can be vital. With this intention, in our previous work [25], we developed a single level clustering based ROI instance verification algorithm for multimedia data mining (MDM) and classification. In [25], the classification was done on cluster level. However, realizing the requirement of more precise and accurate mining performance, instance level analysis can be of great significance. The multiple instance based ROI verification and respective class formation (clustering in individual bag), followed by the multi-level clustering can ensure more effective and accurate mining performance. With this motivation, in this paper a highly robust and efficient Multi-Instance, Multi-Clustering based weakly supervised MIL learning model (MIMC-MIL) has been developed for MDM applications. Generally, a typical clustering based MDM encompasses three phases; segmentation, clustering and classification. These all process introduces huge computational complexity and computation time if executed individually to perform MDM. In case of huge un annotated data; such limitations turn out to be more severe. Hence, to alleviate such limitations, the proposed MIMC-MIL model performs these three processes simultaneously. The proposed mining model performs instance or pixel level segmentation, patch level clustering and bag label (image label) classification simultaneously that enables optimal mining performance for huge un annotated data. Unlike conventional Machine Learning and artificial Neural Network (ANN) algorithm, MIMC-MIL can perform segmentation and classification of multimedia data simultaneously to ensure optimal mining efficiency. The overall proposed model of MIMC based multimedia mining and classification is given in Fig. 1.

In this paper, numerous novelties such as an enhanced loss factor and weight estimation model based soft max approximation techniques has been developed which ensure optimal ROI probability estimation in bags and hence enable more efficient mining and classification accuracy. Here we have considered an assumption that based on certain ROI or concept region, the segmentation and classification can be done using MIL approach. The same concept has been used in our MIMC-MIL based MDM model. As depicted in Fig. 1, the multimedia data SIVAL with 180 positive and equally negative bags have been considered to evaluate the mining and classification efficiency. In this paper, the feature extracted values for the images are taken as input, which is then followed by clustering and ROI verification by our proposed MIMC-MIL model.

a) Multi-Instance, Multi-Cluster Based stance Verifi -cation Model for Multiple Instance Learning

Using multimedia benchmark data as, the MIL approach selects set of features as training data, which is also known as a bag. Mathematically bag can be defined as $\mathcal{X}_i = \{\mathcal{X}_{i1},\ldots,\mathcal{X}_{im}\}$ and for k cluster, the individual bag is associated with a label, which can be defined as $y_{ij}^{k} \in \mathcal{Y} = \{-1, 1\}$. In other words, the individual instance $(x_{ij} \in \mathcal{X})$ in a bag $(\mathcal{X}_{ij} \in \mathcal{X}^m)$ possesses a true label $(y_{ij}^{k} \in \mathcal{Y})$ as a hidden variable that remains unknown during feature mining and training for further classification.

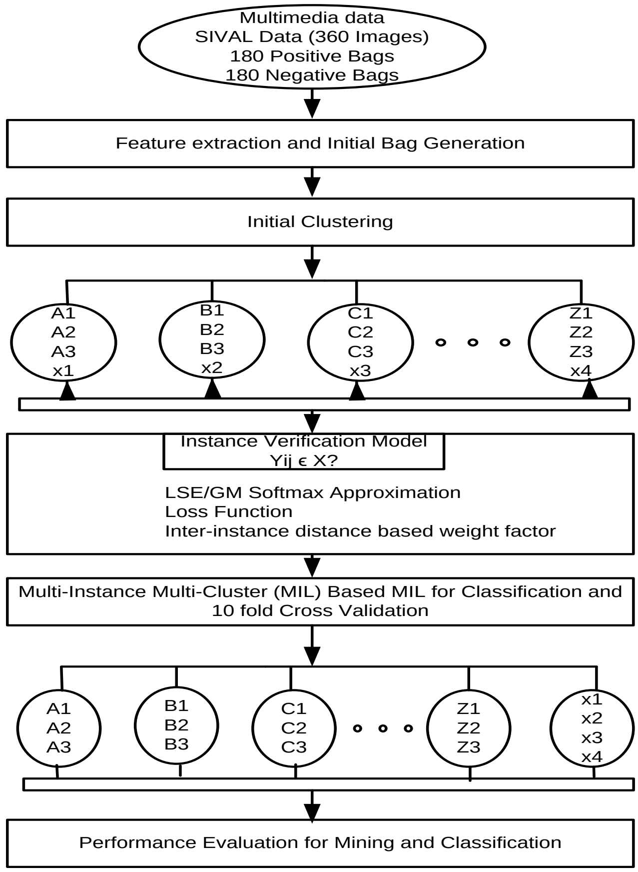


Fig.1 : Proposed MIMC based MIL model for Multimedia data mining

A bag is labelled as positive when x_{ij} belongs to the k^{th} cluster, i.e. $y_{ij}^{k} = 1$. As already stated, a bag can be labelled as positive if minimum one instance is positive and belongs to the k^{th} cluster. Mathematically,

$$y_i = \max_j \left(y_{ij}^{k}\right) \tag{1}$$

where max is similar to an OR operator because $y_{ij}^{k} \in \mathcal{Y}, max_j\left(y_{ij}^{k}\right) = 1 \Longleftrightarrow \exists_j$, provided $y_{ij} = 1$.

In general, the predominant objective of an MIL algorithm is to perform learning at instance-level classifier $h(x_{ij}) : \mathcal{X} \rightarrow \mathcal{Y}$. MIL intends to provide an efficient learning mode for splitting the positive instances into $\mathcal{K}$ clusters by means of $\mathcal{K}$ instance level classifiers $h(x_{ij}) : \mathcal{X} \rightarrow \mathcal{Y}$. In this process, the labelled bags y_i are used in such a manner that $max_j \max_k h(x_{ij}) = y_i$. Unlike conventional MIL approaches [21, 22, 23], we have introduced a loss function to estimate the optimal weak classifier response $h_t^{k}: \mathcal{X} \rightarrow \mathcal{Y}$ that significantly reduces the loss on training data. Mathematically, the loss function is given by:

$$\mathcal{L}_{\mathcal{A}}(h) = -\sum_{i=1}^{n} w_i \left(1(y_i = 1)\log p_i + \right.$$
$1 y j = -1 \log 1 - p i$, and

$$\mathcal{L}_{\mathcal{B}}(h) = \sum_{i=1}^{n} w_i \sum_{(j,m)\in\mathcal{E}_i} v_{jm} \left\|p_{ij} - p_{im}\right\|^2 \tag{2}$$

where w_i represents the initial weight of the i^{th} training data and $1(\cdot)$ states certain index function. The variable $\mathcal{E}_i$ represents the group of the pairs of all the neighbouring instances in i^{th} bag or training data. Here, v_{jm} represents the weight on the patches, which is nothing else but the pair of instances (features). The variables d_{jm} represents the relative distance between j and m. If the instances are closer, then they are assigned with higher weights. To estimate the respective weights (v_{jm}) of the instances and patches, we have used $v_{jm} = exp(-d_{jm})$. Thus, estimating the value of v_{jm}, the cumulative loss function (CLF) (2) has been estimated using equation (3).

$$CLF = \mathcal{L}(h) = \mathcal{L}_{\mathcal{A}}(h) + \lambda \mathcal{L}_{\mathcal{B}}(h) \tag{3}$$

Here, $\mathcal{L}_{\mathcal{B}}(h)$ plays significant role to eliminate the ambiguity during training by imposing an efficient contextual constraint over the instances and thus enabling neighbouring images (patches formed by instances) to share analogous classes. The other loss function, $\mathcal{L}_{\mathcal{A}}(h)$ states the typical negative log likelihood. Variable λ represents the weight associated with the supplementary item that signifies the importance of the inter-relationship between the neighbouring instances (instance represents unit features of the image). Thus, the proposed mining and classification system can be considered as resilient of noise as well as robust for effective segmentation purposes. In our proposed model, the training of h_t^k has been performed by reducing error associated with the training data, which is estimated by weight factor w_{ij}^k

$$\left| w_{ij}^k | : h_t^k = arg\min_{h} \sum_{i,j} \left(h(w_{ij}^k) \neq y_i{}^k\right) |w_{ij}^k \right| \qquad (4)$$

where, $w_{ij}^k \equiv -\frac{\partial \mathcal{L}(h)}{\partial h_{ij}^k}$.

Here, a soft max function $g_\ell(v_\ell)$ has been considered that performs approximations of max value over $v = \{v_1, .., v_m\}$. There are a number of approximation approaches, such as noisy-OR (NOR), generalized mean (GM), log-sum-exponential (LSE), and integrated segmentation and recognition (ISR). Unlike our previous work [25], where NOR model was used, in this paper we have applied GM and LSE approximation techniques individually to perform approximation over $v = \{v_1, .., v_m\}$. In addition, a factor named sharpness control factor (SCF), r has been introduced to enhance the classification efficiency by means of controlling the sharpness during approximation for instance probability estimation. The mathematical presentation of the soft max approximation of GM and LSE models are given in Table I.

Table 1 : Soft max approximation models

Model	$g_\ell(v_\ell)$	$\frac{\partial g_\ell(v_\ell)}{\partial v_i}$	Domain
GM	$\left(\frac{1}{m}\sum_l v_\ell^r\right)^{\frac{1}{r}}$	$g_\ell(v_\ell) = \frac{v_i^{r-1}}{\sum_\ell}$	[0, ∞]
LSE	$\frac{1}{r} ln \frac{1}{m} \sum_{exp} (r v_\ell)$	$\frac{\exp(r v_i)}{\sum_\ell \exp(r)}$	[-∞,∞]

Since $r \to \infty$, soft max approximations can be observed as $g(_\ell v_\ell) \approx max(_\ell v_\ell) g(_\ell v_\ell) \to v$.* Thus, for m variables($v = \{v_1, \dots, v_m\}$), the respective softmax function $g(_\ell v_\ell)$ can be obtained by

$$g_\ell(v_\ell) \approx \max_\ell (v_\ell) = v, \frac{\partial g_\ell(v_\ell)}{\partial v_i} \approx \frac{1(v_i = v^*)}{\sum_l 1(v_i = v^*)} \qquad (5)$$

where, $m = |v|$. To maintain simplified presentation, in rest of the paper, the variable $g(_\ell v_\ell)$ has been represented by g, while v_ℓ is represented in terms of ℓ. In order to enhance the loss function $\mathcal{L}$, at first the probability p_i of bag is required to be estimated, which is stated to be the highest over p_{ij}^k. Here, the probability that an instance x_{ij} belongs to the k^{th} cluster, is given by

$$p_{ij}^k = \sigma(2h_{ij}^k) \qquad (6)$$

where $h_{ij}^k = h^k(x_{ij})$.

Now, substituting max with g, the instance probability p_i in a class can be obtained as

$$p_i = g_j\left(g_k(p_{ij}^k)\right) = g_{jk}(p_{ij}^k) = g_{jk}\left(\sigma(2h_{ij}^k)\right) \qquad (7)$$

where

$$\sigma(v) = \frac{1}{1 + \exp(-v)}$$

The optimal weighted error factor (w_{ij}) and the derivative $\frac{\partial \mathcal{L}}{\partial h_{ij}^k}$ can be obtained as

$$w_{ij}^k = \frac{\partial \mathcal{L}(h)}{\partial h_{ij}^k} = -\frac{\partial \mathcal{L}(h)}{\partial p_i}\frac{\partial p_i}{\partial p_{ij}^k}\frac{\partial p_{ij}^k}{\partial h_{ij}^k}. \qquad (8)$$

Thus, performing the optimization of weighed error factor $|w_{ij}^k|$, the weak classifier h_{ij}^k has been trained efficiently. Finally, a string classifier has been obtained as

$$h^k \leftarrow h^k + \alpha_t^k h_t^k \qquad (9)$$

where α_t assess weighing of the relative significance of the weak learner. Thus, implementing our proposed MIMC-MIL, the instance verification in each bag can be done and respective accurate clustering based classification can be performed.

b) Mimc-Mil Based Multimedia Mining

Multimedia data can be of different types and in huge quantity. The conventional systems suffer from extraction or classification, particularly with huge un annotated data. In addition to the annotation issues, unclear type and nature of multimedia data requires efficient approaches for mining. In a number of MDM systems, clustering has been used for mining and classification. The existing cluster based approaches do apply single level of clustering to perform classification, but considering critical applications, where the misplacement of a single instance or feature can alter the prediction and further decision process, the conventional clustering based mining schemes requires multilevel instance verification. In other words, the probability estimation of an instance of multimedia data in certain class can enable better clustering accuracy and hence can enable enhanced mining classification. In this paper, the MIMC based MIL scheme has been applied for mining and classification, where each instance and its probability of belongingness to certain class or cluster has been done. In general, most of the existing MDM techniques use three different approaches; segmentation, clustering, and classification. The execution of these all approaches with the huge data, turns out to be highly complicate and time consuming. Therefore, to deal with such limitation, we have used the proposed MIMC-MIL scheme that performs clustering, segmentation and classification simultaneously.

In this paper, to perform multimedia mining and classification a benchmark multimedia data containing huge images with different features has been considered from which the training data ($\mathcal{X}_i = \mathcal{X}i1, \ldots, \mathcal{X}im$ has been prepared and respective labelling of bags ($y_i \in \mathcal{Y} = \{-1,1\}$) has been done. Performing the initial clustering and bag formation from benchmark data the proposed MIMC algorithm has been applied as presented in Fig. 1. Table II represents the training data (input) and learning objective definition.

Table 2 : Training data and its objective formulations

TECHNIQUE	TRAINING DATA	OBJECTIVE $x_i \rightarrow Classification$ $x_{ij} \rightarrow Segmentation$ $y_{ij}^k \rightarrow Clustering$
Standard Classifier	x_i	$x_i \rightarrow \{-1,1\}$
Conventional MIL	$x_i = \{x_{i1}, \ldots, x_{im}\}$ $x_{ij} \in x$	$x_i \rightarrow \{-1,1\}$; $x_{ij} \rightarrow \{-1,1\}$
Proposed MIL	$x_i = \{x_{i1}, \ldots, x_{im}\}$ $x_{ij} \in x$	$x_i \rightarrow \{-1,1\}$; $x_{ij} \rightarrow \{-1,1\}$ $y_{ij}^k \rightarrow \{y_{ij}^1, \ldots, y_{ij}^k\}$;

Table II depicts that the proposed MIMC-MIL scheme is capable of performing patch level clustering($x_{ij} \rightarrow \{y_{ij}^1, \ldots, y_{ij}^k\}$; $y_{ij}^k \in \{-1,1\}$), segmentation$\left(x_{ij} \rightarrow \{-1,1\}\right)$ at pixel-level, and classification at bag or image level($x_i \rightarrow \{-1,1\}$). To perform MDM at first feature vectors have been prepared from benchmark data which has been fed as the input of MIMC-MIL algorithm where the learning for multilevel ($\mathcal{K}$ instance-level) classification has been done $h^k(x_{ij}): \mathcal{X} \rightarrow \mathcal{Y}$ for $\mathcal{K}$clusters. Consequently, the bag-level classifier for certain k^{th} cluster has been formed as$h^k(x_i) : \mathcal{X}^m \rightarrow \mathcal{Y}$. Thus, the overall classification approach for MDM can be stated as$\mathcal{H}(x_i) : \mathcal{X}^m \rightarrow \mathcal{Y}$.

$$\mathcal{H}(x_i) = \max_k h^k(x_i) \max_k \max_j h^k(x_{ij}) \qquad (10)$$

As an optimization of our previous work [25], in this paper the ROI probability factor p_i has been estimated in terms of the softmax of $p_{ij} \equiv p(y_{ij} = 1|y_{ij})$ for all the associated instances in the bags (image dataset). ROI instance probability (p_{ij}) in a bag (bag represents the image having multiple clusters, where clusters are formed by instances) has been estimated $\left(p_{ij}^k = p(y_{ij}^k = 1|x_{ij})\right)$ using LSE and GM based soft max approximation technique. The eventual instance probability is obtained as:

$$p_i = g_j(p_{ij}) = g(g_k(p_{ij}^k)) \qquad (11)$$

where p_{ij}^k represents the probability that the ROI or the concept region instance x_{ij} belongs to the k^{th} cluster.

The overall MIMC-MIL based mining and classification model is given in Fig. 1.

Input: Multimedia data extracted features or Bags $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}, \{y_1, \ldots, y_n\}, \mathcal{K}$ cluster, $\mathcal{T}$ Threshold
Output: $h^1, \ldots, h^k$
for $t = 1 \rightarrow \mathcal{T}$ do
for $k = 1 \rightarrow \mathcal{K}$ do
Calculate $\mathcal{L}_{\mathcal{A}}(h)$ and $\mathcal{L}_{\mathcal{B}}(h)$
Calculate weights $w_{ij}^k = \mathcal{L}(h) = \mathcal{L}_{\mathcal{A}}(h) + \lambda\mathcal{L}_{\mathcal{B}}(h)$

$$-\frac{\partial \mathcal{L}(h)}{\partial h_{ij}^k} = -\frac{\partial \mathcal{L}_{\mathcal{A}}(h)}{\partial p_i}\frac{\partial p_i}{\partial p_{ij}^k}\frac{\partial p_{ij}^k}{\partial h_{ij}^k} + \lambda\frac{\partial \mathcal{L}_{\mathcal{B}}(h)}{\partial p_i}\frac{\partial p_{ij}^k}{\partial h_{ij}^k}$$

Perform training of the weak classifier h_t^k using weights w_{ij}^k

$$h_t^k = arg \min_h \sum_{ij} 1\left(h(x_{ij}^k) \neq y_i^k\right)\left|w_{ij}^k\right|$$

Calculate α_t by means of the line search so as to reduce CLF $\mathcal{L}\left(.,h^k + \alpha\, h_t^k,.\right)$
Update strong classifier $h^k \leftarrow h^k + \alpha_t^k\, h_t^k$
Form final cluster with ROI/ verified instances
end for
end for

Fig. 2 : Algorithm for proposed MIMC based mining and classification

A brief of the three significance functional phases of MIMC-MIL is given as follows:

i. *Classification*

In this paper, initially the image level classification has been done that exploits the developed instance verification and clustering approach. Here, the overall features or instances x_{ij} of complete image data have been used to perform training as per [22]. The training approach uses our developed Multi-Instance Multi-Cluster (MIMC) instance features or instance-level labels retrieved from the labels prepared on bag-level ($y_{ij} = y_i$, $i = 1, \ldots, n, j = 1, \ldots, m$) and thus based on the final clustering output the classification has been done.

ii. *Segmentation*

In multimedia mining applications, especially when there are huge data, it becomes too intricate, ambiguous and computationally complex to perform annotations for all the data (image). The proposed MIMC-MIL scheme doesn't demands huge annotation or even any instance-level supervision. The proposed algorithm selects few ROI data, also called concept data randomly along with some other non-ROI data to form a training subset. Our proposed algorithm generates probability mapping for all instances (p_i) associated with bag X_i. Thus, implementing MIMC-MIL classifier, the parameters such as accuracy, recall and F-measures have been estimated. F-measure factor $2.\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ can be used for segmentation.

iii. *Clustering*

As discussed in previous sections, the proposed MIMC-MIL approach performs clustering while performing instance verification or ROI classification for mining. Furthermore, the proposed system performs pixel level segmentation that can be further inter-related with patch level (collection of the instances having similar dimensions and features) clustering. The standard boosting has been applied to perform instance level segmentation, which can then be followed by K-means algorithm to perform clustering of the positive instances (concept region or ROI).

III. Results and Discussion

With an objective to perform multimedia data mining, in this paper a robust and enhanced clustering based multi-instance multi-cluster MIL (MIMC-MIL) scheme has been developed. The overall proposed model has been developed using MATLAB 2014b software tool. To evaluate the performance SIVAL dataset has been used. The considered datasets encompasses 360 bags containing 180 bags each for positive and negative type. The images in SIVAL dataset are presented in Table III. To evaluate the performance of the proposed system, the 10-fold cross validation has been done and performance evaluation has been done in terms of classification accuracy and area under ROC (AUC) curve. As already stated, in the proposed algorithm, two distinct soft max approximation algorithms have been used and hence the proposed algorithm has been evaluated with the both generalized mean (GM) and log-sum-exponential (LSE) algorithm. The results obtained for accuracy and AUC are given in the following figures.

Table 3 : Images in SIVAL dataset

SIVAL IMAGE DATA	
Positive Bag	Negative Bag
Smiley face doll Blues crunge Green tea box	Checker edscarf Dirty running shoe Felt flower rug

In this paper the multimedia data mining and classification has been performed with ROI verification and clustering by means of GM and LSE soft max approximation techniques individually. Fig. 3 and Fig. 4 represent the mining and classification accuracy using proposed MICL-MIL algorithm with log-sum-exponential (LSE) and generalized model (GA) soft max approximation techniques respectively. Here, it can be observed that LSE model performs better with our proposed MIMC algorithm. Interestingly, LSE model with our proposed MIMC algorithm performs better than with conventional boosting based MIL scheme.

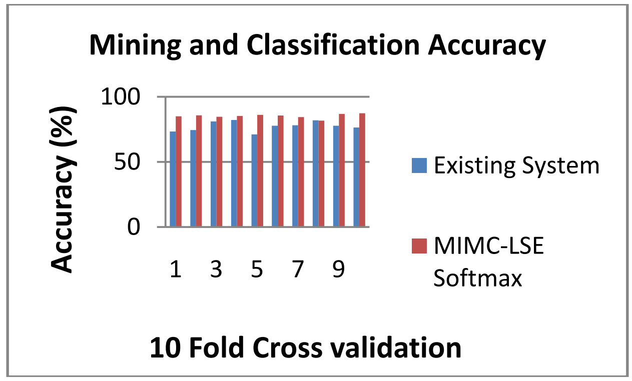


Fig. 3 : Mining and classification accuracy using LSE Soft max approximation

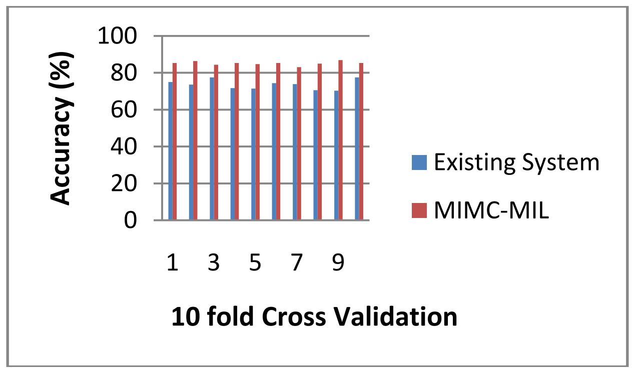


Fig. 4 : Mining and classification accuracy using GM Soft max approximation

Fig. 5 and Fig. 6 affirms that LSE soft max performs better with the proposed MIMC based MIL for multimedia data mining.

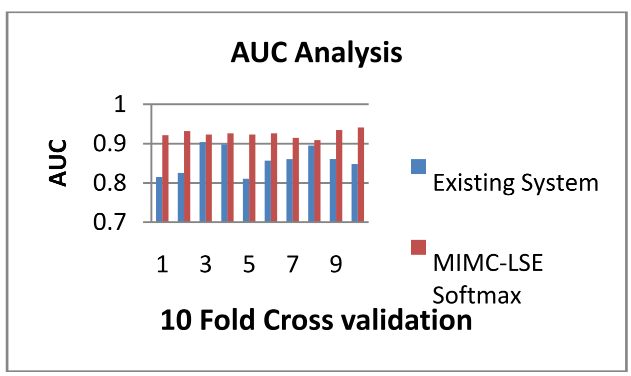


Fig. 5 : AUC analyses using LSE Soft max approximation

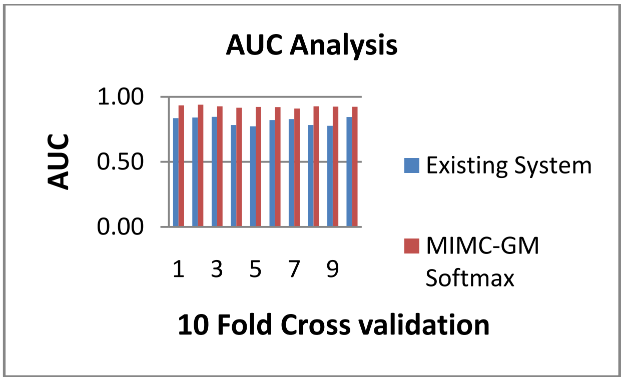


Fig. 6 : AUC analyses using GM Soft max approximation

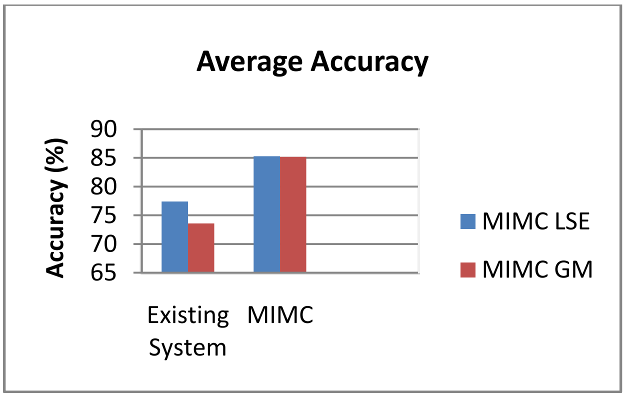


Fig. 7 : Comparative average mining and classification accuracy using GM and LSE Soft max

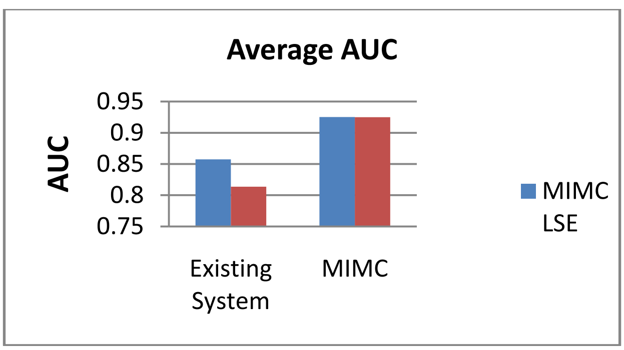


Fig. 8 : Comparative average AUC analysis using GM and LSE Soft max

The average performance analysis (Fig. 7 and Fig. 8) affirms that the proposed MIMC-MIL performs better with log-sum-exponential (LSE) soft max approximation than generalized model (GM) based approximation for ROI instance probability estimation. Overall performance exhibits that the proposed multi-instance multi-cluster (MIMC) algorithm with LSE soft max approximation for MIL can provide a novel solution for large scale multimedia data mining (MDM).

Table 4 : Comparative classification accuracy analysis

Mil Based MiningTechniques	Accuracy(%)
DD-SVM [5]	85.4
MILIS [16]	85.8
MIForest [26]	88.6
mi-SVM [27]	85.0
EM-DD [28]	87.4
MILES [29]	84.8
MILD [15]	83.3
Intra Clustering_DMIL [25]	84.2
Proposed MIMC-MIL	87.5

As depicted in Table IV, the proposed system exhibits better mining and resulting classification accuracy as compared to the other existing systems. The developed system with different benchmark data exhibits the MIMC-MIL based approach outperforms conventional MIL based boosting and hence affirms that our proposed MIMC-MIL scheme can significantly perform with huge un annotated data for multimedia mining applications. Literatures state that other algorithms such as MKL [24] usually takes several days of time to train a classifier even for 60 images, while our proposed system performs optimized classification of 360 images just within 20 minutes.

IV. Conclusion

The exponential rise in un annotated multimedia data has demanded researchers to develop certain efficient multimedia data mining (MDM) algorithm that can provide optimal mining performance with minimal complexity and computational overheads. With these motivations, in this paper a robust multi-instance, multi-cluster (MIMC) multiple instances learning (MIL) algorithm has been developed. With an intension to assure optimal mining and classification efficiency a robust region of interest (ROI) identification and verification model has been developed. To perform ROI verification, two soft max approximation techniques, generalized mean (GM) and log-sum-exponential (LSE) algorithm have been applied. These approximation models have been used to estimate the probability of an instance, whether it belongs to a bag or not. In addition, a weight factor has been introduced that signifies inter-relationship between neighbouring instances. It enables effective clustering, segmentation as well as classification. Interestingly, the proposed system justifies its robustness by segmentation, clustering and classification simultaneously. The performance evaluation with multimedia image datasets with 10 fold cross validation affirms that the proposed system performs better than existing clustering based approaches. Thus, the proposed mining model and classification system can be considered to be resilient to noise as well as more robust in terms of more effective segmentation and classification. The overall performance affirms that the proposed system can be effective to perform mining and classification for different multimedia data types.

References Références Referencias

1. Dietterich, T.G., Lathrop, R.H., Lozano-P´erez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. Vol. 89 (1-2), (1997) 31–71
2. ndrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS. (2003) 1073–1080
3. Zhang, Q., Goldman, S.A., Yu, W., Fritts, J.: Content-based image retrieval using multiple instance learning. In: ICML. (2002) 682–689
4. Chen, Y., Bi, J., Wang, J.Z.: MILES: Multiple-instance learning via embedded instance selection. IEEE Trans. Vol. 28 (12) Pattern Anal. Mach. Intell. (2006), 1931–1947
5. Chen, Y., Wang, J.Z.: Image categorization by learning and reasoning with regions. J. Mach. Learn. Res. Vol.5, (2004) 913–939
6. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines (2001), software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm
7. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: CVPR. (2009) 983–990
8. Leistner, C., Saffari, A., Bischof, H.: MI Forests: Multiple-instance learning with randomized trees. In: ECCV. (2010) 29–42
9. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of key points. In: ECCV Int. Workshop Stat. Learning in Comp. Vis. (2004)
10. Zhang, Q., Goldman, S.A.: EM-DD: An improved multi-instance learning technique. In: NIPS. (2002) 561–568
11. G¨artner, T., Flach, P.A., Kowalczyk, A., Smola, A.J.: Multi-instance kernels. In: ICML. (2002) 179–186
12. Li, M., Kwok, J., Lu, B.L.: Online multiple instance learning with no regret. In: CVPR. (2010) 1395 –1401
13. Wang, J., Zucker, J-D.: Solving multiple-instance problem: A lazy learning approach. In: ICML (2000)
14. Viola, P., Platt, J.C., Zhang, C.: Multiple instance boosting for object detection. In: NIPS. (2006) 1419–1426
15. Li, W.J., Yeung, D.Y.: MILD: Multiple-instance learning via disambiguation. IEEE Trans. on Knowl. and Data Eng. (2010) Vol.22, 76–89
16. Fu, Z., Robles-Kelly, A., Zhou, J.: MILIS: Multiple instance learning with instance selection. IEEE Trans. Pattern Anal. Mach. Intell (2010).
17. D. Zhang, F.Wang, L. Si, and T. Li. M3IC: maximum margin multiple instance clustering. In IJCAI, 2009.
18. P. Doll´ar, B. Babenko, S. Belongie, P. Perona, and Z. Tu. Multiple component learning for object detection. In ECCV, 2008.
19. Z.-H. Zhou and M.-L. Zhang. Multi-instance multilabel learning with application to scene classification. In NIPS, 2007.
20. Z.-J. Zha, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi instance learning for image classification. In CVPR, 2008.
21. P. A. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. In NIPS, 2005.
22. L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithms as gradient descent. In NIPS, 2000.
23. B. Babenko, P. Doll´ar, Z. Tu, and S. Belongie. Simultaneous learning and alignment: Multi-instance and multi-pose learning. In ECCV workshop on Faces in Real-Life Images, 2008.
24. A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In ICCV, 2009.
25. G. S., Girisha, K. Udaya Kumar. An Enhanced Semi-Supervised Multiple Instance Learning Scheme for Multimedia Data Mining. International Journal of Applied Engineering Research. Vol. 10 No.86 (2015). 348-354.
26. Kelly, D., McDonald, J., Markham, C.: Weakly Supervised Training of a Sign Language Recognition System Using Multiple Instance Learning Density Matrices. Systems, Man, and Cybernetics, Part B: Cybernetics, vol.41, no.2, IEEE Transactions on April (2011) 526-541
27. Cheng, H. L., Gondra, I.: A Novel Neural Network-Based Approach for Multiple Instance Learning. Computer and Information Technology (CIT), IEEE 10th International Conference on (2010) 451-456
28. Yuan, X., Wang, M., Song Y.: Concept-dependent image annotation via existence-based multiple-instance learning. Systems, Man and Cybernetics, IEEE International Conference on (2009) 4112-4117
29. Dijia W., Jinbo B., Boyer, K.: A min-max framework of cascaded classifier with multiple instance learning for computer aided diagnosis," Computer Vision and Pattern Recognition, IEEE Conference on (2009) 1359-1366.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Isotropic Dynamic Hierarchical Clustering

By Victor Sadikov & Oliver Rutishauser

Abstract- We face a business need of discovering a pattern in locations of a great number of points in a high-dimensional space. We assume that there should be a certain structure, so that in some locations the points are close while in other locations the points are more dispersed. Our goal is to group the close points together. The process of grouping close objects is known under the name of clustering.

Keywords: clustering; hierarchical clustering; dynamic clustering; isotropic clustering; multi-dimensional space; b-tree; factor analysis.

GJCST-C Classification : H.3.3, I.5.3

Isotropic Dynamic Hierarchical Clustering

Victor Sadikov [α] & Oliver Rutishauser [σ]

Abstract- We face a business need of discovering a pattern in locations of a great number of points in a high-dimensional space. We assume that there should be a certain structure, so that in some locations the points are close while in other locations the points are more dispersed. Our goal is to group the close points together. The process of grouping close objects is known under the name of clustering.

1. We are particularly interested in a hierarchical structure. A plain structure may reduce the number of objects, but the data are still difficult to manage or present.
2. The classical technique suited for the task at hand is a B-Tree. The key properties of the B-Tree are that it is hierarchical and balanced, and it can be dynamically constructed from the input data. In these terms, B-Tree has certain advantages over other clustering algorithms, where the number of clusters needs to be defined *a priori*. The BTree approach allows to hope that the structure of input data will be well determine without any supervised learning.
3. The space is Euclidean and isotropic. This is the most challenging part of the project, because currently there are no B-Tree implementations processing indices in a symmetrical and isotropical way. Some known implementations are based on constructing compound asymmetrical indices from point coordinates, where the main index works as a key, while the function of other (999!) indices is lost; and the other known implementations split the nodes along the coordinate hyper-planes, sacrificing the isotropy of the original space. In the latter case the clusters become coordinate parallelepipeds, which is a rather artificial and unnecessary assumption. Our implementation of a B Tree for a high-dimensional space is based directly on concepts of factor analysis.
4. We need to process a great deal of data, something like tens of millions of points in a thousand-dimensional space. The application has to be scalable, even though, technically, out task is not considered a true Big Data problem. We use dispersed data structures, and optimized algorithms. Ideally, a cluster should be an ellipsoid in a high-dimensional space, but such implementation would require to store O(n2) ellipse axes, which is impractical. So, we are using multi-dimensional balls defined by the centers and radii. On the other hand, calculation of statistical values like the mean and the average deviation, can be done in an incremental way. This mean that when adding a point to a tree, the statistical values for nodes of all levels may be recalculated in O(1) time. The node statistical values are used to split the overloaded nodes in an optimal way. We support both, brute force O(2n) and greedy O(n2) split algorithms. Statistical and aggregated node information also allows to manipulate (to search, to delete) aggregated sets of closely located points.
5. Hierarchical information retrieval. When searching, the user is provided with the highest appropriate nodes in the tree hierarchy, with the most important clusters emerging in the hierarchy automatically. Then, if interested, the user may navigate down the tree to more specific points. The system is implemented as a library of Java classes representing Points in multi-dimensional space, Sets of points with aggregated statistical information (mean, standard deviation,) B-tree, and Nodes with a support of serialization and storage in a MySQL data base.

Author α: 200 S Laurel Av. A5-2D20 Middletown, NJ 07748 (732) 420-7453. e-mail: vic@att.com
Author σ: Atlantic Style P.O.Box 9 Oakhurst, NJ 07755 (732) 455-2081 e-mail: rutishauser@yandex.com

CCS Concepts

- Theory of computation→Theory and algorithms for applicationdomains→Machinelearningtheory→Unsupervised learningand clustering • Mathematics of computing→Mathematical software → Statistical software •Information systems→Information retrieval →Retrieval tasks and goals→Clusteringand classification

Keywords: clustering; hierarchical clustering; dynamic clustering; isotropic clustering; multi-dimensional space; b-tree; factor analysis.

I. Points, Implementation

In a high-dimensional space we assume that a considerable number of coordinates will contain zero values. To optimize the memory and storage space we would like to keep non-zero coordinates only. Thus, a Point object contains 3 fields: the number of non-zero coordinates, the array of sorted coordinate indices, and the array of corresponding coordinate values.

class Point

The Point class provides methods for calculating the Euclidean length of the point vector, getting a particular coordinate (or zero,) adding another point to the given point, calculating a few useful functions like distance to another point, dot product, and serializing to a base-64 string. Some functions, e.g. adding a point, may change the number of non-zero coordinates. The main loop for adding two points looks like the following.

```
int Ix = 0; int Ip = 0; int Iz = 0;
for(; Ix != this.N && Ip != p.N ;)
{
if(this.key[Ix] == p.key[Ip])
{
z.key[Iz] = this.key[Ix];
z.val[Iz] = this.val[Ix] + p.val[Ip];
Ix++; Ip++; Iz++;
}
else if(this.key[Ix] < p.key[Ip])
{
z.key[Iz] = this.key[Ix];
z.val[Iz] = this.val[Ix];
Ix++; Iz++;
}
else
{
z.key[Iz] = p.key[Ip];
z.val[Iz] = p.val[Ip];
Ip++; Iz++;
}
}
```

II. Sets Mathematics

The next step of our approach is the introduction of Sets of Points. The Sets allow calculation of aggregated statistical values. The most important value is the number of points (N) in the Set. It needs to be corrected each time a new point is added to the Set. The obvious way of calculating the new number of points is to increment the current number by one.

$$N = N + 1;$$

Other important statistical values of the set of points are arithmetic mean and the standard deviation. These values should also be adjusted every time a point is added to the Set. We could recalculate the arithmetical mean (**M**) from scratch, but we would like to follow the incremental approach and move it towards the newly added point (**P**) by the 1/N of the distance.

$$\mathbf{M} = \mathbf{M} + (\mathbf{P} - \mathbf{M}) / N;$$

As for standard deviation, at first sight, it seems to be a value that requires the full recalculation. Fortunately, this is not the case. We can store and adjust the standard deviation in an incremental way too, based on the following formula.

$$E[X – E(X)]2 = E[X2] – (E[X]) 2$$

This means that to calculate the standard deviation it is enough to store the sum of the squares of point coordinates (S,) which can be adjusted incrementally.

$$S = S + |\mathbf{P}|2;$$

And when we need to calculate the standard deviation we will do the following.

$$D = sqrt(S/N - |\mathbf{M}|2);$$

III. Clustering Example

Clustering basically means grouping similar objects together. If the objects have a number of numerical attributes they may be represented as points in a multidimensional space. The clustering will mean to partition the whole set of points into a number of disjoint sub-sets.

Let's consider an example in a one-dimensional space. The example is free of the challenges related to multidimensional clustering and is easy to comprehend. Assume we are given the set of five numbers {0.4.5. 9.13.}

Figure 1

We can see that initially the points occupy the segment [0,13.] Assume we need to split the set of the given points on the number line into two **clusters** then the resulting clusters will be sub-segments. Even in this simple example two different splits are possible. The first splits makes egments [0, 5] and [9, 13] while the other makes segments [0, 4] and [5, 13] (see Figure 1.)

In terms of segments, the first split looks better, because it finds two compact segments rich of points, while the second split covers almost the whole initial segment. In terms of statistical variables, the first split is better too, because the sum of the deviations of the resulting sets is minimal.

IV. Dynamic Clustering, Approach

If we continue adding new points, then we need to decide which sub-segment each new point belongs to. Points below 5 we can add to the first sub-segment, and points above 9 we add to the second sub-segment. We may place points between 5 and 9 into either sub segment. Our criterion here is to avoid big segments, or (which is the same) to keep the sum of deviations minimal. But we always need to update the boundaries of the sub segments so that we can exactly know in which segment a particular point is to be found. If the points are well spread, knowing exact boundaries of the segments may also improve unsuccessful search.

After adding a certain number of points to a sub-segment, we will need to split this sub-segment into two sub segments with a smaller number of points. This can be done exactly in the same manner as we split the initial segment. After that we will be adding new points to the set of three sub-segments. Then we will need to split another sub-segment. And the number of sub-segments will increase again.

Unlike the static approach where the set of all objects exists before the procedure of clustering starts; dynamic approach assumes that clusters are incrementally adjusted each time a new object gets added into the set. This eliminates the dedicated step of clustering for the price of a longer time needed to include objects. Dynamic clustering provides better flexibility and can be performed with less *a priori* known information about the data, e.g. when the total number of target clusters is unknown.

V. Hierarchical Clustering, Approach

The bigger the number of points in our data set, the bigger the number segments. Soon it gets big enough, and we may need to introduce the next level of hierarchy, when smaller clusters are, in their turn, grouped into clusters of the higher level. The approach of adding points, splitting segments, and adding new levels when necessary is quite similar to adding objects to a B-tree [1.] The tree starts with the root node, responsible for all points stored in the tree. At each level, the parent node consist of a number of sub-nodes. The points in each sub-node are close one to another.

A classical one-dimensional B-tree design focuses on minimization of the information stored at the node level. Namely, a parent node stores a number of adjacent values in ascending order, with the sub-nodes being placed between adjacent values, plus one at the beginning and one at each end. Thus we know that the elements of each subtree are greater than the left adjacent value and less than the right one.

On the contrary, our design prefers to store excessive boundary and statistical information about the sub-nodes. In the case of one-dimensional space, each node occupies a segment and sub-nodes of one parent do not intersect. The boundary segment can be defined by its two ends, or by the middle point (**C**) and the distance (R) to the ends. The letter way will occupy less memory in a general multidimensional case. We also keep the number of points in each sub-tree, their arithmetic mean and standard deviation.

As we stated above, statistical values facilitate splitting the nodes in the optimal way, while the boundary information allows us in some cases to terminate an unsuccessful search
heuristically.

VI. Isotropic b-tree, Approach

B-trees do their work great, as long as the attributes of the objects are one-dimensional. Unfortunately, there is some problem with direct extension of B-tree to a multidimensional case. The two common approaches are the following.

The mix approach assumes composing the compound index based on the component indices, and then making use of a one-dimensional B-tree. The problem with this approach is that the component indices are treated by far not equally. One index plays the main role, while the role of the others is insignificant.

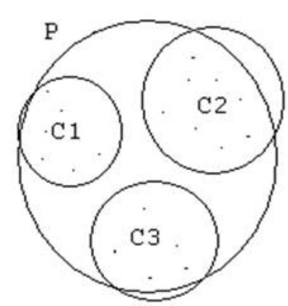
The other approach is more complicated. R-tree [2] is a variant of B-tree where the nodes are bound by coordinate rectangles. This approach is more symmetrical in terms of using indices. But the directions along the coordinate axes are still different from arbitrary ones.

We would like to build a variant of B-tree, where the nodes are bound with circles or ellipses. This decision ensures that the essential property of isotropy of the physical space is not ignored.

VII. (Isotropic Dynamic Hierarchical Clustering, Two-Dimensional Case)

The circles are defined by the center point (**C**) and the radius (R.) It the 2-dimensional space, the center is defined by the two coordinates, and so we need to store 3 real values. Alternatively, if we decided to present clusters as ellipses in general orientation, we will need to store the semi-principal axes and the angles, which would require O(n^2) memory, with n being the number of dimensions in the space.

In the picture to the right, the circles corresponding to the sub-nodes of one and the same parent do not intersect. The biggest circle (P) corresponds to the parent node, while C1, C2, and C3 correspond to the sub-nodes. Keeping on adding new points to the tree we cannot avoid the situation where the sub-nodes start to intersect. We will dwell on the intersecting areas later in this paper.



a) *Selecting a Sub-Node*

As mentioned in section 4, while adding a new point outside any bounding circle, we need to select the most appropriate sub-node. To do this, we will try to add the new point to each sub-node, and calculate the new bounding radii. Then we will select a sub-node so that the sum of the new squared radii should be minimal, because our goal is to make eventually all sub-node circles of about the same absolute size in the given space.

Let's assume that the old radius was R_i, then the new radius will be $R_i+H_i/2$, where Hi is the distance from the new point to the circle C_i. If we select the i-th sub-node, the sum of new squared radii will grow by $(R_i+H_i/2)2 – R_i^2$, i.e. by $R_iH_i+H_i^2/4$, so we need to select the sub-node where $T_i= 4R_iH_i+H_i^2$ is minimal.

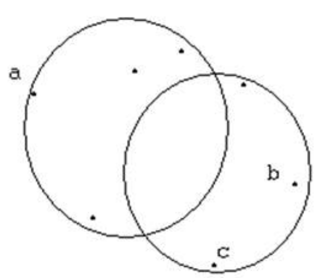
b) Splitting the Overloaded Node

When a node gets overloaded, i.e. it includes more than the maximum number of sub-nodes or points; the node needs to be split into two nodes at the same level. In the classical B-tree the node is split into two nodes with the equal number of elements.

In our case, we can split the node taking into account the following criteria:

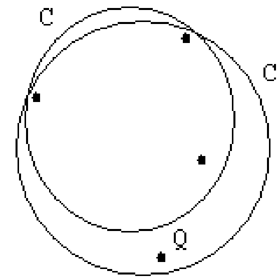
- the maximum radius of the two new circles is minimal;
- the new nodes intersect with the minimal area; the sum of standard deviations of the new nodes is
- minimal.

To this goal, first we will find the pair of points at the *longest* distance. In the figure to the right, such a pair consists of points **a** and **b**. If these points both belonged to one and the same bounding circle, the radius of this circle would be greater than the distance Da,b/2, which, as we assume, is the maximum distance. So, to minimize the maximum radius we need to distribute points **a** and **b** to the different bounding circles Ca and Cb. Now we will find point **c**, so that the distance from **c** to either of circles Ca and Cb is the longest. In the picture above, it is the distance between point **c** and circle Ca (which now consists of just one point **a**.) Once again, to minimize the would-be radii, we need to distribute point **c** to the other circle, Cb. Continuing in the same manner, we will ultimately get circles Ca and Cb, as shown in the picture.



c) Adjusting Bounding Circles

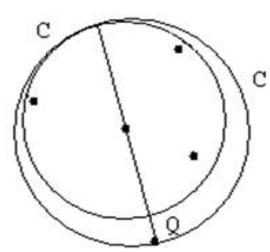
Let's assume that we need to add a new point (**Q**) to the tree. We start from the root and go down the tree, level by level. At each level we need to select the most appropriate sub-node. E.g. at the parent level P, we need to select one of the sub-nodes, C1, C2, or C3. Logically, there are two different cases. If the new point belongs to a particular bounding circle, no adjustment is needed. But if the new point is outside of any bounding circle, we need to select the most appropriate subnode; to add the new point to the selected sub-node; and to adjust the corresponding bounding circle, so that it should include the new point as well as all old points. See the figure below.



In the 2-dimensional case, the new minimal bounding circle can be exactly calculated. In the figure above, the minimal circle is the circle circumscribed about the triangle of the first two points and point **Q**. It can be calculated from the coordinates of these points. E.g. the radius of the circumscribed circle is $L_1*L_2*L_3/sqrt((L_1+L_2+L_3)*(L_2+L_3-L_1)*(L_1L_2+L_3)*(L_1+L_2$ $L_3)$), where L_1, L_2 and L_3 are the lengths of the sides.

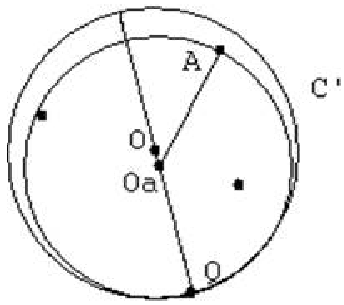
In a multi-dimensional space, the situation is much more complicated. Fortunately, we do not need the exact minimal bounding circles. The bounding circles are very useful in many procedures where they heuristically allow to reduce the amount of calculation, but fortunately they are not critical. So, we would recommend to use a less exact but easier to calculate approximation.

We can easily construct the new bounding circle (C') about the old bounding circle (C) and the newly added point (**Q**) as shown in the figure to the right. First, we calculate the distance (H) from the point **Q** to the circle C. Then we move the center of the circle C towards **Q** by H/2. This will be the center of C'. The radius of the new circle will be the old radius (R) plus H/2. Thus, the circle C' will surely contain all old points as well as the new point.



Moreover, the new bounding circle (C') can be easily optimized. By construction, the new circle has to lie through the point **Q**, but it doesn't probably lie through any other actual points. So, we can shrink the circle, so that it lies through yet another point.

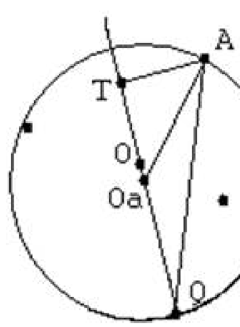
Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016



Let's shift the center (O) of the circle towards the point Q, so that the new circle with the center Oa lies both through points Q and A. See the picture above. The value of the shift can be calculated based on the points Q, O and A. When we shift the center for the point A, this doesn't necessarily mean that any other point B will belong to the new shrunken circle. But we can repeat this procedure for all points of the set, and find the largest shrunken circle, corresponding to the shortest allowed shift. That circle will work for all the points of the set.

d) Calculation of a Quasi-Minimal Bounding Circle

Let's assume that point T is the foot of the perpendicular dropped from point A on the line (Q,O.) Then $|A\,T|^2 + |T\,Q|^2 = |A\,Q|^2$. And $|A\,T|^2 + |T\,Oa|^2 = |A\,Oa|^2$. But $|T\,Oa| = |T\,Q| - |Oa\,Q|$ and $|Oa\,Q| = |Oa\,A|$. So, finally, $|Oa\,Q|$ should be $|A\,Q|^2 / 2\,|T\,Q|$, where $|T\,Q|$ is the projection of the vector [Q A] on line (Q,O) and can be calculated by means of the dot product.



e) Exact Minimal Bounding Circle

Please notice that, in some cases, the exact minimal bounding circle may be slightly smaller than the quasi-minimal circle constructed above. Moreover, the shrunken circle (C') depends on the initial circle C.

If the set consists of just one point, the exact minimal circle is the circle with the center in that point and the radius of zero. If the set contains two points, the exact minimal circle is the circle from the center of mass and the radius of half the distance between the points.

If the set contains three points, there are two cases. Either, the exact minimal circle is the circle circumscribed around these three points. Or, the circle built on the two of the three points as a diameter, provided that the third point lies inside it. If we want to build the circle for the second case, we need first to find the two (out of the three) points with the longest distance between them. Then, we need to check that the third point (X) will make an obtuse-angled triangle. I.e. $|A\,X|^2 + |X\,B|^2 < |A\,B|^2$. Sets of more than three points are quite similar.

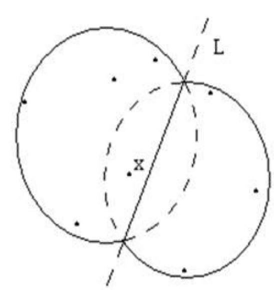
If we do not want to build the exact minimal circle, we may use the shrinking technique. But we will need a good first approximation for the minimal circle. As shown above, the circle build on the longest segment plays an essential role in construction of minimal circles; besides, and it is easy to calculate. This makes it a good first approximation for building quasi-minimal bounding areas.

VIII. Multi-Dimensional s-tree, full Details

So far we have clearly described what we would like to achieve in our multi-dimensional S-Tree. A real implementation is not so smooth, and requires solutions to a number of complicated issues.

a) Overlapped Circle

As mentioned in Section 7, some sub-nodes of a given node may overlap. Such cases may occur when a new point is added to a sub-node, which results in adjusting the bounding circle of this sub-node, or when a sub-node is split into two sub-nodes at the same level, as discussed in Section 7.2. In this case the actual areas of sub-nodes are not circle, but they are rather "cut" circles, as shown in the picture below.



The reason for the sub-node areas to be nothing but the "cut" circles is that the areas need to be convex. Now, the more complicated form of the areas makes us change the way we calculate the appropriate sub-node when the given point **x** belongs to both bounding circles.

In the two-dimensional space we use line L, to determine where point **x** belongs to. All points at one side of line L belong to one sub-node, while all the points at the other side belong to the other sub-node. Analogically, in the multi-dimensional space, the border line L will become a plane. All points at one side of the plane will belong to one sub-node, and all points at the other side of the plane L will belong to the other sub-node.

Every plane in the multi-dimensional space can be defined by its normal vector and the distance from the point of origin. The only drawback is that theoretically we will need a border plane for each pair of sub-nodes. E.g. if a parent node has, say, 5 sub-nodes, we will need to store 10 border plains. A more memory-effective way would be to calculate the equation of the plane L on the fly, based on the bounding circles.

Namely, as we discussed, the circle C1 can be defined by its center, O_1, and the radius R1. Analogically, we define the circle C2 by the center O_2 and the radius R2. It is easy to see that the border plane L is the set of all the point such as the difference between their squared distances to points O_1 and O_2 is constant.

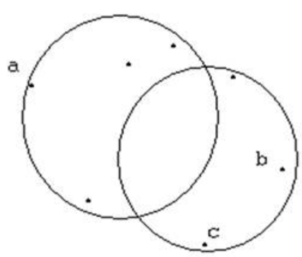
$$|O_1 - x|^2 - |O_2 - x|^2 = F_{1,2}$$

To be precise, the constant F1,2 is the difference between the squared radii of the circles in question, i.e. $F_{1,2} = R_1^2 - R_2^2$.

So, to find whether point **x** belongs to the circle C1, we need to calculate $|O_1 - x|^2 - |O_2 - x|^2 - R_1^2 + R_2^2$, and to compare it with zero.

b) Splitting the Node

When a node gets overloaded, i.e. it includes more than the maximum number of sub-nodes or points; the node needs to be split into two nodes at the same level.



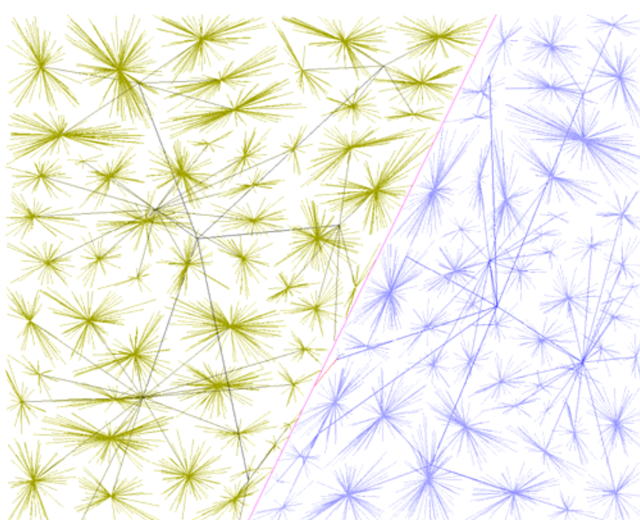
While splitting the node, we expect that the results nodes will have bounding areas with smaller radii. Let's assumes that **a** and **b** are the points with the greatest distance between them, as in the picture above. If we ultimately put these points into one and the same node, the radius of the bounding circle for that node can not be less than |**a b**|, which is almost the radius of the bounding circle for the original node. To avoid this, we have to put **a** and **b** into different nodes. Now we have nodes Ca and Cb, consists of points **a** and **b**, correspondingly.

At the next step let's consider, say, the point **c**. We can either put it into Ca or Cb. And we need to estimate how good or bad it would be to put it into a particular node. E.g. we can try to optimize (to keep minimal) the maximum radius of Ca and Cb. But, now we know that there is something more unpleasant than just big radii; it is overlapped circles. So we may want to keep the circles overlapped in the minimal possible measure.

In the previous section we have introduced the expression $L^2 - (R^2 - r^2)$. It shows in what manner the circles are overlapped, and should be greater than zero. We will try to optimize (to keep maximal) this expression. It gives the same rules for selecting points **a**, **b**, and **c**, but gives better results at next steps.

IX. Example, Two-Dimensional Case

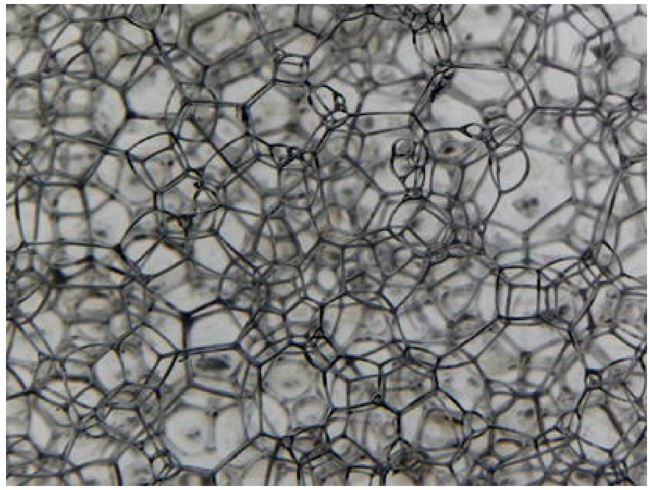
The most challenging task of implementing a B-tree is a split of overloaded nodes. When a node is split into two nodes, there appears a new boundary. All sub-nodes of the given node may need to be recursively split by the new boundary. The picture below illustrates the result of a split for a two-dimensional case.



Here the new boundary is highlighted in red. The new sub-trees are painted olive and violet. The higher levels are darker than lower levels, so that it would be easier to trace the centers of clusters.

X. Illustration, Three-Dimensional Case

As we discussed in Section 8.1, ideally, the clusters of a particular level should form isotropic ellipsoids. The result we would like to get should look like soap foam bubbles.



XI. Example Multi-Dimensional Case

As an example we will cluster wiki pages.

1. The first step is to parse an HTML page and to extract pure text. There are several tools available for this purpose. We use Java CC, an open source parser and lexical analyzer generator. The generator accepts a formal grammar definition, written in .JJ file, which also allows to define additional custom code. The input to the lexical analyzer is a sequence of characters; the output is a sequence of tokens.
   In our example, we are interested in skipping HTML tags and parsing the text further, so that it become a list of words. At this stage we are dropping everything what is not a word, i.e. numbers, email addresses, references to web pages, expressions, identifiers, etc.
2. Then, we analyze the text and map it to a point in a semantic space. For each word in text we will find the root. Basically, for verbs we drop endings as -s, -ed, -ing; for nouns we drop ending–s. Actually, the procedure is a bit more complicated due to language exceptions. Secondly, we calculate the weight of the word. We assume that more frequent words should have a lighter weight than infrequent words. So, we distributed all the words to 256 sets with about equal frequencies. Thirdly, we find the meaning of the word in question. We have split all words to 1024 groups with similar meanings. Please notice, that one word can have more than one meanings, with only one of them being actualized in the text. Without knowing what the actual meaning is, we have to add all meaning with the same weights depending on the frequency of the word.
   Now we can define a target point in a 1024-dimensional space where each dimension corresponds to a meaning. The coordinates of the target point are calculated by accumulating all weights corresponding to particular meanings. It also seems reasonable to divide the coordinates by the number of the words in the text, so that repletion of sentences or words does not affect the meaning of the text.

XII. Conclusions

Arbitrary points in multi-dimensional space can be isotropically clustered into a balanced hierarchical structure, similar to a B-tree.

Clustering into a multi-dimensional B-tree does not require any supervision or any *a priori* given information, like the number of clusters.

Clustering into a multi-dimensional tree can be done dynamically and efficiently. Adding new points to the tree requires only incremental updates of statistical values associated with nodes.

Text pages can be mapped to points into 1000-dimentional semantic space.

The search of pages close to a given semantic point can return a hierarchically ordered results, allowing the user to select more general or more specific topics.

XIII. Acknowledgments

This research was not sponsored by National Science Foundation or any other financial source.

References Références Referencias

1. Donald E. Knuth, 1973, *Sorting and Searching*, volume 3, *The Art of Computer Programming*, Addison-Wesley.
2. Antonin Guttman, 1984, *R-trees: a dynamic index structure for spatial searching,* ACM, New York, USA. (1984-r-treeguttman. pdf)
3. Sean Owen et al. 2011, *Mahout in Action,* Manning Publications Co., New York, USA. (ISBN 9781935182689, Mahout. in. Action.pdf)
4. Richard A. Reyment, K. G. Joereskog, 1993, *Applied Factor Analysis in the Natural Sciences*, Cambridge University Press, UK.
5. George A. Miller, 2003, *WordNet Lexical Database of English Language*, Cognitive Science Laboratory of Princeton University
6. Roget's Thesaurus, 2006, Electronic Lexical Knowledge Base (ELKB) http://www.nzdl.org/ELKB
7. Adam Kilgarriff, 1995, *BNC Database and Word Frequency Lists*, http://www.kilgarriff.co.uk/bnc-readme.html
8. Brian Goetz's, 2003, *HTML Parser.*

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Evolution of Object-Oriented Database Systems

By Hibatullah Alzahrani

Clark Atlanta University, United States

Abstract- Data bases are quintessential part of most modern web and mobile applications. In most part, relational databases dominate the database market but the evolution of object-oriented databases has provided users and developers with an alternative option. Object-oriented databases provide a number of advantages over relational databases like ease of extensibility, custom data models, provision for modelling complex data structures and faster access time. But they do lack in certain areas and have no strict standards and implementation mostly depends upon the vendor. Nevertheless, object-oriented databases are slowly finding their way into database market, especially in the area of large-scale databases. But the long history of relational databases keeps them alive as tough competitor and the future seems to be going towards object-relational databases.

Keywords: object-oriented, database, relational, data- base management system, evolution, advantages, disadvantages.

GJCST-C Classification : *H.2.4*

EVOLUTIONOFOBJECTORIENTEDDATABASESYSTEMS

Strictly as per the compliance and regulations of:

Evolution of Object-Oriented Database Systems

Hibatullah Alzahrani

Abstract- Data bases are quintessential part of most modern web and mobile applications. In most part, relational databases dominate the database market but the evolution of object-oriented databases has provided users and developers with an alternative option. Object-oriented databases provide a number of advantages over relational databases like ease of extensibility, custom data models, provision for modelling complex data structures and faster access time. But they do lack in certain areas and have no strict standards and implementation mostly depends upon the vendor. Nevertheless, object-oriented databases are slowly finding their way into database market, especially in the area of large-scale databases. But the long history of relational databases keeps them alive as tough competitor and the future seems to be going towards object-relational databases.

Keywords: object-oriented, database, relational, database management system, evolution, advantages, disadvantages.

I. Introduction

Databases are the nuts and bolts of the modern information systems. Every major application on Internet and smartphones uses them in one way or another. They are ubiquitously used in data centers and for maintaining records in hospitals, universities and all kinds of government and private institutions. Strictly speaking, there is a distinction between a database and a database management system (DBMS) – database is an organized collection of data whereas DBMS is a software which interacts with the database and the user and acts as an interface between them. But usually database is used to refer to both the database itself and the DBMS. Most commonly used DBMS is Relational DBMS (RDBMS) which is based on relational data model in which data is stored as tables or "relations" consisting of rows and columns. With the advent of object-oriented programming paradigm and the rise of object-oriented programming languages, the concept of object-oriented databases was conceived in which data is represented as objects rather than as tables. Figure 1 provides a mapping between the relational and object-oriented database model. In this article, we will briefly discuss what object-oriented databases are, trace the evolution of object-oriented databases, their use in modern systems and their advantages and disadvantages over traditional Relational Databases.

Author: e-mail: mesfer66@gmail.com

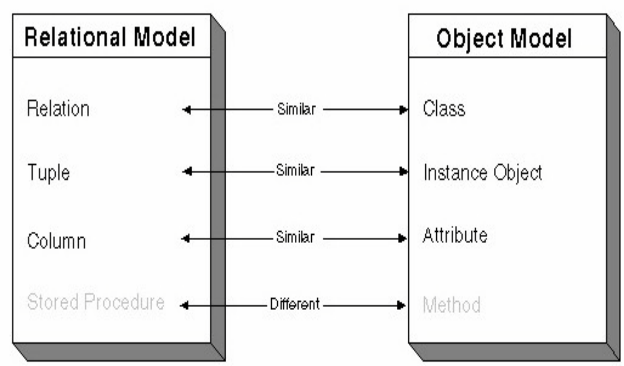


Figure 1 : Relational vs Object-Oriented Data Model

II. What are Object-Oriented Databases

Object-oriented databases are designed and built according to the object-oriented paradigm in which everything is modeled as objects including the data. This type of data model helps in tackling complex data structures, for instance multimedia content, in a more natural way and provides a seamless transition from design to conception. According to object-oriented database system manifesto [1], an Object-Oriented Database Management System (OODBMS) must satisfy two criterion:

i) It should be a Database Management System (DBMS)
ii) It should be an object-oriented system

The first criterion means that the OODBMS should provide the five features which are must for any database system – persistence, concurrency, data recovery, secondary storage management and ad hoc query facility. The second criterion means that the database system should support all the requisite features of an object-oriented system like encapsulation, complex objects, inheritance, polymorphism, extensibility etc. Hence the data in OODBMS is represented as collection of interacting objects instead of collection of inter-related tables. Usage of object-oriented concepts like polymorphism and inheritance make the interaction between the objects a trivial task. Figure 2 provides an example of how the same data, customer account information for a banking system, is represented in two different formats. Whereas data is stored as tables in the relational database and we need to relate of "join" tables to perform a query, it is stored as a collection of objects in object-oriented database, and query can be easily performed by following the pointer from parent object to its children.

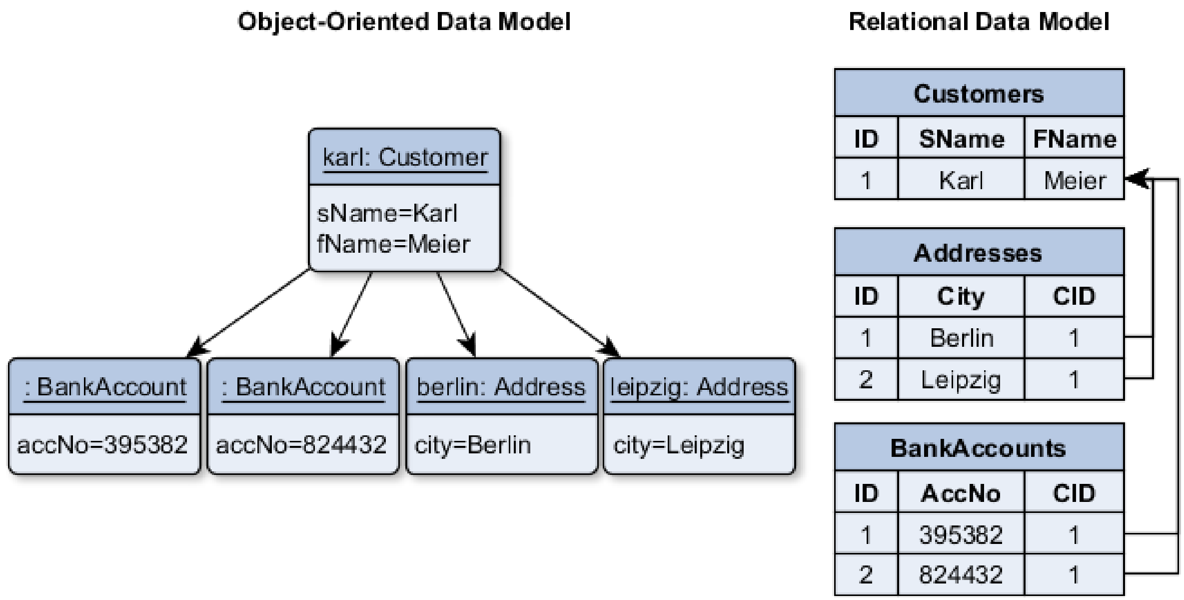


Figure 2 : Example of Data Representation in Relational vs Object-Oriented Data Model

II. Evolution of Object-Oriented Databases

The term "object-oriented database system" was first introduced in 1985 in [2] and [3]. Orion Research Project at MCC headed by Won Kim was the first major project initiated for the development of OODBMS and several papers were published during its course, best of which were compiled by Won Kim in the form of book [4] in 1990. Meanwhile Servo-Logic [5] began work on one of the first commercial projects which was later renamed as Gem Stone [6] in 1995 and was based on Small Talk object-oriented language. Another Lisp-based system, Graphael, was introduced at around the same time in France and it was followed by O2 [7] in 1992 which was later acquired by IBM. Tom Atwood at Onto logic produced V base which later became ONTOS and was the first to support C++. Drew Wade at same time produced Objectivity/DB [8], [9]. In 1989, first OODBS manifesto [1] was published by Malcolm Atkinson which criticized relational model for being inadequate in meeting the demands of new applications and laid down the criterion for the object-oriented database system. A year later, second manifesto [10] was published, which went the opposite way by supporting relational model and wanting to support SQL for OODBS. The third and final manifesto was published in 1995 [11] and it presented OO model as an extension of relational model and proposed to extend the relational model to incorporate the object-oriented characteristics by allowing for custom user-defined data types.

Meanwhile, work on standardization of OODBMS began in 1991 when Rick Cattell of Sun Microsystems formed a consortium of 5 major OODBMS vendors, named ODMG (Object Database Management Group). As a result of these efforts for standardization, standards were published for ODL (Object Definition Language), OQL (Object Query Language) and OML (Object Manipulation Language). First standard was released in 1993 [12], which was mostly designed with C++ in mind as primary object-oriented language but in the final release 3.0 in 2001 [13], Java bindings were added to the standard to support Java and later bindings for Small Talk were also added. Afterwards, the ODMG was disbanded. Meanwhile, some object-oriented features were also included in SQL: 1999 [14] and were then revised in its 2003 version and then in the latest 2011 version [15]. SQL: 2011 supports Object Language Bindings (SQL/OLB) for performing queries to Object-oriented databases and various other object oriented features.

In 2004, db4o [16] was released as the first free open-source OODBMS and it was the first DBMS to implement native queries in the programming language itself like Java and C#. Also Perst and DTS/S1 were made available under dual open-source and commercial licenses. In 2005, Microsoft implemented Native Queries in its .NET framework by introducing LINQ and DLINQ. As a result, many object-oriented languages now support native queries. Though this is not the same as a full-fledge open-source database system, but it does provide the user or developer with an alternative option other than commercial databases, important since commercial OO databases are usually costlier than relational ones.

III. Advantages of OO Databases

Since Relational Databases have been the norm of the day for a long period of time, any new emerging technology is compared against this benchmark to measure its usefulness. Some of the advantages of OODBMS over traditional RDBMS are:

a) Complex Data Models

In relational databases, all data is modeled as relations or tables and it is extremely complicated to model complex data structures. Object-oriented databases model all data as objects which can model even complex data structures very easily. In fact, one of the very first application for object-oriented databases were graph-based data structures which were modelled easily using object-oriented concepts but were a nightmare to model in relational databases.

b) Real-World Modelling

One of the strengths of object-oriented paradigm is its ability to model the real-world objects in a natural way. This becomes useful in maintaining data when we can store it as an object rather than a table and then perform manipulations on it. For instance, a very important application of OO databases is maintaining multimedia content. While using relational databases, it is cumbersome to store this content in form of tables but OO database can model the whole multimedia document as an object and store it easily.

c) Extensibility (Provision of New Custom Data Types)

OO databases provides the ability to form new data types from existing ones using the object-oriented concepts of inheritance and polymorphism. No such feature is available in relational databases.

d) No Impedance Mismatch

In OODBMS, there is no mismatch between data represented by database system and the data representation required by application. We can model our data as required by the application and subsequently use it directly in our application. This is in contrast to RDBMS in which all data has to be stored in form of tables and then at runtime, data is manipulated into the form required by the application.

e) Ease of Design and Implementation

Relational Data Models are not very descriptive in nature and in general, the database is first designed using Entity Relationship (ER) model and then implemented using relational model. Object-oriented databases do away with this hassle by designing the data directly as objects and then implementing it as such.

f) Faster Data Access and Improved Performance

OODBMS are usually much faster than relational ones since they have a many-to-many relationship and objects can be accessed using pointers only. Furthermore, there is no need for 'join' as objects are linked through pointers and any specific object can be found following the chain of pointers.

g) Easy handling of very large data

Object-oriented databases can very easily handle very large amounts of complex data and as a result, very large databases are often built using OODBMS. In fact, world's largest database is an OODBMS – Stanford Linear Accelerator Center (SLAC) BaBar database [17] uses Objectivity/DB [9] and currently sizes almost 900 TB.

IV. Disadvantages of OO Databases

Despite all its advantages, OODBMS are still not as popular as relational databases due to following reasons:

a) No Universal Data Model

Relational databases have a fixed data model in which data is always represented in the form of tables. No such standard exists for object-oriented databases in which data is modeled as custom objects and depends on the type of data and need of application.

b) No Standard Query Language

There is no standard query language for OODBMS like SQL for Relational Databases. Even though ODMG standardized OQL (Object Query Language) but it is still widely unimplemented. More recently, trend has been towards implementing Native Queries in programming languages like LINQ in C# or the database vendor provides separate bindings for most popular object-oriented languages like Java, C++ etc.

c) No Mathematical Foundation

Relational Databases are based on the solid foundations of Relational Algebra and Relational Calculus. No such mathematical foundation exists for object-oriented databases.

d) Lack of Ad hoc Queries or Closure

Closure is a property of relational databases which enables nested queries where new tables are created by joining existing ones and then querying the new table. Since there are no joins in OODBMS, there are no nested queries and the nature of query that can be performed is highly dependent on the design of database. Hence strictly speaking, some OODBMS can violate an essential criterion for database management systems which requires that all database management system should support ad hoc query.

V. Conclusion

Object-oriented databases have been around for quite some time now but they haven't found widespread acceptance like relational databases. They provide some very nice features which are absent in relational databases like custom data types, faster access and support for modeling real-world complex data structures. But still, they are not as ubiquitous as relational databases. The present trend is towards incorporating the good features of both types of databases forming a hybrid one, called *Object-Relational Database*, which is intrinsically a relational database with object-oriented features. Even the modern standards of SQL provide some object-oriented features. Thus integration of object-oriented features in relational databases highlight their importance and the competition they pose. Already, very big databases are being designed using object-oriented paradigm and as databases become larger and larger, it is inevitable that the trend would go towards object-oriented databases in future.

ReferencesRéférences Referencias

1. M. Atkinson, F. Bancilhon, D. Dewitt, K. Dittrich, D. Maier, S. Zdonik, "The Object-Oriented Database System Manifesto". *In Proc. of the First International Conference on Deductive and Object-Oriented Databases*, pages 223-40, Kyoto, Japan, December 1989.
2. T. Atwood, "An Object-Oriented DBMS for Design Support Applications," *Proceedings of the IEEE COMPINT 85*, pp. 299-307, September 1985
3. N. Derrett, W. Kent, and P. Lyngbaek, "Some Aspects of Operations in an Object-Oriented Database," *Database Engineering*, vol. 8, no. 4, IEEE Computer Society, December 1985
4. Kim, Won. *Introduction to Object-Oriented Databases*. The MIT Press, 1990. ISBN 0-262-11124-1.
5. D. Maier, A. Otis, and A. Purdy, "Object-Oriented Database Development at Servio Logic," *Database Engineering*, vol. 18, no.4, December 1985.
6. https://gemtalksystems.com/products/gss32/
7. Bancilhon, Francois; Delobel,Claude; and Kanellakis, Paris. *Building an Object-Oriented Database System: The Story of O_2.* Morgan Kaufmann Publishers, 1992.
8. *Angela Guess (February 6, 2013). "Objectivity Launches Objectivity/DB 11.0". DATAVERSITY. Retrieved December 2, 2014.*
9. http://www.objectivity.com/
10. The Committee for Advanced DBMS Function, Third Generation Database System Manifesto, *In Computer Standards and Interfaces 13 (1991)*, pages 41-54. North Holland.
11. H. Darwen and C.J. Date, *The Third Manifesto*, SIGMOD RECORD 24(1):39-49, March 1995.
12. R.G.G. Cattell (ED), *The Object Database Standard: ODMG-93, Release 1.1* The Morgan Kaufmann Series in Data Management Systems, 1994
13. *The Object Data Standard: ODMG 3.0.* Edited by R.G.G. Cattell and Douglas K. Barry, with contributions by Mark Berler, Jeff Eastman, David Jordan, Craig L. Russell, Olaf Schadow, Torsten Stanienda, and Fernando Velez. Morgan Kaufmann Publishers, Inc., 2000.
14. Jim Melton, *Advanced SQL, 1999: Understanding Object-Relational and Other Advanced Features.* Morgan Kaufmann Publishers, Inc., 2003.
15. Zemke, Fred. "*What's new in SQL:2011*". ACM SIGMOD Record 41.1 (2012): 67-73.
16. Stefan Edlich, Jim Paterson, Henrik Hörning, Reidar Hörning, "*The definitive guide to db4o*", Apress, 2004
17. http://www.slac.stanford.edu/BFROOT/www/Public/Computing/Databases/

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Frame Work for Text Mining using Learned Information Extraction System

By M. Vasavi & Sathish Kuppani

SV University

Abstract- Text mining is a very exciting research area as it tries to discover knowledge from unstructured texts. These texts can be found on a computer desktop, intranets and the internet. The aim of this paper is to give an overview of text mining in the contexts of its techniques, application domains and the most challenging issue. The Learned Information Extraction (LIE) is about locating specific items in natural-language documents. This paper presents a framework for text mining, called DTEX (Discovery Text Extraction), using a learned information extraction system to transform text into more structured data which is then mined for interesting relationships. The initial version of DTEX integrates an LIE module acquired by an LIE learning system, and a standard rule induction module. In addition, rules mined from a database extracted from a corpus of texts are used to predict additional information to extract from future documents, thereby improving the recall of the underlying extraction system. Applying these techniques best results are presented to a corpus of computer job announcement postings from an Internet newsgroup.

GJCST-C Classification : I.2.4, D.3.3

A Frame Work for Text Mining using Learned Information Extraction System

M. Vasavi [α] & Sathish Kuppani [σ]

Abstract- Text mining is a very exciting research area as it tries to discover knowledge from unstructured texts. These texts can be found on a computer desktop, intranets and the internet. The aim of this paper is to give an overview of text mining in the contexts of its techniques, application domains and the most challenging issue. The Learned Information Extraction (LIE) is about locating specific items in natural-language documents. This paper presents a framework for text mining, called DTEX (Discovery Text Extraction), using a learned information extraction system to transform text into more structured data which is then mined for interesting relationships. The initial version of DTEX integrates an LIE module acquired by an LIE learning system, and a standard rule induction module. In addition, rules mined from a database extracted from a corpus of texts are used to predict additional information to extract from future documents, thereby improving the recall of the underlying extraction system. Applying these techniques best results are presented to a corpus of computer job announcement postings from an Internet newsgroup.

I. Introduction

In this modern culture, text is the most common vehicle for the formal exchange of information. Although extracting useful information from texts is not an easy task, it is a need of this modern life to have a business intelligent tool which is able to extract useful information as quick as possible and at a low cost. Text mining is a new and exciting research area that tries to take the challenge and produce the intelligence tool. The tool is a text mining system which has the capability to analyse large quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to extract meaningful and useful information [1]. The aim of text mining tools is to be able to answer sophisticated questions and perform text searches with an element of intelligence. Technically, text mining is the use of automated methods for exploiting the enormous amount of knowledge available in text documents. Text Mining represents a step forward from text retrieval. It is a relatively new and vibrant research area which is changing the emphasis in text-based information technologies from the level of retrieval to the level of analysis and exploration. Text mining, sometimes alternately referred to as text data mining, refers generally to the process of deriving high quality information from text. Researchers like [2], [3] and others pointed that text mining is also known as Text Data The problem of text mining, i.e. discovering useful knowledge from unstructured or semi-structured text, is attracting increasing attention [4, 18, 19, 21, 22, 27]. This paper suggests a new framework for text mining based on the integration of Learned Information Extraction (LLIE) and Knowledge Discovery from Databases (KDD), a.k.a. data mining. KDD and LIE are both topics of significant recent interest. KDD considers the application of statistical and machine-learning methods to discover novel relationships in large relational databases. LIE concerns locating specific pieces of data in natural-language documents, thereby extracting structured information from free text. However, there has been little if any research exploring the interaction between these two important areas. In this paper, we explore the mutual benefit that the integration of LLIE and KDD for text mining can provide.

Traditional data mining assumes that the information to be “mined” is already in the form of a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural-language documents rather than structured databases. Since LLIE addresses the problem of transforming a corpus of textual documents into a more structured database, the database constructed by an LLIE module can be provided to the KDD module for further mining of knowledge as illustrated in Figure 1. Information extraction can play an obvious role in text mining as illustrated.

Author α: *Asst.professor, Department of Computer Applications, RVR & JC College of Engineering, Chowdavaram .Guntur-19.*
Author σ: *College Engineering, S.V. University, Tirupati.*

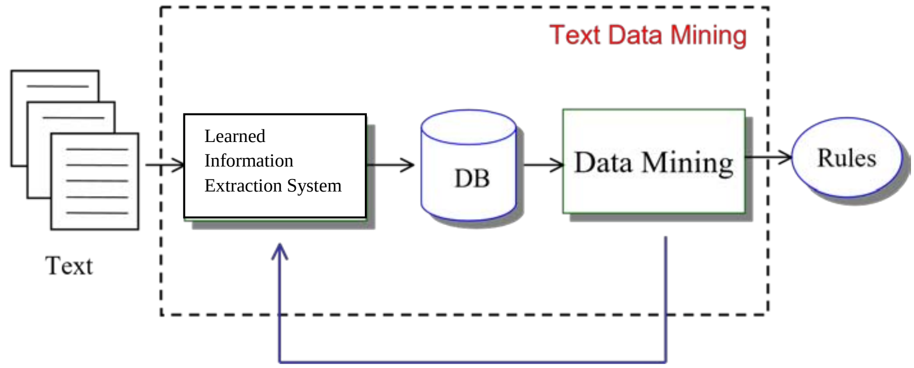


Figure 1 : Overview of LIE-based text mining framework

The constructing an LIE system is a difficult task, there has been significant recent progress in using machine learning methods to help automate the construction of LIE systems [5, 7, 9, 23]. By manually annotating a small number of documents with the information to be extracted, a reasonably accurate LIE system can be induced from this labelled corpus and then applied to a large corpus of text to construct a database. However, the accuracy of current LIE systems is limited and therefore an automatically extracted database will inevitably contain significant numbers of errors. An important question is whether the knowledge discovered from this "noisy" database is significantly less reliable than knowledge discovered from a cleaner database. This paper presents experiments showing that rules discovered from an automatically extracted database are close in accuracy to that discovered from a manually constructed database.

A less obvious interaction is the benefit that KDD can in turn provide to LIE. The predictive relationships between different slot fillers discovered by KDD can provide additional clues about what information should be extracted from a document. For example, suppose we discovered that computer-science jobs requiring "My SQL" skills are "database" jobs in many cases. If the LIE system manages to locate "My SQL" in the language slot but failed to extract "database" in the area slot, we may want to assume there was an extraction error. Since typically the *recall* (percentage of correct slot fillers extracted) of an LIE system is significantly lower than its *precision* (percentage of extracted slot fillers which are correct) [13], such predictive relationships might be productively used to improve recall by suggesting additional information to extract. This paper reports experiments in the computer-related job-posting domain demonstrating that predictive rules acquired by applying KDD to an extracted database can be used to improve the recall of information extraction.

The remainder of the paper is organized as follows. Section 2 presents some background information on text mining and LIE. Section 3 describes a system called DTEX (Discovery from Text EXtraction) that combines LIE and KDD for text mining. Section 4 presents and discuss performance gains obtained in LIE by exploiting mined prediction rules. Section 5 discusses some related work, Section 6 outlines directions for future research, and Section 7 presents our conclusions.

II. Background: Text Mining and Information Extraction

"Text mining" is used to describe the application of data mining techniques to automated discovery of useful or interesting knowledge from unstructured text [20]. Several techniques have been proposed for text mining including conceptual structure, association rule mining, episode rule mining, decision trees, and rule induction methods. In addition, Information Retrieval (IR) techniques have widely used the "bag-of-words" model [2] for tasks such as document matching, ranking, and clustering.

The related task of information extraction aims to find specific data in natural-language text. DARPA's Message Understanding Conferences (MUC) have concentrated on LIE by evaluating the performance of participating LIE systems based on blind test sets of text documents [13]. The data to be extracted is typically given by a template which specifies a list of slots to be filled with substrings taken from the document. Figure 2 shows a (shortened) document and its filled template for an information extraction task in the job-posting domain. This template includes slots that are filled by strings taken directly from the document. Several slots may have multiple fillers for the job-posting domain as in programming languages, platforms, applications, and areas.

We have developed machine learning techniques to automatically construct information extractors for job postings, such as those listed in the USENET newsgroup misc. jobs. offered [6]. By extracting information from a corpus of such textual job postings, a structured, searchable database of jobs can be automatically constructed; thus making the data in online text more easily accessible. LIE has been shown to be useful in a variety of other applications, e.g.

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016

seminar announcements, restaurant guides, university web pages, apartment rental ads, and news articles on corporate acquisitions [5, 9, 23].

The most related system to our approach is probably DOCUMENT EXPLORER [14] which uses automatic term extraction for discovering new knowledge from texts. However, DOCUMENT EXPLORER assumes semi-structured documents such as SGML text unlike DTEX developed for general natural-language text. Similarly, automatic text categorization has been used to map web documents to pre-defined concepts for further discovery of relationships among the identified concepts [24]. One of the limitations for these approaches is that they require a substantial amount of domain knowledge.

Several rule induction methods and association rule mining algorithms have been applied to databases of corporations or product reviews automatically extracted from the web [17, 16, 33]; however, the interaction between LIE and rule mining has not been addressed. Recently a probabilistic framework for unifying information extraction and data mining has been proposed [25]. In this work, a graphical model using conditional probability theory is adopted for relational data, but experimental results on this approach are yet to be gathered. A boosted text classification system based on link analysis [12] is related to our work in spirit in that it also trLIEs to improve the underlying learner by utilizing feedback from a KDD module.

III. Integrating Data Mining and Information Extraction

In this section, it discusses the details of our proposed text mining framework, DTEX (Discovery from Text Extraction). We consider the task of first constructing a database by applying a learned information-extraction system to a corpus of natural-language documents. Then, we apply standard data-mining techniques to the extracted data, discovering knowledge that can be used for many tasks, including improving the accuracy of information extraction.

a) The DTEX System

In the proposed framework for text mining, LIE plays an important role by pre-processing a corpus of text documents in order to pass extracted items to the data mining module. In our implementations, we used two state-of-the-art systems for learning information extractors, RAPLIER (Robust Automated Production of Information Extraction Rules) [6] and BWI (Boosted Wrapper Induction) [15]. By training on a corpus of documents annotated with their filled templates, they acquire a knowledge base of extraction rules that can be tested on novel documents. RAPLIER and BWI

Document

Title: Web Development Engineer

Location: Beaverton, Oregon

This individual is responsible for design and implementation of the web-interfacing components of the Access Base server, and general back-end development duties.

A successful candidate should have experience that includes:

One or more of: **Solaris, Linux, IBM AIX, plus Windows/NT**

Programming in **C/C++, Java**

Database access and integration: **Oracle, ODBC**

CGI and scripting: **one or more of JavaScript, VBScript, Perl, PHP, ASP**

Exposure to the following is a plus: **JDBC, Flash/Shockwave, FrontPage and/or Cold Fusion.**

A BSCS and 2+ years' experience (or equivalent) is required.

Filled Template

- title: "Web Development Engineer"
- location: "Beaverton, Oregon"
- languages: "C/C++", "Java", "Javascript", "VBScript", "Perl", "PHP", "ASP"
- platforms: "Solaris", "Linux", "IBM AIX", "Windows/NT"
- applications: "Oracle", "ODBC", "JDBC", "Flash/Shockwave", "FrontPage", "Cold Fusion"
- areas: "Database", "CGI", "scripting"
- degree required: "BSCS"
- years of experLIEnce: "2+ years"

Figure 2 : Sample text and filled template for a job posting

Table 1 : Synonym dictionary (partially shown)

Standard Term	Synonyms
"Access"	"MS Access", "Microsoft Access"
"ActiveX"	"Active X"
"AI"	"Aritificial Intelligence"
"Animation"	"GIF Animation", "GIF Optimization/Animation"
"Assembly"	"Assembler"
"ATM"	"ATM Svcs"
"C"	"ProC", "Objective C"
"C++"	"C ++", "C+ +"
"Client/Server"	"Client Server", "Client-Server", "Client / Server"
"Cobol"	"Cobol II", "Cobol/400", "Micro focus Cobol"

Job postings (600)

- Oracle $\in$ application and QA partner $\in$ application $\rightarrow$ SQL $\in$ language
- HTML $\in$ language and Windows $\in$ platform and Active Server pages $\in$ application $\rightarrow$ data base $\in$ area.
- Java $\in$ language and Active X $\in$ area and Graphics $\in$ area $\rightarrow$ Web $\in$ area
- UNIX $\notin$ platform and Windows $\notin$ platform and Games $\in$ are $\rightarrow$ 3D $\in$ area
- AIX $\in$ platform and Sybase $\notin$ application and DB2 $\in$ application $\rightarrow$ Lotus Notes $\in$ application
- C++ $\in$ language and C $\in$ language and CORBA $\in$ application and Title = Software Engineer $\rightarrow$ Windows $\in$ platform.

Figure 3 : Sample mined prediction rules for computer-science jobs

have been demonstrated to perform well on realistic applications such as USENET job postings and seminar announcements.

After constructing an LIE system that extracts the desired set of slots for a given application, a database can be constructed from a corpus of texts by applying the LIE extraction patterns to each document to create a collection of structured records. Standard KDD techniques can then be applied to the resulting database to discover interesting relationships. Specifically, we induce rules for predicting each piece of information in each database field given all other information in a record. In order to discover prediction rules, we treat each slot-value pair in the extracted database as a distinct binary feature, such as "graphics $\in$ area", and learn rules for predicting each feature from all other features.

Similar slot fillers are first collapsed into a pre-determined standard term. For example, "Windows XP" is a popular filler for the platforms slot, but it often appears as "Win XP", "Win XP", 'MS Win XP", and so on. These terms are collapsed to unique slot values before rules are mined from the data. In our experiment, a manually-constructed synonym dictionary with 111 entries was employed. Table 1 shows the first 10 entries of the dictionary.

We have applied C4.5 RULES [34] to discover interesting rules from the resulting binary data.

Resume posting (600)

- HTML $\in$ language and DHTML $\in$ language $\rightarrow$ HML $\in$ languages
- Illustrator $\in$ application $\rightarrow$ Flash $\in$ application
- Dreamweaver 4 $\in$ application and Web Design $\in$ area $\rightarrow$ Photoshop 6 $\in$ application
- MS Excel $\in$ application $\Rightarrow$ MS Access $\in$ application
- ODBC $\in$ application $\Rightarrow$ JSP $\in$ language
- Perl $\in$ language and HTML $\in$ language $\Rightarrow$ Linux $\in$ plat form

Figure 4 : Sample rules of D TEX for computer-science resume posting

SF Book descriptions (1,500)

- Sign of the Unicorn $\in$ related books and American Science Fiction $\in$ subject $\Rightarrow$ Knight of Shadows $\in$ related books.

- Spider Robinson $\in$ author $\Rightarrow$ Jeanne Robinson $\in$ author
- Roger Zelany $\in$ author $\Rightarrow$ 5 $\in$ average rating

Figure 5 : Sample rules of DTEX for book descriptions

Discovered knowledge describing the relationships between slot values is written in the form of production rules. If there is a tendency for "Web" to appear in the area slot when "Director" appears in the applications slot, this is represented by the production rule, "Director.

Web". Rules can also predict the absence of a filler in a slot; however, here it focusses on rules predicting the presence of fillers. Since any LIE or KDD module can be plugged into the DTEX system, we also tested a highly-accurate information extractor (wrapper) manually developed for a book recommending system [28] to find interesting patterns from a corpus of book descriptions. Sample association rules mined from a collection of 1,500 science fiction (SF) book descriptions from the online Amazon.com bookstore are shown in Figure 5. Slots such as authors, titles, subjects, related books, and average customer ratings are identified from the corpus.

a) Evaluation

Discovered knowledge is only useful and informative if it is accurate. Therefore, it is important to measure the accuracy of discovered knowledge on independent test data. The primary question we address in the experiments of this section is whether knowledge discovered from automatically extracted data (which may be quite noisy due to extraction errors) is relatively reliable compared to knowledge discovered from a manually constructed database.

For the dataset, 600 computer-science job postings to the newsgroup austin. jobs were collected and manually annotated with correct extraction templates. Ten-fold cross validation was used to generate training and test sets. RAPLIER was used to learn the LIE component and RIPPER was used as the KDD component. Rules were induced for predicting the fillers of the languages, platforms, applications, and areas slots, since these are usually filled with multiple discrete-valued fillers and have obvious potential relationships between their values (See [30] for more details on this experiment).

In order to test the accuracy of the discovered rules, they are used to predict the information in a database of user-labelled examples. For each test document, each possible slot-value is predicted to be present or absent given information on all of its other slot-values. Average performance across all features and all test examples were then computed.

The classification accuracy for predicting the absence or presence of slot fillers is not a particularly informative performance metric since high accuracy can be ach LI Eved by simply assuming every slot filler is absent. This is because the set of potential slot fillers is very large and only a small fraction of possible fillers is present in any given example. Therefore, we evaluate the performance of DTEX using the LIE performance metrics of precision, recall, and F-measure with regard to predicting slot fillers. These metrics are defined as follows:

$$Precision = \frac{Numberactualslotvaluescorrectlypredicted}{Numberslotvaluespredictedtobepresent} \tag{1}$$

$$\text{recall}=\frac{\text{Number of actual slot values correctly predicted}}{\text{Number of actual slot values}} \tag{2}$$

We also report F-measure which is the harmonic mean of recall and precision:

$$\text{F-measures}=\frac{2\times precision\times recall}{precision\times recall} \tag{3}$$

Before constructing a database using an LIE system, we filtered out irrelevant documents from the newsgroup using a bag-of-words Naive-Bayes text categorizer [26]. 200 positive documents (computer-science job postings) and 20 negative examples (spam postings, resume´s, or non-cs job postings) are provided to the classifier for training. The performance of the classifier trained to predict the class" relevant" was reasonably good; precision is about 96% and recall is about 98%.

RAPLIER was trained on only 60 labelled documents, at which point its accuracy at extracting information is somewhat limited; extraction precision is about 91.9% and extraction recall is about 52.4% . We purposely trained RAPLIER on a relatively small corpus in order to demonstrate that labelling only a relatively small number of documents can result in a good set of extraction rules that is capable of building a database from which accurate knowledge can be discovered.

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016

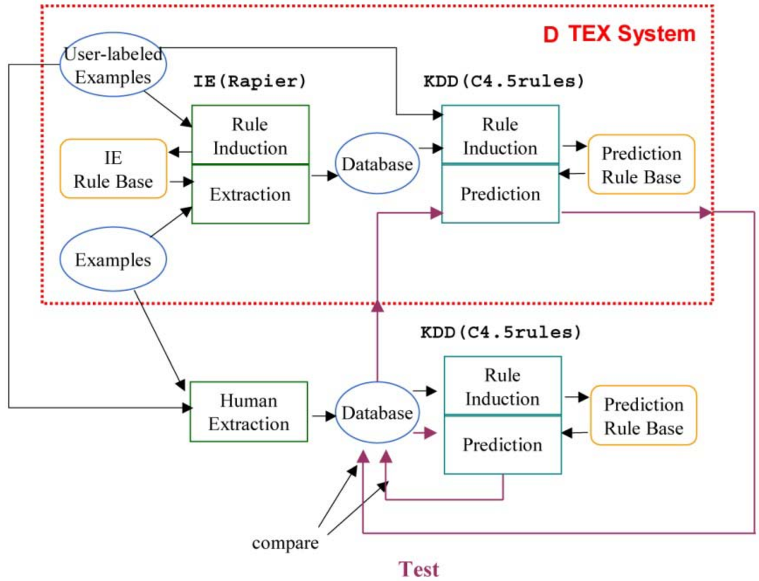


Figure 6 : The system architecture - training and testing

Because of the two different training phases used in DTEX, there is a question of whether or not the training set for LIE should also be used to train the rule-miner. To clearly illustrate the difference between mining human-labelled and LIE-labelled data, the LIE training data are thrown away once they have been used to train RAPLIER and ten-fold cross-validation is performed on the remaining 540 examples for evaluation of the data mining part. The same set of training examples was provided to both KDD systems, whereas the only difference between them is that the training data for DTEX is automatically extracted by RAPLIER after being trained on a disjoint set of 60 user-labelled examples. The overall architecture of the final system is shown in Figure 6.

Figure 7 shows the learning curves for precision, recall, and F-measure of both system as well as a random guessing strategy used as a baseline. The random guessing method predicts a slot value based on its frequency of occurrence in the training data. Even with a small amount of user-labelled data, the results indicate that DTEXachieves a performance fairly comparable to the rule-miner trained on a manually constructed database.

IV. Mined Rules to Improve Lie

After mining knowledge from extracted data, DTEX can predict information missed by the previous extraction using discovered rules. In this section, we discuss how to use mined knowledge from extracted data to aid information extraction itself.

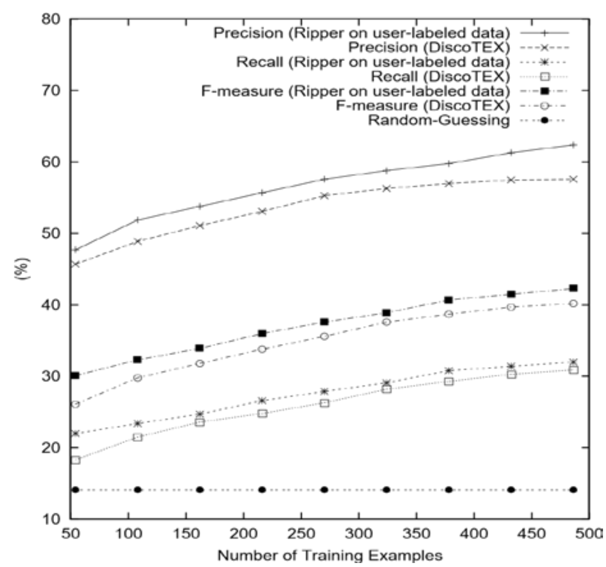


Figure 7 : User-labelled data vs. LIE-labelled data in rule accuracy

a) The Algorithm

Tests of LIE systems usually consider two performance measures, *precision* and *recall* defined as:

$$Precision = \frac{Numberactualslotvaluescorrectlypredicted}{Numberslotvaluespredictedtobep \quad resent} \tag{4}$$

$$\text{recall} = \frac{\text{Number of actual slot values correctly predicted}}{\text{Number of actual slot values}}$$

Many extraction systems provide relatively high precision, but recall is typically much lower. Previous experiments in the job postings domain showed RAPLIER's precision (e.g. low 90%'s) is significantly higher than its recall (e.g. mid 60%'s) [6]. Currently, RAPLIER's search focuses on finding high-precision rules and does not include a method for trading-off precision and recall. Although several methods have been developed for allowing a rule learner to trade-off precision and recall [11], this typically leaves the overall F-measure unchanged.

By using additional knowledge in the form of prediction rules mined from a larger set of data automatically extracted from additional unannotated text, it may be possible to improve recall without unduly sacrificing precision. For example, suppose we discover the rule "Voice XML

" "Mobile". If the LIE system extracted "VoiceXML" but failed to extract "Mobile", we may want to assume there was an extraction error and add "Mobile" to the area slot, potentially improving recall. Therefore, after applying extraction rules to a document, DTEXapplies its mined rules to the resulting initial data to predict additional potential extractions.

First, we show the pseudocode for the rule mining phase in Figure 8. A final step shown in the figure is filtering the discovered rules on both the training data and a disjoint set of labeled validation data in order to retain only the most accurate of the induced rules. Currently, rules

Input: *D* is the set of document.

Output: *RB* is the set of prediction rules.

Function Rule Mining (*D*)

Determine *T*, a threshold value for rule validation

Create a database of labelled examples (by applying LIE to the document corpus, *D*)

For each labelled example $D \in D$ **do**

F := set of slot fillers of *D*

Convert *F* to binary features

Build a prediction rule base, *RB* (by applying rule miner to the binary data, *F*)

For each prediction rule $R \in RB$ **do**

Verify *R* on training data and validation data

If the accuracy ofis lower than *T*

Delete *R* from *RB*

Return *RB*.

Figure 8 : Algorithm specification: rule mining

Input: *RB* is the set of prediction rules.

D is the set of documents.

Output: *F* is the set of slot fillers extracted.

Function Information Extraction *(RB, D)*

For each example $D \in D$ **do**

Extract fillers from using extraction rules and add them to *F*

For each rule in the prediction rule base *RB* do

If *R* fires on the current extracted fillers

If the predicted filler is a substring of *D*

Extract the predicted filler and add it to *F*

Return F.

Figure 9 : Algorithm specification: LIE

that make *any* incorrect predictions on either the training or validation extracted templates are discarded. Since association rules are not intended to be used together as a set as classification rules are, we focus on mining prediction rules for this task.

The extraction algorithm which attempts to improve recall by using the mined rules is summarized in Figure 9. Note that the final decision whether or not to extract a predicted filler is based on whether the filler (or any of its synonyms) occurs in the document as a substring. If the filler is found in the text, the extractor considers its prediction confirmed and extracts the filler.

One final issue is the order in which prediction rules are applLI Ed. When there are interacting rules, such as "XML Semantic Web" and "Semantic Web $\notin$ areas $\rightarrow$ .NET $\in$ *areas*, different rule-application orderings can produce different results. Without the first rule, a document with "XML $\in$ *languages*" but without "Semantic Web $\in$ *area*" in its initial filled template will

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016

make the second rule fire and predict ".NET $\in$ *areas* ". However, if the first rule is executed first and its prediction is confirmed, then "Semantic Web" will be extracted and the second rule can no longer fire. In DTEX, all rules with negations in their antecedent conditions are applied first. This ordering strategy attempts to maximally increase recall by making as many confirmable predictions as possible.

To summarize, documents which the user has annotated with extracted information, as well as unsupervised data which has been processed by the initial LIE system (which RAPLIER has learned from the supervised data) are all used to create a database. The rule miner then processes this database to construct a knowledge base of rules for predicting slot values. These prediction rules are then used during testing to improve the recall of the existing LIE system by proposing additional slot fillers whose presence in the document are confirmed before adding them to final extraction template.

a) *Evaluation*

To test the overall system, 600 hand-labelled computer-science job postings to the newsgroup austin.jobs were collected. 10-fold cross validation was used to generate training and test sets. In addition, 4,000 unannotated documents were collected as additional optional input to the text miner. Rules were induced for predicting the fillers of the languages, platforms, applications, and areas slots, since these are usually filled with multiple discrete-valued fillers and have obvious potential relationships between their values. Details of this experiment are described in [29].

Figure 10 shows the learning curves for recall and F-measure. Unlabeled examples are not employed in these results. In order to clearly illustrate the impact of the amount of training data for both extraction and prediction rule learning, the same set of annotated data was provided to both RAPLIER and the rule miner. The results were statistically evaluated by a two-tailed, paired *t*-test. For each training set size, each pair of systems were compared to determine if their differences in recall and were statistically significant ($P < 0.05$).

DTEX using prediction rules performs better than RAPLIER. As hypothesized, DTEX provides higher recall, and although it does decrease precision somewhat, overall F-measure is moderately increased. One interesting aspect is that DTEX retains a fixed recall advantage over RAPLIER as the size of the training set increases. This is probably due to the fact that the increased amount of data provided to the text miner also continues to improve the quality of the acquired prediction rules. Overall, these results demonstrate the role of data mining in improving the performance of LIE.

Table 2 shows results on precision, recall and F-measure when additional unlabeled documents are used to construct a larger database prior to mining for prediction rules. The 540 labelled examples used to train the extractor were always provided to the rule miner, while the number of additional unsupervised examples were varied from 0 to 4,000. The results show that the more unsupervised data supplied for building the prediction rule base, the higher the recall and the overall F-measure. Although precision does suffer, the decrease is not as large as the increase in recall.

Although adding information extracted from unlabeled documents to the database may result in a larger database and therefore more good prediction rules, it may also result in noise in the database due to extraction errors and consequently cause some inaccurate prediction rules to be discovered as well. The average F-measure without prediction rules is 86.4%, but it goes up to 88.1% when DTEX is provided with 540 labeled examples and 4,000 unlabeled examples. Unlabeled examples do not show as much power as labeled examples in producing good predic-

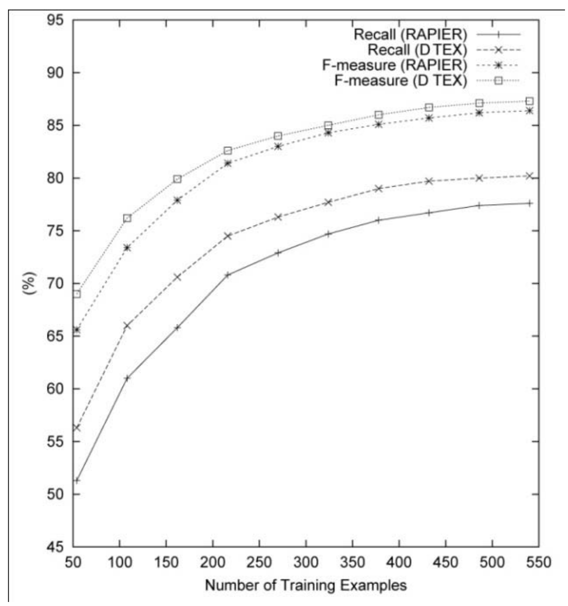


Figure 10 : Recall and F-measures on job postings

Number of Examples for Rule Mining	Precision	Recall	F-Measure
0	97.4	77.6	86.4
540(Labelled)	95.8	80.2	87.3
540+1000(Unlabeled)	94.8	81.5	87.6
540+2000(Unlabeled)	94.5	81.8	87.7
540+3000(Unlabeled)	94.2	82.4	87.9
540+4000(Unlabeled)	93.5	83.3	88.1
Matching Fillers	59.4	94.9	73.1

Table 2 : Performance results of DTEX with unlabeled examples

tion rules, because only 540 labeled examples boost recall rate and F-measure more than 4,000 unlabeled examples. However, unlabeled examples are still helpful since recall and F-measure do slowly increase as more unlabeled examples are provided.

As a baseline, in the last row of Table 2, we also show the performance of a simple method for increasing recall by always extracting substrings that are known fillers for a particular slot. Whenever a known filler string, e.g. "C#", is contained in a test document, it is extracted as a filler for the corresponding slot, e.g. language. The reason why this works poorly is that a filler string contained in a job posting is not necessarily the correct filler for the corresponding slot. For instance, "HTML" can appear in a newsgroup posting, not in the list of required skills of that particular job announcement, but in the general instructions on submitting resume´s.

V. Conclusions

In this paper, it is presented an approach that uses an automatically learned LIE system to extract a structured database from a text corpus, and then mines this database with existing KDD tools. Our preliminary experimental results demonstrate that Learned information extraction and data mining can be integrated for the mutual benefit of both tasks. LIE enables the application of KDD to unstructured text corpora and KDD can discover predictive rules useful for improving LIE performance.

Text mining is a relatively new research area at the intersection of natural-language processing, machine learning, data mining, and information retrieval. By appropriately integrating techniques from each of these disciplines, useful new methods for discovering knowledge from large text corpora can be developed. In particular, the growing interaction between computational linguistics and machine learning [8] is critical to the development of effective text-mining systems.

References Références Referencias

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Databases (VLDB-94)*, pages 487–499, Santiago, Chile, Sept. 1994.
2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information RetrLIEval*. ACM Press, New York, 1999.
3. S. Basu, R. J. Mooney, K. V. Pasupuleti, and J. Ghosh. Evaluating the novelty of text-mined rules using lexical knowledge. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, pages 233–239, San Francisco, CA, 2001.
4. M. W. Berry, editor. *Proceedings of the Third SIAM International Conference on Data Mining(SDM-2003) Workshop on Text Mining*, San Francisco, CA, May 2003.
5. M. E. Califf, editor. *Papers from the Sixteenth National Conference on Artificial Intelligence (AAAI-99) Workshop on Machine Learning for Information Extraction*, Orlando, FL, 1999. AAAI Press.
6. M. E. Califf and R. J. Mooney. Relational learning of pattern-match rules for information extraction. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 328–334, Orlando, FL, July 1999.
7. C. CardLIE. Empirical methods in information extraction. *AI Magazine*, 18(4):65–79, 1997.
8. C. CardLIE and R. J. Mooney. Machine learning and natural language (Introduction to special issue on natural language learning). *Machine Learning*, 34:5–9, 1999.
9. F. Ciravegna and N. Kushmerick, editors. *Papers from the 14th European Conference on Machine Learning(ECML-2003) and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD-2003) Workshop on Adaptive Text Extraction and Mining*, Cavtat-Dubrovnik, Croatia, Sept. 2003.
10. W. W. Cohen. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML-95)*, pages 115–123, San Francisco, CA, 1995.
11. W. W. Cohen. Learning to classify English text with ILP methods. In L. De Raedt, editor, *Advances in Inductive Logic Programming*, pages 124–143. IOS Press, Amsterdam, 1996.
12. W. W. Cohen. Improving a page classifLIEr with anchor extraction and link analysis. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1481–1488, Cambridge, MA, 2003. MIT Press.
13. DARPA, editor. *Proceedings of the Seventh Message Understanding Evaluation and Conference (MUC-98)*, Fairfax, VA, Apr. 1998. Morgan Kaufmann.
14. R. Feldman, M. Fresko, H. Hirsh, Y. Aumann, O. Liphstat, Y. Schler, and M. Rajman. Knowledge management: A text mining approach. In U. Reimer, editor, *Proceedings of Second International Conference on Practical Aspects of Knowledge Management (PAKM-98)*, pages 9.1–9.10, Basel, Switzerland, Oct. 1998.
15. D. Freitag and N. Kushmerick. Boosted wrapper induction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 577–583, Austin, TX, July 2000. AAAI Press / The MIT Press.
16. R. Ghani and A. E. Fano. Using text mining to infer semantic attirbutes for retail data mining. In *Proceedings of the 2002 LIEEE International Conference on Data Mining (ICDM-2002)*, pages 195–202, Maebash City, Japan, Dec. 2002.
17. R. Ghani, R. Jones, D. Mladenic´, K. Nigam, and S. Slattery. Data mining on symbolic knowledge

extracted from the Web. In D. Mladenic´, editor, *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining*, pages 29–36, Boston, MA, Aug. 2000.
18. M. Grobelnik, editor. *Proceedings of LIEEE International Conference on Data Mining (ICDM2001) Workshop on Text Mining (TextDM'2001)*, San Jose, CA, 2001.
19. M. Grobelnik, editor. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence(IJCAI-2003) Workshop on Text Mining and Link Analysis (TextLink-2003)*, Acapulco, Mexico, Aug. 2003.
20. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2000.
21. M. A. Hearst. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 3–10, College Park, MD, June 1999.
22. M. A. Hearst. What is text mining? http://www.sims. berkeley.edu/˜heast/text-mining.html,Oct. 2003.
23. N. Kushmerick, editor. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001) Workshop on Adaptive Text Extraction and Mining*, Seattle, WA, Aug. 2001. AAAI Press.
24. S. Loh, L. K. Wives, and J. P. M. de Oliveira. Concept-based knowledge discovery in texts extracted from the Web. *SIGKDD Explorations*, 2(1):29–39, July 2000.
25. A. McCallum and D. Jensen. A note on the unification of information extraction and data mining using conditional-probability, relational models. In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*, Acapulco, Mexico, Aug. 2003.
26. McCallum and K. Nigam. A comparison of event models for naive Bayes text classification. In *Papers from the AAAI-98 Workshop on Text Categorization*, pages 41–48, Madison, WI, July 1998.
27. D. Mladenic´, editor. *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining*, Boston, MA, Aug. 2000.
28. R. J. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the Fifth ACM Conference on Digital LibrarLI Es*, pages 195–204, San Antonio, TX, June 2000.
29. U. Y. Nahm and R. J. Mooney. A mutually beneficial integration of data mining and information extraction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 627–632, Austin, TX, July 2000.
30. U. Y. Nahm and R. J. Mooney. Using information extraction to aid the discovery of prediction rules from texts. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining*, pages 51–58, Boston, MA, Aug. 2000.
31. U. Y. Nahm and R. J. Mooney. Mining soft-matching rules from textual data. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, pages 979–984, Seattle, WA, July 2001.
32. U. Y. Nahm and R. J. Mooney. Mining soft-matching association rules. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM2002)*, pages 681–683, McLean, VA, Nov. 2002.
33. J. M. PLIErre. Mining knowledge from text collections using automatically generated metadata. In D. Karagiannis and U. Reimer, editors, *Proceedings of the Fourth International Conference on Practical Aspects of Knowledge Management (PAKM-2002)*, pages 537–548, VLIEnna, Austria, Dec. 2002. Springer. Lecture Notes in Computer Sicnece Vol. 2569.
34. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016

Global Journals Inc. (US) Guidelines Handbook 2016

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

Global Journals Incorporate (USA) is accredited by Open Association of Research Society (OARS), U.S.A and in turn, awards “FARSC” title to individuals. The 'FARSC' title is accorded to a selected professional after the approval of the Editor-in-Chief/Editorial Board Members/Dean.

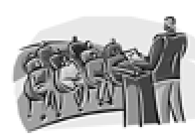
- The “FARSC” is a dignified title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

FARSC accrediting is an honor. It authenticates your research activities. After recognition as FARSC, you can add 'FARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, and Visiting Card etc.

The following benefits can be availed by you only for next three years from the date of certification:

FARSC designated members are entitled to avail a 40% discount while publishing their research papers (of a single author) with Global Journals Incorporation (USA), if the same is accepted by Editorial Board/Peer Reviewers. If you are a main author or co-author in case of multiple authors, you will be entitled to avail discount of 10%.

Once FARSC title is accorded, the Fellow is authorized to organize a symposium/seminar/conference on behalf of Global Journal Incorporation (USA).The Fellow can also participate in conference/seminar/symposium organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent.

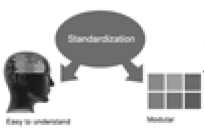


You may join as member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. In addition, it is also desirable that you should organize seminar/symposium/conference at least once.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The FARSC can go through standards of OARS. You can also play vital role if you have any suggestions so that proper amendment can take place to improve the same for the benefit of entire research community.

As FARSC, you will be given a renowned, secure and free professional email address with 100 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.



The FARSC will be eligible for a free application of standardization of their researches. Standardization of research will be subject to acceptability within stipulated norms as the next step after publishing in a journal. We shall depute a team of specialized research professionals who will render their services for elevating your researches to next higher level, which is worldwide open standardization.

The FARSC member can apply for grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A. Once you are designated as FARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria. After certification of all your credentials by OARS, they will be published on your Fellow Profile link on website https://associationofresearch.org which will be helpful to upgrade the dignity.

The FARSC members can avail the benefits of free research podcasting in Global Research Radio with their research documents. After publishing the work, (including published elsewhere worldwide with proper authorization) you can upload your research paper with your recorded voice or you can utilize chargeable services of our professional RJs to record your paper in their voice on request.

The FARSC member also entitled to get the benefits of free research podcasting of their research documents through video clips. We can also streamline your conference videos and display your slides/ online slides and online research video clips at reasonable charges, on request.

The FARSC is eligible to earn from sales proceeds of his/her researches/reference/review Books or literature, while publishing with Global Journals. The FARSC can decide whether he/she would like to publish his/her research in a closed manner. In this case, whenever readers purchase that individual research paper for reading, maximum 60% of its profit earned as royalty by Global Journals, will be credited to his/her bank account. The entire entitled amount will be credited to his/her bank account exceeding limit of minimum fixed balance. There is no minimum time limit for collection. The FARSC member can decide its price and we can help in making the right decision.

The FARSC member is eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get remuneration of 15% of author fees, taken from the author of a respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

The ' MARSC ' title is accorded to a selected professional after the approval of the Editor-in-Chief / Editorial Board Members/Dean.
The “MARSC” is a dignified ornament which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., MARSC or William Walldroff, M.S., MARSC.

MARSC accrediting is an honor. It authenticates your research activities. After becoming MARSC, you can add 'MARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, Visiting Card and Name Plate etc.

The following benefitscan be availed by you only for next three years from the date of certification.

MARSC designated members are entitled to avail a 25% discount while publishing their research papers (of a single author) in Global Journals Inc., if the same is accepted by our Editorial Board and Peer Reviewers. If you are a main author or co-author of a group of authors, you will get discount of 10%.

As MARSC, you will be given a renowned, secure and free professional email address with 30 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The MARSC member can apply for approval, grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A.

Once you are designated as MARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria.

It is mandatory to read all terms and conditions carefully.

Institutional Fellow of Open Association of Research Society (USA)-OARS (USA)

Global Journals Incorporation (USA) is accredited by Open Association of Research Society, U.S.A (OARS) and in turn, affiliates research institutions as “Institutional Fellow of Open Association of Research Society” (IFOARS).

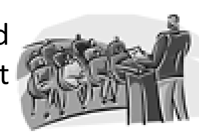
The “FARSC” is a dignified title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

The IFOARS institution is entitled to form a Board comprised of one Chairperson and three to five board members preferably from different streams. The Board will be recognized as “Institutional Board of Open Association of Research Society”-(IBOARS).

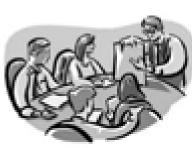
The Institute will be entitled to following benefits:

The IBOARS can initially review research papers of their institute and recommend them to publish with respective journal of Global Journals. It can also review the papers of other institutions after obtaining our consent. The second review will be done by peer reviewer of Global Journals Incorporation (USA) The Board is at liberty to appoint a peer reviewer with the approval of chairperson after consulting us.

The author fees of such paper may be waived off up to 40%.



The Global Journals Incorporation (USA) at its discretion can also refer double blind peer reviewed paper at their end to the board for the verification and to get recommendation for final stage of acceptance of publication.

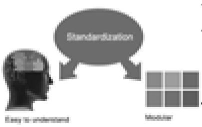


The IBOARS can organize symposium/seminar/conference in their country on behalf of Global Journals Incorporation (USA)-OARS (USA). The terms and conditions can be discussed separately.

The Board can also play vital role by exploring and giving valuable suggestions regarding the Standards of “Open Association of Research Society, U.S.A (OARS)” so that proper amendment can take place for the benefit of entire research community. We shall provide details of particular standard only on receipt of request from the Board.

The board members can also join us as Individual Fellow with 40% discount on total fees applicable to Individual Fellow. They will be entitled to avail all the benefits as declared. Please visit Individual Fellow-sub menu of GlobalJournals.org to have more relevant details.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.



After nomination of your institution as "Institutional Fellow" and constantly functioning successfully for one year, we can consider giving recognition to your institute to function as Regional/Zonal office on our behalf.
The board can also take up the additional allied activities for betterment after our consultation.

The following entitlements are applicable to individual Fellows:

Open Association of Research Society, U.S.A (OARS) By-laws states that an individual Fellow may use the designations as applicable, or the corresponding initials. The Credentials of individual Fellow and Associate designations signify that the individual has gained knowledge of the fundamental concepts. One is magnanimous and proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice.

Open Association of Research Society (US)/ Global Journals Incorporation (USA), as described in Corporate Statements, are educational, research publishing and professional membership organizations. Achieving our individual Fellow or Associate status is based mainly on meeting stated educational research requirements.

Disbursement of 40% Royalty earned through Global Journals : Researcher = 50%, Peer Reviewer = 37.50%, Institution = 12.50% E.g. Out of 40%, the 20% benefit should be passed on to researcher, 15 % benefit towards remuneration should be given to a reviewer and remaining 5% is to be retained by the institution.

We shall provide print version of 12 issues of any three journals [as per your requirement] out of our 38 journals worth $ 2376 USD.

Other:

The individual Fellow and Associate designations accredited by Open Association of Research Society (US) credentials signify guarantees following achievements:

- The professional accredited with Fellow honor, is entitled to various benefits viz. name, fame, honor, regular flow of income, secured bright future, social status etc.

- In addition to above, if one is single author, then entitled to 40% discount on publishing research paper and can get 10%discount if one is co-author or main author among group of authors.
- The Fellow can organize symposium/seminar/conference on behalf of Global Journals Incorporation (USA) and he/she can also attend the same organized by other institutes on behalf of Global Journals.
- The Fellow can become member of Editorial Board Member after completing 3yrs.
- The Fellow can earn 60% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.
- Fellow can also join as paid peer reviewer and earn 15% remuneration of author charges and can also get an opportunity to join as member of the Editorial Board of Global Journals Incorporation (USA)
- • This individual has learned the basic methods of applying those concepts and techniques to common challenging situations. This individual has further demonstrated an in–depth understanding of the application of suitable techniques to a particular area of research practice.

Note :

"

- In future, if the board feels the necessity to change any board member, the same can be done with the consent of the chairperson along with anyone board member without our approval.
- In case, the chairperson needs to be replaced then consent of 2/3rd board members are required and they are also required to jointly pass the resolution copy of which should be sent to us. In such case, it will be compulsory to obtain our approval before replacement.
- In case of “Difference of Opinion [if any]” among the Board members, our decision will be final and binding to everyone.

"

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.Online Submission: There are three ways to submit your paper:

(A) (I) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer, then Direct Submission through Homepage is also available.

(C) If these two are not convenient, and then email the paper directly to dean@globaljournals.org.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Swis 721 Lt BT.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, “Abstract” word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be three lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.
Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript’s Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals Inc. (US) welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals Inc. (US) is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global

Journals Inc. (US) are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals Inc. (US) and Editorial Board, will become the copyright of the Global Journals Inc. (US).

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals Inc. (US) follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads:

Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications.

Research letters: The letters are small and concise comments on previously published matters.

5. STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also.Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

You must strictly follow above Author Guidelines before submitting your paper or else we will not at all be responsible for any corrections in future in any of the way.

Before start writing a good quality Computer Science Research Paper, let us first understand what is Computer Science Research Paper? So, Computer Science Research Paper is the paper which is written by professionals or scientists who are associated to Computer Science and Information Technology, or doing research study in these areas. If you are novel to this field then you can consult about this field from your supervisor or guide.

TECHNIQUES FOR WRITING A GOOD QUALITY RESEARCH PAPER:

1. Choosing the topic: In most cases, the topic is searched by the interest of author but it can be also suggested by the guides. You can have several topics and then you can judge that in which topic or subject you are finding yourself most comfortable. This can be done by asking several questions to yourself, like Will I be able to carry our search in this area? Will I find all necessary recourses to accomplish the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. Pick a good study spot: To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. Think technically: Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page
- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including <u>definite statistics</u> - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.

- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.
- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.
- Do not give permission to anyone else to "PROOFREAD" your manuscript.
- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

Criterion for Grading a Research Paper (Compilation) by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

Index

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

ISSN 9754350