ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING 178

Natarajan Meghanathan
Dhinaharan Nagamalai
Nabendu Chaki (Eds.)

Advances in Computing and Information Technology

Springer

Advances in Intelligent Systems and Computing 178

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/11156

Natarajan Meghanathan, Dhinaharan Nagamalai,
and Nabendu Chaki (Eds.)

Advances in Computing and Information Technology

Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) July 13–15, 2012, Chennai, India – Volume 3

Springer

Editors
Dr. Natarajan Meghanathan
Department of Computer Science
Jackson State University
Jackson
USA

Dr. Dhinaharan Nagamalai
Wireilla Net Solutions PTY Ltd
Melbourne
VIC
Australia

Dr. Nabendu Chaki
Department of Computer Science &
Engineering
University of Calcutta
Calcutta
India

ISSN 2194-5357 e-ISSN 2194-5365
ISBN 978-3-642-31599-2 e-ISBN 978-3-642-31600-5
DOI 10.1007/978-3-642-31600-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012940793

© Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The Second International Conference on Advances in Computing and Information Technology (ACITY-2012) was held in Chennai, India, during July 13–15, 2012. ACITY attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West. The goal of this conference series is to bring together researchers and practitioners from academia and industry and share cutting-edge development in the field. The conference will provide an excellent international forum for sharing knowledge and results in theory, methodology and applications of Computer Science and Information Technology. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of Computer Science and Information Technology.

The ACITY-2012 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the conference. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer-review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. The overall acceptance rate of ACITY-2012 is less than 20%. Extended versions of selected papers from the conference will be invited for publication in several international journals. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in various research areas of Computer Science and Information Technology. In closing, ACITY-2012 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. We would like to thank the General and Program Chairs, organization staff, the members of the Technical

Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research.

It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Natarajan Meghanathan
Dhinaharan Nagamalai
Nabendu Chaki

Organization

General Chairs

David C. Wyld	Southeastern Louisiana University, USA
E.V. Krishnamurthy	Australian National University, Australia
Jae Kwang Lee	Hannam University, South Korea
Jan Zizka	SoNet/DI, FBE, Mendel University in Brno, Czech Republic
V.L. Narasimhan	Pentagram R&D Intl. Inc., New Bern, USA
Michal Wozniak	Wroclaw University of Technology, Poland

Steering Committee

Abdul Kadhir Ozcan	Karatay University, Turkey
Brajesh Kumar Kaushik	Indian Institute of Technology-Roorkee, India
Dhinaharan Nagamalai	Wireilla Net Solutions PTY LTD, Australia
Eric Renault	Institut Telecom - Telecom SudParis, Evry, France
Jacques Demerjian	Communication & Systems, France
James Henrydoss	AT&T and University of Colorado, USA
Krzysztof Walkowiak	Wroclaw University of Technology,Poland
Murugan D.	Manonmaniam Sundaranar University, India
Nabendu Chaki	University of Calcutta, India
Natarajan Meghanathan	Jackson State University, USA
Raja Kumar M.	Taylor's University, Malaysia
Salah Al-Majeed	University of Essex, UK
Selma Boumerdassi	Conservatoire National des Arts Et Metiers (CNAM), France
Sundarapandian Vaidyanathan	VelTech Dr. RR & Dr. SR Technical University, India

Program Committee Members

A.H.T. Mohammad	University of Bradford, UK
A.P. Sathish Kumar	PSG Institute of Advanced Studies, India
AAA. Atayero	Covenant University, Nigeria
Abdul Aziz	University of Central Punjab, Pakistan
Abdul Kadhir Ozcan	Karatay University, Turkey
Abdul Kadir Ozcan	The American University, Cyprus
Abdulbaset Mohammad	University of Bradford, United Kingdom
Ahmad Saad Al-Mogren	King Saud University, Saudi Arabia
Ahmed M. Khedr	Sharjah University, Sharjah, UAE
Ahmed Nada	Al-Quds University, Palestinian
Ajay K. Sharma	Dr. B R Ambedkar National Institute of Technology, India
Alaa Ismail Elnashar	Taif University, KSA
Alejandro Garces	Jaume I University, Spain
Alejandro Regalado Mendez	Universidad del Mar - México, USA
Alfio Lombardo	University of Catania, Italy
Ali El-Rashedy	University of Bridgeport, CT, USA
Ali M.	University of Bradford, United Kingdom
Ali Maqousi	Petra University, Jordan
Alireza Mahini	Islamic Azad University-Gorgan, Iran
Alvin Lim	Auburn University, USA
Amandeep Singh Thethi	Guru Nanak Dev University Amritsar, India
Amit Choudhary	Maharaja Surajmal Institute,India
Anand Sharma	MITS-Rajasthan, India
Anjan K.	RVCE-Bangalore, India
Ankit Thakkar	Nirma University, India
Ankit	BITS, PILANI India
Anthony Atayero	Covenant University, Nigeria
Aravind P.A.	Amrita School of Engineering India
Arun Pujari	Sambalpur University, India
Arunita Jaekel	University of Windsor, Canada
Ashok Kumar Das	IIT Hyderabad, India
Ashok kumar Sharma	YMCA Institute of Engineering, India
Ashutosh Dubey	NRI Institute of Science & Technology, Bhopal
Ashutosh Gupta	MJP Rohilkhand University, Bareilly
Athanasios Vasilakos	University of Western Macedonia, Greece
Azween Bin Abdullah	Universiti Teknologi Petronas, Malaysia
B. Srinivasan	Monash University, Australia
Babak Khosravifar	Concordia University, Canada
Balakannan S.P.	Chonbuk Nat. Univ., Jeonju
Balasubramanian K.	Lefke European University, Cyprus
Balasubramanian Karuppiah	Dr. MGR University, India
Bari A.	University of Western Ontario, Canada

Beatrice Cynthia Dhinakaran	TCIS, South Korea
Bela Genge	European Commission Joint Research Centre, Belgium
Bharat Bhushan Agarwal	I.F.T.M University, India
Bhupendra Suman	IIT Roorkee , India
Biju Pattnaik	University of Technology, India
Bikash singh	Islamic University-Kushtia, Bangladesh
Binod Kumar Pattanayak	Siksha O Anusandhan University, India
Bobby Barua	Ahsanullah University of Science and Technology, Bangladesh
Bong-Han	Kim, Chongju University, South Korea
Boo-Hyung Lee	KongJu National University, South Korea
Brajesh Kumar Kaushik	Indian Institute of Technology, India
Buket Barkana	University of Bridgeport, USA
Carlos E. Otero	University of South Florida Polytechnic, USA
Charalampos Z. Patrikakis	National Technical University of Athens, Greece
Chin-Chih Chang	Chung Hua University ,Taiwan
Cho Han Jin	Far East University, South Korea
Choudhari	Bhagwati Chaturvedi College of Engineering, India
Christos Politis	Kingston University, UK
Cristina Ribeiro	University of Waterloo, Canada
Cristina Serban	Ovidius University of Constantza, Romania
Danda B. Rawat	Old Dominion University, USA
David C. Wyld	Southeastern Louisiana University, USA
Debasis Giri	Haldia Institute of Technology, India
Debdatta Kandar	Sikkim Manipal University, India
Dhinaharan Nagamalai	Wirella Net Solutions PTY Ltd, Australia
Diego Reforgiato	University of Catania, Italy
Dimitris Kotzinos	Technical Educational Institution of Serres, Greece
Doreswamyh hosahalli	Mangalore University, India
Durga Toshniwal	Indian Institute of Technology, India
E. Martin	University of California, Berkeley, USA
E.V. Krishnamurthy	ANU College of Engg & Computer Science, Austraila
Emmanuel Bouix	iKlax Media, France
Eric Renault	Institut Telecom - Telecom SudParis, Evry, France
Ermatita Zuhairi	Sriwijaya University, Indonesia
Farag M. Sallabi	United Arab Emirates University, UAE
Farshad Safaei	Shahid Beheshti University, Iran
Ford Lumban Gaol	University of Indonesia
Genge Bela	Joint Research Centre, European Commission, Italy
Ghalem Belalem	University of Oran, Algeria
Giovanni Cordeiro Barroso	Universidade Federal do Ceara, Brasil
Giovanni Schembra	University of Catania, Italy
Girija Chetty	University of Canberra, Australia

Gomathi Kandasamy	Avinashilingam Deemed University for Women, India
Gopalakrishnan Kaliaperumal	Anna University, Chennai
Govardhan A.	JNTUH College of Engineering, India
Guo Bin	Institute TELECOM SudParis, France
H.V. Ramakrishnan	Dr. MGR University, India
Haider M. Alsabbagh	Basra University, Iraq
Haller Piroska	Petru Maior University-Tirgu Mures, Romania
Hao Shi	Victoria University, Australia
Hao-En Chueh	yuanpei University, Taiwan
Hari Chavan	National Institute of Technology, Jamshedpur, India
Henrique J.A. Holanda	UERN - Universidade do Estado do Rio Grande do Norte, Brasil
Henrique Joao Lopes Domingos	University of Lisbon, Portugal
Hiroyuki Hisamatsu	Osaka Electro-Communication University, Japan
Ho Dac Tu	Waseda University, Japan
Homam Reda El-Taj	Universiti Sains Malaysia, Malaysia
Hong yu	Capitol College, USA
Huosheng Hu	University of Essex, UK
Hussein Al-Bahadili	Petra University, Jordan
Hussein Ismail Khalaf Al-Bahadili	Petra University, Jordan
Hwangjun Song	Pohang University of Science and Technology,South Korea
Ignacio Gonzalez Alonso	University of Oviedo, Europe
Indrajit Bhattacharya	Kalyani Govt. Engg. College, India
Intisar Al-Mejibli	University of Essex, UK
Ioannis Karamitsos	Itokk Communications, Canada
J.K. Mandal	University of Kalyani, India
Jacques Demerjian	Communications & Systems, France
Jae Kwang Lee	Hannam University, South Korea
Jalel Akaichi	University of Tunis, Tunisia
Jan Zizka	SoNet/DI, FBE, Mendel University in Brno, Czech Republic
Jeong-Hyun Park	Electronics Telecommunication Research Institute, South Korea
Jeyanthy N.	VIT University, India
Jifeng Wang	University of Illinois at Urbana Champaign, USA
Johann Groschdl	University of Bristol, UK
Jose Enrique Armendariz-Inigo	Universidad Publica de Navarra, Spain
Juan Li	North Dakota State University, USA
Jyoti Singhai	Electronics and Communication Deptt-MANIT, India
Jyotirmay Gadewadikar	Alcorn State University, USA
Kai Xu	University of Bradford, United Kingdom
Kamalrulnizam Abu Bakar	Universiti Teknologi Malaysia, Malaysia

Karim Konate	University Cheikh Anta DIOP, Dakar
Kaushik Chakraborty	Jadavpur University, India
Kayhan Erciyes	Izmir University, Turkey
Khaled Shuaib	United Arab Emirates University, UAE
Khamish Malhotra	University of Glamorgan, UK
Khoa N. Le	University of Western Sydney, Australia
Krishnamurthy E.V.	ANU College of Engg & Computer Science, Australia
Krzysztof Walkowiak	Wroclaw University of Technology, Poland
Kuribayashi	Seikei University, Japan
L. Nirmala Devi	Osmania University - Hyderabad, India
Laiali Almazaydeh	University of Bridgeport, USA
Lu Yan	University of Hertfordshire, UK
Lus Veiga	Technical University of Lisbon, Portugal
Lylia Abrouk	University of Burgundy, France
M. Aqeel Iqbal	FUIEMS, Pakistan
M. Rajarajan	City University, UK
M. Ali	University of Bradford, UK
Maode Ma	Nanyang Technological University, Singapore
Marco Folli	University of Pavia, Italy
Marco Roccetti	Universty of Bologna, Italy
Massimo Esposito	ICAR-CNR, Italy
Md. Sipon Miah	Islamic University-Kushtia, Bangladesh
Michal Wozniak	Wroclaw University of Technology, Poland
Michel Owayjan	American University of Science & Technology, Lebanon
Miguel A. Wister	Juarez Autonomous University of Tabasco, Mexico
Mohamed Hassan	American University of Sharjah, UAE
Mohammad Ali Jabreil Jamali	Islamic Azad University, Iran
Mohammad Hadi Zahedi	Ferdowsi University of Mashhad, Iran
Mohammad Hajjar	Lebanese University, Lebanon
Mohammad Kaghazgaran	Islamic Azad University, Iran
Mohammad Mehdi Farhangia	Universiti Teknologi Malaysia, Malaysian
Mohammad Momani	University of technology Sydney, Australia
Mohammad Talib	University of Botswana, Botswana
Mohammad Zaidul Karim	Daffodil International University, Bangladesh
Mohammed Feham	University of Tlemcen, Algeria
Mohammed M. Alkhawlani	University of Science and Technology, Yemen
Mohsen Sharifi	Iran University of Science and Technology, Iran
Muhammad Sajjadur Rahim	University of Rajshahi, Bangladesh
Murty	Ch A S, JNTU, Hyderabad
Murugan D.	Manonmaniam Sundaranar University, India
Mydhili Nair	M S Ramaiah Institute of Technology, India
N. Krishnan	Manonmaniam Sundaranar University, India
Nabendu Chaki	University of Calcutta, India

Nadine Akkari	King abdulaziz University, Saudi Arabia
Naohiro Ishii	Aichi Institute of Technology, Japan
Nasrollah M. Charkari	Tarbiat Modares University, Iran
Natarajan Meghanathan	Jackson State University, USA
Nicolas Sklavos	Technological Educational Institute of Patras, Greece
Nidaa Abdual Muhsin Abbas	University of Babylon, Iraq
Nour Eldin Elmadany	Arab Acadmy for Science and Technology, Egypt
Ognjen Kuljaca	Alcorn State University, USA
Olakanmi Oladayo	University of Ibadan, Nigeria
Omar Almomani	Universiti Utara Malaysia, Malaysia
Orhan Dagdeviren	Izmir University, Turkey
Osman B. Ghazali	Universiti Utara Malaysia, Malaysia
Othon Marcelo Nunes Batista	Universidade Salvador, Brazil
Padmalochan Bera	Indian Institute of Technology, Kharagpur, India
Partha Pratim Bhattacharya	Mody Institute of Technology & Science, India
Patricia Marcu	Leibniz Supercomputing Centre, Germany
Patrick Seeling	University of Wisconsin - Stevens Point, USA
R. Thandeeswaran	VIT University, India
Phan Cong Vinh	London South Bank University, UK
Pinaki Sarkar	Jadavpur University, India
Polgar Zsolt Alfred	Technical University of Cluj Napoca, Romania
Ponpit Wongthongtham	Curtin University of Technology, Australia
Quan (Alex) Yuan	University of Wisconsin-Stevens Point, USA
Rafael Timoteo	University of Brasilia - UnB, Brazil
Raied Salman	Virginia Commonwealth University, USA
Rajendra Akerkar	Technomathematics Research Foundation, India
Rajeswari Balasubramaniam	Dr. MGR University, India
Rajkumar Kannan	Bishop Heber College, India
Rakhesh Singh Kshetrimayum	Indian Institute of Technology, Guwahati, India
Raman Maini	Punjabi University, India
Ramayah Thurasamy	Universiti Sains Malaysia, Malaysia
Ramayah	Universiti Sains Malaysia, Malaysia
Ramin karimi	University Technology Malaysia
Razvan Deaconescu	University Politehnica of Bucharest, Romania
Reena Dadhich	Govt. Engineering College Ajmer
Reshmi Maulik	University of Calcutta, India
Reza Ebrahimi Atani	University of Guilan, Iran
Rituparna Chaki	West Bengal University of Technology, India
Robert C. Hsu	Chung Hua University, Taiwan
Roberts Masillamani	Hindustan University, India
Rohitha Goonatilake	Texas A&M International University, USA
Rushed Kanawati	LIPN - Universite Paris 13, France
S. Geetha	Anna University - Tiruchirappalli, India
S. Hariharan	B.S. Abdur Rahman University, India

S. Venkatesan	University of Texas at Dallas - Richardson, USA
S.A.V. Satyamurty	Indira Gandhi Centre for Atomic Research, India
S. Arivazhagan	Mepco Schlenk Engineering College, India
S. Li	Swansea University, UK
S. Senthil Kumar	Universiti Sains Malaysia, Malaysia
Sajid Hussain	Acadia University, Canada
Salah M. Saleh Al-Majeed	University of Essex, United Kingdom
Saleena Ameen	B.S.Abdur Rahman University, India
Salem Nasri	ENIM, Monastir University, Tunisia
Salim Lahmiri	University of Qubec at Montreal, Canada
Salini P.	Pondichery Engineering College, India
Salman Abdul Moiz	Centre for Development of Advanced Computing, India
Samarendra Nath Sur	Sikkim Manipal University, India
Sami Ouali	ENSI, Compus of Manouba, Manouba, Tunisia
Samiran Chattopadhyay	Jadavpur University, India
Samodar reddy	India school of mines , India
Samuel Falaki	Federal University of Technology-Akure, Nigeria
Sanjay Singh	Manipal Institute of Technology, India
Sara Najafzadeh	University Technology Malaysia
Sarada Prasad Dakua	IIT-Bombay, India
Sarmistha Neogy	Jadavpur University, India
Satish Mittal	Punjabi University, India
S.C. SHARMA	IIT - Roorkee, India
Seetha Maddala	CBIT, Hyderabad
Selma Boumerdassi	Cnam/Cedric, France
Sergio Ilarri	University of Zaragoza, Spain
Serguei A. Mokhov	Concordia University, Canada
Shaoen Wu	The University of Southern Mississippi, USA
Sharvani G.S.	RV College of Engineering, Inida
Sherif S. Rashad	Morehead State University, USA
Shin-ichi Kuribayashi	Seikei University, Japan
Shivan Haran	Arizona state University, USA
Shobha Shankar	Vidya vardhaka College of Engineering, India
Shrikant K. Bodhe	Bosh Technologies, India
Shriram Vasudevan	VIT University, India
Shrirang Ambaji Kulkarni	National Institute of Engineering, India
Shubhamoy Dey	Indian Institute of Management Indore, India
Solange Rito Lima	University of Minho, Portugal
Souad Zid	National Engineering School of Tunis, Tunisia
Soumyabrata Saha	Guru Tegh Bahadur Institute of Technology, India
Sridharan	CEG Campus - Anna University, India
Sriman Narayana Iyengar	VIT University, India
Srinivasulu Pamidi	V R Siddhartha Engineering College Vijayawada, India

Sriram Maturi	Osmania University, India
Subhabrata Mukherjee	Jadavpur University, India
Subir Sarkar	Jadavpur University, India
Sundarapandian Vaidyanathan	VelTech Dr. RR & Dr. SR Technical University, India
Sunil Singh	Bharati vidyapeeth's College of Engineering, India
Sunilkumar S. Manvi	REVA Institute of Technology and Management Kattigenhalli, India
SunYoung Han	Konkuk University, South Korea
Susana Sargento	University of Aveiro, Portugal
Swarup Mitra	Jadavpur University, Kolkata, India
T. Ambaji Venkat Narayana Rao	Hyderabad Institution of Technology and Management , India
T.G. Basavaraju	National Institute of Technology Karnataka (NITK), India
Thomas Yang	Embry Riddle Aeronautical University, USA
Tri Kurniawan Wijaya	Technische Universitat Dresden, Germany
Tsung Teng Chen	National Taipei Univ., Taiwan
Utpal Biswas	University of Kalyani, India
V.M. Pandharipande	Dr. Babasaheb Ambedkar Marathwada University, India
Valli Kumari Vatsavayi	AU College of Engineering, India
Vijayalakshmi S.	VIT University, India
Virgil Dobrota	Technical University of Cluj-Napoca, Romania
Vishal Sharma	Metanoia Inc., USA
Wei Jie	University of Manchester, UK
Wichian Sittiprapaporn	Mahasarakham University, Thailand
Wided Oueslati	l'institut Superieur de Gestion de Tunis, Tunisia
William R. Simpson	Institute for Defense Analyses, USA
Wojciech Mazurczyk	Warsaw University of Technology, Poland
Xiaohong Yuan	North Carolina A & T State University, USA
Xin Bai	The City University of New York, USA
Yahya Slimani	Faculty of Sciences of Tunis, Tunisia
Yannick Le Moullec	Aalborg University, Denmark
Yaser M. Khamayseh	Jordan University of Science and Technology, Jordan
Yedehalli Kumara Swamy	Dayanand Sagar College of Engineering, India
Yeong Deok Kim	Woosong University, South Korea
Yogeshwar Kosta	Marwadi Education Foundations Group of Institutions, India
Yuh-Shyan Chen	National Taipei University, Taiwan
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Zaier Aida	National Engeneering School of GABES, Tunisia
Zakaria Moudam	Université sidi mohammed ben Abdellah, Morocco
Zuqing Zhu	Cisco Systems, USA

External Reviewers

A. Kannan	K.L.N. College of Engineering, India
Martin	Sri Manakula Vinayagar Engineering College, India
Abhishek Samanta	Jadavpur University, Kolkata, India
Ayman Khalil	Institute of Electronics and Telecommunications of Rennes, France
Cauvery Giri	RVCE, India
Ch. V. Rama Rao	Gudlavalleru Engineering College, India
Chandra Mohan	Bapatla Engineering College, India
E.P. Ephzibah	VIT University-Vellore, India
Hameem Shanavas	Vivekananda Institute of Technolgy, India
Kota Sunitha	G. Narayanamma Institute of Technology and Science, Hyderabad
Kunjal B. Mankad	ISTAR, Gujarat, India
Lakshmi Rajamani	Osmania University, India
Lavanya	Blekinge Institute of Technology, Sweden
M.P. Singh	National Institute of Technology, Patna
M. Tariq Banday	University of Kashmir, India
M.M.A. Hashem	Khulna University of Engineering and Technology, Bangladesh
Mahalinga V. Mandi	Dr. Ambedkar Institute of Technology, Bangalore, Karnataka, India
Mahesh Goyani	G H Patel College of Engineering and Technology, India
Maragathavalli P.	Pondicherry Engineering College, India
M.P. Singh	National Institute of Technology, Patna
M. Tariq Banday	University of Kashmir, India
M.M.A. Hashem	Khulna University of Engineering and Technology, Bangladesh
Mahalinga V. Mandi	Dr. Ambedkar Institute of Technology, India
Monika Verma	Punjab Technical University, India
Moses Ekpenyong	University of Uyo, Nigeria
Mini Patel	Malwa Institute of Technology, India
N. Kaliammal	NPR College of Engg &Tech, India
N. Adhikari	Biju Pattnaik University of Technology, India
N.K. Choudhari	Bhagwati Chaturvedi College of Engineering, India
Naga Prasad Bandaru	PVP Siddartha Institute of Technology, India
Nagamanjula Prasad	Padmasri Institute of Technology, India
Nagaraj Aitha	I.T, Kamala Institute of Tech & Science, India
Nana Patil	NIT Surat, Gujrat
Nitiket N. Mhala	B.D. College of Engineering - Sewagram, India
P. Ashok Babu	Narsimhareddy Engineering College, India
P. Sheik Abdul Khader	B.S. Abdur Rahman University, India

Pesn Krishna Prasad	Aditya Engineering College, India
Partha Pratim Bhattacharya	Mody Institute of Technology & Science, India
Pappa Rajan	Anna University, India
Pooja Jain	JUIT Waknaghat, India
Prabu Dorairaj	NetApp Inc, India
Pradeepini Gera	Jawaharlal Nehru Technological University, India
Pravin P. Karde	HVPM's College of Engineering & Technology - Amravati, India
Premanand K. Kadbe	Vidya Pratishthan's College of Engineering, India
Priti Sajja	Sardar Patel University, India
R. Baskaran	Anna University - Chennai, India
R. Murali	Dr. Ambedkar Institute of Technology, Bangalore
R.M. Suresh	Mysore University
Rabindranath Bera	Sikkim Manipal Inst. of Technol., India
Rajashree Biradar	Ballari Institute of Technology and Management, India
Rajesh Kumar Krishnan	Bannari Amman Inst. of Technol., India
Rajesh Kumar P.	The Best International, Australia
Rajeshwari Hegde	BMS College of Engineering, India
Rajagopal Palsonkennedy	Dr. MGR University, India
Selvakumar Ramachandran	Blekinge Institute of Technology, Sweden

Contents

Artificial Intelligence, Soft Computing and Applications

Web and Semantic Technology

Ad Hoc, Sensor, Ubiquitous Computing and VLSI Design

Soft Computing Approach for Modeling Genetic Regulatory Networks

Khalid Raza and Rafat Parveen

Department of Computer Science, Jamia Millia Islamia (Central University),
New Delhi, India
kraza@jmi.ac.in, rafatparveen@yahoo.co.in

Abstract. Interactions among the cellular components determine the behaviour of the complex biological system. The major challenge of the post-genomic era is to understand how interactions among various molecules in a cell determine its form and function. Several computational techniques for modeling biological systems, particularly gene regulatory networks (GRNs), has been proposed in order to understand the complex biological interactions and behaviours. Gene regulatory models has been proved to be the most widely used mechanism to model, analyze and predict the behaviour of an organism. In this paper, we have reviewed the role of soft computing techniques, such as fuzzy logic, artificial neural networks, evolutionary algorithms and their hybridization, for modeling GRNs. In addition, recent developments in this area are introduced and various challenges and opportunities for future research are discussed.

1 Introduction

Networks play an important role in biological investigations and used to represent processes in biological systems. It captures the interactions and dependencies between molecular biological entities such as genes, transcripts, proteins and metabolites [22]. Systems biology is rapidly growing research area which aims at the system level understanding of biological systems [1]. Systems biology is one of the large application areas for network-centred analysis and visualization of biological entities. With the availability of complete genome sequences and high-throughput post-genomics experimental data, last decade have witnessed a viable interest in the study of networks of macromolecular interactions such as gene regulatory networks, metabolic networks, protein-protein interaction networks, or signal transduction networks. Today computational modeling of biological systems has become rather essential in order to understand the complex biological interactions and behaviour. Many theoretical models have been proposed to model, analyze and infer complex regulatory interactions and provide hypothesis for experimental verification.

A genetic regulatory network (GRN) is a network depicting interactions between genes and model causal relationship between gene activities. A GRN denotes the assembly of regulatory effects and gene interactions in a biological system. The GRN helps us understand the intricate interactions of multiple genes under various stimuli or environmental conditions [3]. Modeling GRNs enables us to decipher the gene interaction mechanism for a particular stimulation and further we can utilize this

N. Meghanathan et al. (Eds.): Advances in Computing & Inf. Technology, AISC 178, pp. 1–11.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

information to predict adverse effects of new drugs or to determine a new drug target [20]. Due to improved understanding of gene regulation processes modeling efforts are increasingly being used for generating the hypotheses that are then tested with experimental data. Generally, the process of GRNs modeling consists of a few main steps: (i) selection of an appropriate model (ii) inferring parameters from data (iii) validating the model and (iv) conducting simulation of GRNs, to predict its behaviour under various conditions [48]. Hence, there is a need for efficient computational tools for the qualitative modeling of GRNs so as to understand the experimental data in the context of the dynamical behavior of a cell and generates hypotheses with the assistance of computational tools [4, 5].

Some review papers on GRNs modeling exists in the literature [1, 2, 18, 19, 21, 48], but we have approached in a different way. We have done survey of soft computing based techniques for modeling GRNs. In addition, recent developments and future challenges in the area are discussed.

2 Basic Modeling Techniques

There are several techniques for modeling GRNs including Directed graph, Petri nets [16, 17], Boolean networks [6–8, 17], generalized Bayesian networks [9, 10], linear and non-linear ordinary differential equations (ODEs) [11–15], machine learning approach, etc. *Directed graph* is a straightforward and most simple way to model a GRN, where vertices represent genes and edges interactions among the genes. A directed edge is defined as a tuple (*i*, *j*, *s*), where *i* denotes the head, *j* the tail of the edge and *s* is equal to either + or – indicating whether *i* is activated or inhibited by *j*. The graphical representations of GRNs permit a number of operations that can be carried out to make prediction about biological processes [1]. Petri nets are an extension of graph models that represents a well-established technique for modeling regulatory systems. *Petri net* is a non-deterministic method which has successfully been applied for simulating GRN, allowing simple quantitative representation of dynamic processes. The limitation of Petri nets model is that it does not support hierarchical structuring, which makes them difficult to be use for large-scale networks. *Boolean networks* are deterministic method based on logical functions. The Boolean method assumes the expression level of each gene is either expressed (ON) of not expressed (OFF). In the network, each node's logical function is determined by finding the minimum set of nodes whose expression level can explain the observed changes in the state of a given node. The advantages of Boolean methods are its simplicity and finite state space. Boolean methods are also more computationally tractable. The algorithm, REVEAL (reverse engineering algorithm) [17] was first step towards modeling large-scale network using Boolean network. However, these models ignore the effect of genes at intermediate levels and impractically assume that transitions between states are synchronous.

Bayesian networks (BNs) uses a graphical representations of multivariate joint probability distribution, having two parts, a directed acyclic graph and a set of local joint probability distributions. These models can deal with the stochastic aspects of gene regulation and able to handle noisy and incomplete data which is prevalent in microarray technology. However, these models can not deal with dynamic aspects of gene regulation. Dynamic Bayesian networks have been formulated to overcome the problem of dynamicity. *Ordinary differential equations* (ODEs) formalism have been mostly used method for modeling dynamic biochemical networks, particularly, GRNs. The ODEs approach is able to capture detailed information about the network's dynamics but it needs high-quality data on kinetic parameters and hence it is currently appropriate for a few systems only. A detailed discussion about various differential equation-based approaches can be found in [1] and [19].

3 Soft Computing Techniques

Prof. L. A. Zadeh coined the term "soft computing" (SC) in 1992 which is an evolving collection of methodologies, that aims to exploit tolerance for imprecision, uncertainty, and partial truth to achieve robustness, tractability, and low cost. Fuzzy logic (FL), neural networks (NN), and evolutionary computation (EC) are the core methodologies of SC. Each of these methodologies has their own strength, for example, FL is capable of representing knowledge via fuzzy rules, ANNs can be used for learning and adaptation and EAs for the optimization. However, FL, NN, and EC should not be viewed as rival of each other rather synergistic and complementary instead. Soft computing is causing a breakthrough in engineering and science fields since it can solve problems that have not been able to be solved by traditional hard-computing methods [25]. In Zadeh's own words, *"Soft computing is an emerging approach to computing which parallel the remarkable ability of the human mind to reason and learn in an environment of uncertainty and imprecision"* [23].

4 Role of Soft Computing in GRN Modeling

Soft computing is gradually opening up several opportunities in bioinformatics, especially by generating low-cost, low-precision (approximate) and good solutions. It provides us efficient solutions to the various challenging problems from bioinformatics such as protein structure prediction, microarray data analysis, gene sequence analysis, modeling genetic and biochemical networks [24]. Soft computing techniques, particularly, FL, ANNs, EAs and their hybridization have been successfully used for modeling GRNs.

Fuzzy Logic

The biological systems behave in a fuzzy manner. FL provides a mathematical framework for modeling and describing biological systems. Literature reports that FL has been successfully used for modeling GRNs due to its capability to represent non-linear systems, its friendly language to incorporate and edit domain knowledge in the form of fuzzy rules. Woolf and Wang [28] proposed a novel algorithm for analysing gene expression data using FL. The model was designed to find triplets (activators, repressors, targets) in yeast gene expression data set. The model was implemented using C-language and executed on an 8-processor SGI Origin 2000 system, which took ~200 hours to analyse the relationships between 1,898 genes. Later, Ressom, *et. al.* [39] has extended and improved the work of Woolf and Wang [28] in terms of reducing computation time and generalizing the gene regulatory model to accommodate co-activator and co-repressors. Reduction in computation time is achieved by using clustering as a pre-processing step. The improved algorithm achieves a reduction of 50% computation time. Later on R. Ram, *et.al.* [33] has also improved the fuzzy logic model developed by Woolf and Wang [28] to predict changes in expression values and infer causal relationship between genes. They have improved the searching activator/repressor regulatory relationship between gene triplets in the microarray data. A pre-processing technique for the fuzzy model has also been proposed to remove redundant computations due to presence of similar expression profiles in the microarray data. The pre-processing technique groups the genes based on similarity in their expression profile variations and yeast expression data has been used to test the model but the limitation is that interactions extracted from the microarray data are not necessarily causative but are likely to be associated in a similar biological pathway.

Pan Du, *et.al.* [32] has applied fuzzy weights for modeling the interactions between genes in a GRN. The interaction in the network is modelled as fuzzy function that depends on the detail known about the network. The analysis and creation of GRNs involves first clustering of data using multi-scale fuzzy k-means clustering and then searching for weighted time correlation between the cluster centre time profiles. The link validity and strength is then evaluated using fuzzy metric based on evidence strength and co-occurrence of similar gene function within a cluster. Experimental results on the carbohydrate metabolism of the model plant *Arabidopsis thaliana* have been illustrated. GO database has been used to evaluate gene regulatory relationships from a biological viewpoint.

Y. Sun, *et.al.* [3] has applied dynamic fuzzy modeling approach by incorporating structural knowledge to model GRNs. This technique infers information on gene interactions in the form of fuzzy rules and considers the dynamic aspects of gene regulation. It is able to reveal more biological relationships among genes and their products. It has used two sets of data to validate the models, synthetic data from a numerical example and real *SOS DNA repair network* data with structural knowledge. The distinguishing feature of this model is that (a) prior structural knowledge on GRN can be incorporated for the purpose of faster convergence of the identification process and (b) non-linear dynamic property of the GRN can be well captured for the better prediction.

Artificial Neural Networks

An artificial neural network (ANN) is a computational model that is inspired by the structural and functional aspects of biological nervous systems. The capabilities of ANNs to learn from the data, approximate any multivariate nonlinear function and its robustness to noisy data make ANN a suitable candidate for modeling gene regulatory interactions from gene expression data. Several types of ANNs have been successfully applied for modeling gene regulatory interactions including perceptrons [40–42], self-organizing maps (SOM) [43, 44] and recurrent neural networks (RNNs) [30, 37].

Ed. Keedwell, *et.al.* [43] has successfully applied ANN in the purest sense for the reconstruction of GRNs from microarray data. The design of the neural network was quite simple when dealing with Boolean networks and standard feed-forward backpropagation method has been applied. The modelled ANN was tested under various conditions and found that resulting networks were able to encode complex relationship between genes. Vohradsky [27] has also proposed an ANN based model assuming that the regulation effect on the gene expression of a particular gene can be expressed as a neural network. Each node in the network represents a particular gene and the wiring between the nodes represents regulatory interactions. Here each layer of the network represents the level of gene expression at time *t* and output of a node at time $t+\Delta t$ can be derived from the expression levels. The regulatory effect is transformed using a sigmoidal transfer function to the interval (0, 1). The main advantage of this model is that it is continuous, uses a transfer function to transform the inputs to a shape close to those observed in natural processes and does not use artificial elements. The drawback is that it consists of large number of parameters that must be computed from experimental data.

Stochastic neural network model in the framework of a coarse-grained approach was proposed by Tiam and Burrage [30] for better description of the GRNs. The model is able to represent both intermediate regulation as well as chance events in gene expression. Poisson random variables are applied to represent chance events. X. Hu *et.al.* [45] has proposed a general recurrent neural network (RNN) model for the reverse-engineering of GRNs and to learn their parameters. RNN has been deployed due to its capability to deal with complex temporal behaviour of genetic networks. The model was tested on *SOS DNA Repair* network of the *e.coli*. The model was able to discover complex regulatory relationships among genes in the SOS network.

Evolutionary Algorithms

Evolutionary algorithms (EAs) are basically optimization algorithm based on Darwin’s theory of evolution. It is basically a search algorithm that is modeled on the mechanics of natural selection and survival for the fittest. It combines survival of the fittest among individuals with a structured yet randomized information exchange to form a search algorithm. In EAs optimization techniques searching from a population are done from a single point and for each iteration a competitive selection is done. The solutions with high “fitness” are recombined with other solutions. The solutions are then “mutated” by making a small change to a single element of the solution. The main purpose of recombination and mutation is to generate new solutions but it is biased towards regions of the space for which good solutions have already been identified. Generally, three evolutionary techniques are distinguished: genetic programming (GP), genetic algorithms (GA) and evolutionary programming (EP). The GP focuses on programs evolution, GA on optimizing general combinatorial problems and EP focuses on optimizing continuous functions without recombination. EAs belong to probabilistic algorithms and they differ from random algorithms in that they combine elements of directed and stochastic search. Due to this reason EAs are more robust than directed search methods. Another merit of EAs is that they maintain a population of potential solutions while other search techniques process a single point of the search space. The limitation of GP and GA-based modeling techniques are that they do not take care of the noise effect which is prevalent in microarray data.

Various constituents of EAs have been successfully applied for modeling GRNs. A combination of GP and Least Mean Square (LMS) method, called LMS-GP, has been applied by Ando *et.al.* [46] to identify a concise form of regulation between genes from time series data. LMS is applied to determine the coefficients of the GPs, which decreases the Mean Squared Error (MSE) between the observed and model time series without complicating the GPs. This model has been tested on artificial as well as real-world data. The proposed LMS-GP model has average MSE of 4.21×10^{-3} over 10 runs, while standard GP averaged MSE is 6.704×10^{-3} over 10 runs. Wang *et.al.* [47] has proposed a joint GP and Kalman filtering (KF) approach to infer GRNs from time series data. Here nonlinear differential equation model is adopted and an iterative algorithm has been proposed to identify the model, where GP is employed to identify the structure of the model and KF is deployed to estimate the parameters in each iteration. The proposed model has been tested using synthetic as well as time-series gene-expression data of yeast protein synthesis. Due to noise in microarray data, the KF may not be appropriate for estimating parameters.

Noman and Iba [50] have applied decoupled S-system formalism for the inference of effective kinetic parameters from time series data and employed Trigonometric Differential Evolution (TDE) as the optimization engine for capturing the dynamics of gene expression data. The fitness function used here is a modified version of Kimaru *et.al.* [51] for reducing the number of false positive predictions. The spare network structure has been identified with the help of hill-climbing local search (HCLS) method within the framework of proposed EA. Experiments on well studied small scale artificial network in noise-free as well as noisy environment is done. The proposed model successfully identifies the network structure and its parameter values. Real-life data has also been used for reconstructing the *SOS DNA repair network* of *e.coli*. The proposed model correctly identified the regulations of gene *lexA* and some other known regulations. Chowdhury and Chetty [52] extended the work of Noman et.al. [50]. In this model, GA is applied for scoring the networks' several useful features for accurate inference of network, such as a Prediction Initialization (PI) algorithm to initialize the individuals, a Flip Operation (FO) for matching the values, and a restricted execution of HCLS over few individuals. A refinement algorithm for optimizing sensitivity and specificity of inferred networks was also proposed.

Hybridized Techniques

Each of the soft computing (SC) constituents has their own advantages. The learning and adaptation capability of ANN, knowledge representation via fuzzy rules through FLs and optimization capability of GAs when joined together, one can exploit the advantages of each in the hybridized model. The most common form of hybridizations are ANN+FL=Neuro-Fuzzy, ANN+GA=Neuro-Genetic and ANN+FL+GA=Neuro-Fuzzy-Genetic. Many hybridized forms of SC techniques has been reported in the literature for modeling GRNs [12, 26, 29, 31, 34-38, 54]. Table 1 summarizes the various types of hybridization used for modeling GRNs.

Neuro-fuzzy is one of the earliest and most widely used forms of hybridization. Liu *et.al.* [26] has proposed a neuro-fuzzy network models with biological knowledge to infer strong regulatory relationships and interrelated fuzzy rules. This model infers fuzzy rules automatically which describes the regulatory conditions in GRNs and explain the meaning of nodes and weight value in the neural network. Vineetha *et.al.* [35] presented a multilayered dynamic neuro-fuzzy network (DNFN) to extract gene regulatory relationship and reconstruct GRN for circulating plasma RNA data from colon cancer patients. This hybridized model combines the features of connectionist and FL to encode the knowledge learned in the form of fuzzy rules and processes data by applying the principles of fuzzy reasoning. A neuro-fuzzy inference system (NFIS) was applied by Jung & Cho [37] for reconstruction of GRNs. Here gene expression profile is first transformed into a mapping form then the transformed data are mapped into the NFIS and resulting fuzzy rules are applied to infer the relationship. The mapping of gene expression profile to fuzzy rules provides NFIS noise filtering capability for noisy and uncertain gene expression profile. Datta *et.al.* [34] tried to model GRN by a combination of RNN and fuzzy membership distribution of weights. A cost function had been applied to match the neurons response with the gene expression data and a differential evolution algorithm applied to minimize the cost function. The model has been used to infer the GRN of *SOS DNA repair network* of *e. coli.*

Table 1. Hybridized techniques for Modeling GRNs

Modeling techniques	Results obtained	References
RNN + PSO + ACO	Reconstructed genetic interaction network of yeast as well as SOS response system of e. coli	K. Kentzoglanakis, 2012 [36]
Neuro-fuzzy	Reconstruction of partial GRN of yeast	Liu et.al., 2011 [26]
Neuro-fuzzy	Extract regulatory relationships & construct GRN	Vineetha et.al., 2010 [35]
RNN+Fuzzy	Extracted GRN from yeast	Maraziotis, et.al., 2010 [12]
RNN+Clustering+PSO	Inferred GRN	Zhang, et.al., 2009 [29]
RNN+Fuzzy	Determine regulatory interaction between genes	Datta et.al., 2009 [34]
RNN + GA	Extracted GRN modules	Chiang & Chao, 2007 [31]
Neuro-fuzzy	Reconstructed GRN from microarray data	Jung & Cho, 2007 [37]
RNN + PSO	Extracted GRN from gene expression profiles.	Xu Rui et.al. 2007 [38]

Maraziotis *et.al.* [12] proposed a multilayer evolutionary trained neuro-fuzzy recurrent network (ENFRN) that select potential regulators of target genes and their regulation type. The recurrent, self-organizing structure and evolutionary training of the network give rise to an optimized collection of gene regulatory relations and its fuzzy nature eliminates noise-related issues. The ENFRN was tested on several benchmark datasets of yeast and it successfully retrieve biologically valid regulatory relationships and provide better insights for understanding the dynamics of GRNs. Chiang & Chao [31] has introduced a GA-RNN hybrid approach for finding feed-forward regulated genes. This GA-RNN hybrid method constructs various kinds of regulatory modules. RNN controls the feed-forward and feed-backward loop in regulatory module and GA provide ability of global searching of common regulated genes. This method extricates new feed-forward connections in gene regulatory models by modified multi-layer RNN architectures.

Zhang *et.al.* [29] proposed a hybridized form of PSO (particle swarm optimization) and RNN, called PSO-RNN. The PSO is a computational method that tries to optimize a problem by iteratively improving a candidate solution with regard to a given measure of quality. In this method, they have tried to integrate gene expression data and gene functional category information for the inference of GRNs. The inference was based on module network model which consists of two parts. The first is module selection part which determines the optimal modules using fuzzy c-means (FCM) clustering technique and incorporate functional category information. The second is network inference part, which uses PSO-RNN, to infer the underlying network between modules. The model was tested on real data from development of rat central nervous system (CNS) and the yeast cell cycle process. Another RNN-PSO (particle swarm optimization) based approach was proposed by X. Rui *et.al.* [38]. In this approach [38], gene interaction is demonstrated through a connection weight matrix and PSO-based searching algorithm is presented to uncover genetic network constructions that best fit with the time series data and analyse possible genetic interactions. PSO is used to train the network and find out the network parameters. For the real data set, this framework provides a meaningful insight into gene interactions in the network. K. Kentzoglanakis [36] has hybridized PSO, ant colony optimization (ACO) and ANN for modeling dynamic behaviour of gene regulatory systems. The ACO is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. ACO has been used for searching the discrete space of network architecture, PSO for searching the corresponding continuous space of RNN model parameters. This framework has been tested for the reconstruction of small artificial network as well as real-world data set of SOS response system of the *e.coli.*

5 Conclusions and Discussions

The gene regulatory networks (GRNs) demonstrate the interactions between genes. Understanding GRNs is essential because (i) it provides a large-scale, coarse-grained view of an organism at the mRNA level (ii) gives valuable indications for the therapeutics of complex diseases (iii) explains how different phenotypes emanate and which groups of genes are responsible for them and (iv) helps in understanding evolution by comparing genetic networks of various genomes. When comparing various methods for modeling GRNs, Boolean networks methods are useful to capture simplified interactions but these methods suffers from the loss of information due to discretisation. Also, it impractically assumes that transitions between activation states of the genes are synchronous. However, despite such limitations, these methods can be applied where accuracy is not the main concern. On the other hand, Bayesian networks methods are capable to deal with the stochastic aspects of gene expression and can handle noisy and incomplete data. However, it cannot deal with the dynamic aspects of gene regulations. Dynamic Bayesian networks were devised to solve dynamicity problem. To overcome information loss due to discretisation, ODE-based approach can be applied. These approaches provide detailed information about the network's dynamics but it requires huge amount of high-quality experimental data. The results of these methods are highly affected by noisy data.

When above methods are compared with soft computing (SC) based approach, SC-based approach are more robust and tolerant to noisy and incomplete data. The learning and adaptation capability of ANNs, knowledge representation through FLs and optimization capability of GAs when joined together, one can exploit the advantages of each of them. Also, different types of hybridization let us incorporate the generic and application-specific properties of these soft computing constituents. However, these SC-based methods require huge computation. The overall picture is that there is no any super model exists covering all aspects of cellular dynamics. We have observed that most of the techniques applied are hybridized forms of various SC techniques and clustering. Clustering is important because it allows preprocess of data and reduce data dimensionally so that computation time can be reduced.

We can improve our understanding of genetic interactions by (i) incorporating prior biological knowledge into the model (ii) integrating multiple biological data sources and (iii) decomposing the problem into smaller modules [29]. Modeling techniques can also be improved by (a) preprocessing gene expression data to reduce noises (b) incorporating clustering techniques to identify biologically meaningful modules which reduces the dimensionality of the data (c) applying soft computing method to capture nonlinear and dynamic relationships between genes.

Most of the proposed methods have various advantages and disadvantages; thus, we perceive a greater need for improving our understanding about the fundamental idea for each method and must consider available input data and constraints in choosing an appropriate modeling technique. Current research focuses on the modeling of GRNs from synthetic data, or on the simulation of small-scale regulatory networks with several genes or gene clusters. The modeling of large-scale genetic networks is yet to be done. Large number of genes, magnitude of the regulatory effect between the genes and speed of their regulatory response should also be incorporated in the model.

References

[1] de Jong, H.: Modeling and simulation of genetic regulatory systems: A literature review. J. Computational Biology 9, 67–103 (2002)

[2] Cho, K.-H., Choo, S.-M., et al.: Reverse engineering of gene regulatory networks. IET Syst. Biol. 1(3), 149–163 (2007)

[3] Sun, Y., Feng, G., Cao, J.: A new approach to dynamic fuzzy modeling of genetic regulatory networks. IEEE Transactions on Nanobioscience 9(4), 263–272 (2010)

[4] Naldi, A., Thieffry, D., Chaouiya, C.: Decision Diagrams for the Representation and Analysis of Logical Models of Genetic Networks. In: Calder, M., Gilmore, S. (eds.) CMSB 2007. LNCS (LNBI), vol. 4695, pp. 233–247. Springer, Heidelberg (2007)

[5] Remy, É., Ruet, P., Mendoza, L., Thieffry, D., Chaouiya, C.: From Logical Regulatory Graphs to Standard Petri Nets: Dynamical Roles and Functionality of Feedback Circuits. In: Priami, C., Ingólfsdóttir, A., Mishra, B., Riis Nielson, H. (eds.) Transactions on Computational Systems Biology VII. LNCS (LNBI), vol. 4230, pp. 56–72. Springer, Heidelberg (2006)

[6] Akutsu, T., Miyano, S., Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Pac. Symp. Biocomput., pp. 17–28 (1999)

[7] Martin, S., Shang, Z., Martino, A., Faulon, J.-L.: Boolean dynamics of genetic regulatory networks inferred from microarray time series data. Bioinformatics 23, 866–874 (2007)

[8] Shmulevich, I., Dougherty, E.R., Kim, S., Zhang, W.: Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks. Bioinformatics 18, 261–274 (2002)

[9] Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. J. Computational Biology 7, 601–620 (2000)

[10] Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 19, 2271–2282 (2003)

[11] Klipp, E.: Systems biology in practice: concepts, implementation and application. Wiley-VCH, Weinheim (2005)

[12] Maraziotis, I.A., Dragomir, A., Thanos, D.: Gene regulatory networks modeling using a dynamic evolutionary hybrid. BMC Bioinformatics 11, 140 (2010)

[13] de Jong, H., Page, M.: Search for steady states of piecewise-linear differential equation models of genetic regulatory networks. IEEE/ACM Trans. Computational Biology and Bioinformatics 5(2), 208–222 (2008)

[14] Chen, T., He, H.L., Churck, G.M.: Modeling gene expression with differential equations. In: Pac. Symp. Biocomput., pp. 29–40 (1999)

[15] Tyson, J.J., Csikasz-Nagy, A., Novak, B.: The dynamics of cell cycle regulation. Bioessays 24(12), 1095–1109 (2002)

[16] Koch, I., Schueler, M., Heiner, M.: STEPP – search tool for exploration of Petri net paths: a new tool for Petri net-based path analysis in biochemical networks. Silico Biol. 5, 129–137 (2005)

[17] Liang, S., Fuhrman, S., Somogyi, R.: REVEAL, a general reverse engineering algorithm for inference of genetic regulatory network architectures. In: Pacific Symposium on Biocomputing, vol. 3, pp. 18–29. World Scientific Publishing (1998)

[18] Mitra, S., Das, R., Hayashi, Y.: Genetic networks and soft computing. IEEE/ACM Trans. on Comp. Biology and Bioinformatics 8(1), 94–107 (2011)

[19] Karlebach, G., Shamir, R.: Modeling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology 9, 770–780 (2008)

[20] Bower, J.M., Bolouri, H.: Computational modeling of genetic and biochemical networks, pp. 1–48. MIT Press, London (2001)

[21] Schlitt, T., Brazma, A.: Current approaches to gene regulatory network modeling. BMC Bioinformatics 8 (suppl. 6), S9 (2007)

[22] Schreiber, F., et al.: A generic algorithm for layout of biological networks. BMC Bioinformatics 10, 375 (2009)

[23] Zadeh, L.A.: Fuzzy logic, neural networks and soft computing. One-page course announcement of CS 294-4. University of California at Berkeley (1992)

[24] Mitra, S., Hayashi, Y.: Bioinformatics with soft computing. IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Rev. 36(5), 616–635 (2006)

[25] Zadeh, L.A.: Fuzzy logic, neural networks, and soft computing. Comm. ACM 37, 77–84 (1994)

[26] Liu, G., et al.: Combination of neuro-fuzzy network models with biological knowledge for reconstructing gene regulatory networks. Journal of Bionic Engineering 8(1), 98–106 (2011)

[27] Vohradsky, J.: Neural network model of gene expression. FASEB J. 15, 846–854 (2001)

[28] Woolf, P.J., Wang, Y.: A fuzzy logic approach to analyzing gene expression data. Physiological Genomics 3, 9–15 (2000)

[29] Zhang, Y., et al.: Reverse engineering module networks by PSO-RNN hybrid modeling. BMC Genomics 10 (suppl. 1), S15 (2009)

[30] Tian, T., Burrage, K.: Stochastic neural network models for gene regulatory networks. In: IEEE Congress on Evolutionary Computation, pp. 162–169 (2003)

[31] Chiang, J.-H., Chao, S.-Y.: Modeling human cancer-related regulatory modules by GA-RNN hybrid algorithms. BMC Bioinformatics 8, 91 (2007)

[32] Du, P., et al.: Modeling gene expression networks using fuzzy logic. IEEE Transcation on Systems, Man and Cybernetic – Part B: Cybernetics 35(6), 1351–1359 (2005)

[33] Ram, R., Chetty, M., Dix Trevor, I.: Fuzzy model for gene regulatory network. In: Proc. of IEEE Congress on Evolutionary Computation, pp. 1450–1455 (2006)

[34] Datta, D., et al.: A recurrent fuzzy neural model of a gene regulatory network for knowledge extraction using differential equation. In: Proc. of IEEE Congress on Evolutionary Computation, pp. 2900–2906 (2009)

[35] Vineetha, S., Chandra, C., Bhat, S., Idicula, S.M.: Gene regulatory network from microarray data using dynamic neural fuzzy approach. In: Proceedings of the International Symposium on Biocomputing (ISB 2010). ACM, New York (2010)

[36] Kentzoglanakis, K.: A swarm intelligence framework for reconstructing gene networks: searching for biologically plausible architectures. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(2), 358–371 (2012)

[37] Jung, S.H., Cho, K.-H.: Reconstruction of gene regulatory networks by neuro-fuzzy inference system. In: Frontiers in the Convergence of Bioscience and Information Technologies, pp. 32–37 (2007)

[38] Rui, X., Wunsch, D.C., Frank, R.L.: Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. IEEE/ACM Transactions on Comp. Biology and Bioinformatics 4(4), 681–692 (2007)

[39] Ressom, H., Wang, D., Varghese, R.S., Reynolds, R.: Fuzzy logic-based gene regulatory network. In: IEEE International Conference on Fuzzy Systems, vol. 2, pp. 1210–1215 (2003)

[40] Kim, S., et al.: Multivariate measurement of gene expression relationships. Genomics 67, 201–209 (2000)

[41] Huang, J., Shimizu, H., Shioya, S.: Clustering gene expression pattern and extracting relationship in gene network based on artificial neural networks. J. Bioscience and Bioeng. 96, 421–428 (2003)

[42] Zhou, X., et al.: A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks. Bioinformatics 20(17), 2918–2927 (2004)

[43] Keedwell, E., Narayanan, A., Savic, D.: Modeling gene regulatory data using artificial neural networks. In: Proc. of the 2002 IEEE/INNS/ENNS International Joint Conference on Neural Networks (IJCNN 2002), pp. 183–189 (2002)

[44] Weaver, D.C., Workman, C.T., Stormo, G.D.: Modeling regulatory networks with weight matrices. In: Proc. Pacific Symp. Biocomputing, pp. 112–123 (1999)

[45] Hu, X., Maglia, A., Wunsch II, D.C.: A general recurrent neural network approach to model genetic regulatory networks. In: Proc. of IEEE Engineering in Medicine and Biology Annual Conference, pp. 4735–4738

[46] Ando, S., Sakamoto, E., Iba, H.: Modeling genetic network by hybrid GP. In: Proc. of the Congress on Evolutionary Computation, CEC 2002, vol. 1, pp. 291–296 (2002)

[47] Wang, H., Qian, L., Dougherty, E.: Inference of gene regulatory networks using genetic programming and Kalman filter. In: IEEE GENSIPS, pp. 27–28 (2006)

[48] Sirbu, A., Ruskin, H.J., Crane, M.: Comparison of evolutionary algorithms in genetic regulatory network model. BMC Bioinformatics 11, 59 (2010)

[49] Maeshiro, T., et al.: An evolutionary system for prediction of gene regulatory networks in biological cells. In: SICE Annual Conference 2007, pp. 1577–1581 (2007)

[50] Noman, N., Iba, H.: Reverse engineering genetic networks using evolutionary computation. Genome Informatics 16(2), 205–214 (2005)

[51] Kimura, S., et al.: Inference of S-system models of genetic networks using cooperative coevolutionary algorithm. Bioinformatics 21(7), 1154–1163 (2005)

[52] Chowdhury, A.R., Chetty, M.: An improved method to infer gene regulatory network using S-System. In: IEEE Congress on Evolutionary Computation, pp. 1012–1019 (2011)

A Novel Algorithm for Hub Protein Identification in *H.Sapiens* Using Global Amino Acid Features

Aswathi B.L., Baharak Goli, and Achuthsankar S. Nair

Department of Computational Biology and Bioinformatics,
University of Kerala,
Trivandrum 695581, India
aswathi.bl@gmail.com

Abstract. Identification of hub proteins solely from amino acids in proteome remains an open problem in computational biology that has been getting increasing deliberations with extensive growth in sequence information. In this context, we have chosen to investigate whether hub proteins can be predicted from amino acid sequence information alone. Here, we propose a novel hub identifying algorithm which relies on the use of conformational, physiochemical and pattern characteristics of amino acid sequences. In order to extract the most potential features, two feature selection techniques, CFS (Correlation-based Feature Selection) and ReliefF algorithms were used, which are widely used in data preprocessing for machine learning problems. The performance of two types of neural network classifiers such as RBF network and multilayer perceptron were evaluated with these filtering approaches. Our proposed model led to successful prediction of hub proteins from amino acid sequences alone with 92.98% and 92.61% accuracy for multilayer perceptron and RBF Network respectively with CFS algorithm and 94.69% and 90.89% accuracy for multilayer perceptron and RBF Network respectively using ReliefF algorithm.

Keywords: Protein hubness, Protein protein interaction networks, Protein protein interaction, feature selection methods, machine learning.

1 Introduction

With the rapid advancement of amino acid sequencing technologies and databases the amount of proteomic data has been increasing almost exponentially. The most important biologically functional parts of amino acid sequence of any organism are its proteins. Proteins are the work horse molecules of the cellular machinery, which mediate a broad range of cellular functions. Proteins usually function through their interactions with other proteins. Such a group of proteins with their interactions form a protein-protein interaction network (PPIN) [1]. In a PPIN, a protein is denoted by a node and a connecting edge represents a protein-protein interaction. The degree of a protein represents the total number of interactions that protein has. Highly interactive proteins are called ‘hubs’ and they literally 'hold the protein interaction networks together' [2]. Hub proteins are known to have high density of binding sites [3], which enable them to have multiple interactions. Most of the protein-protein interaction networks consist

N. Meghanathan et al. (Eds.): Advances in Computing & Inf. Technology, AISC 178, pp. 13–22.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

of small number of hub proteins while the sparsely connected proteins are rich in number [4].

Analysis of hub proteins assumes vital importance, since they are highly interactive and the possibility of their involvements in multiple pathways are higher [3]. When a hub node is deleted, it is more lethal to the organism than the deletion of those nodes which are sparsely in a protein-protein interaction network [2]. Hub characterization is highly crucial for better realization of cellular functions as well as discovering novel drug targets and predicting the side effects in drug discovery by understanding the pathways, topologies and dynamics of them. Most of the well-known and widely examined proteins including p53 are concerned in diseases, are hubs and studying these hub proteins can provide useful information for predicting the possible side effects in drug discovery [1,4,5].

A Large number of computational algorithms have been proposed to predict hub proteins in protein-protein interaction networks using various data such as gene ontology [6], gene proximity [7, 8], gene fusion events [9, 10] and gene co-expression data [11-12]. But most of such computational predictions have been focused on the identification of binary protein-protein interactions with varying degrees of accuracies [1]. One of the major limiting factors for using the above mentioned data is the lack of availability of them for the entire protein interaction data of an organism. Application of existing methods which use structural information is also severely limited as PDB structures are not available for many of the proteins [1].

In order to surmount the limitations of availability of structural and ontology data which are slow in emergence, in this study we have developed a statistics-based approach to discriminate hub and non hub proteins from amino acid sequence information alone using soft computational algorithms.

2 Materials and Methods

2.1 Dataset

For this study, we selected *H.Sapiens* as the model organism, which is well annotated and have modest protein interaction information. The protein interaction data was extracted from IntAct [13] database. These data were then curated to obtain the non-redundant dataset which included 10,578 Protein- protein interactions. Corresponding amino acid sequences of varying lengths were compiled from Uniprot [14]. Total number of protein interactions was 53120 with an average degree of interaction 9.534.

2.2 Identification of the Degree for Hubs

The degree of connectivity of proteins in our PPI dataset ranged from 1 to 450. For classifying a protein as hub, we had to determine a degree threshold. Based on the literature survey, the degree thresholds or connectivity cut-off of hub proteins are species specific [5]. So far, there is no concordance on the exact connectivity threshold values for these proteins [5]. In some of the previous studies, these thresholds were taken based on fold change and the accumulative protein interaction distribution plots in some of the previous studies [5, 3] and we have adopted the fold change approach [1]. The degree fold change was determined as the ratio of the connectivity value and average connectivity. A node with fold change greater than or equal to 2 (cutoff, P-value < 0.001, using distribution of standard normalized fold change values in H.Sapiens) was the criterion applied for considering a protein as hub [1]. To ensure rigorous screening of non-hubs, we considered only those proteins which have degree in a range between 1 and 5 for non-hub test and train set. The final number of highly connected protein was 550 and sparsely connected protein was 2010.

2.3 Feature Transformation

The quantitative characteristics of amino acid sequences that we took into our consideration included 28 Amino acid pattern-features, 3 conformational lineaments and 14 physiochemical properties.

Amino acid pattern- features

This include, amino acid composition (20 features), atomic composition (5 features), the ratio of strong and weak hydrobhobic residues of an amino acid sequence using Chaos Game Representation approach [15](1 feature) and Spectral areas obtained through Fast Fourier transformation for both hydrophobicity and frequency distribution of 6 phosphorylation- prone amino acids (2 features).

Amino acid composition of a protein sequence is comprised of the frequencies of each residue or amino acid. Hence, we got 20 features for all 20 residues for each amino acid sequence. For Atomic composition we extracted 5 features, which were computed by measuring the frequencies of five different atoms, Carbon, Nitrogen, Hydrogen, Sulfur, Oxygen, which constitute an amino acid. For each amino acid sequence, the ratios of strong and weak hydrobhobic residues were obtained using Chaos Game Representation (CGR) approach, which is one of the graphical representation methods for biological sequences [15]. We divided the 20 amino acids into 4 groups as, least hydrophobic (Arginine, Lysine, Asparagine, Glutamine, Glutamic Acid, Histidine, Aspartic Acid), weak hydrophobic (Proline, Tyrisine, Tryptophan, Threonine, Glycine, Serine), medium hydrophobic (Cysteine, Alanine, Phenylalanine, Methionine) and strong hydrophobic (Isoleucine, Leucine, Valine) based on the hydrophobicity values and represent each group at each corner of the CGR Plot. After getting the CGR graph, it is divided by a hyper plane and hence the total amino acid distribution is divided into two groups- Least Hydrophobic and strong Hydrophobic. Linear sum of each group is calculated and the ratio is taken. Fig. 1 illustrates the Hydrophobicity- ratio computation using CGR plots. CGR points can be generated by an iterated function system defined by the following equations,

$$X_i = 0.5\ (X_i\text{-}1+g_{ix})$$

$$Y_i = 0.5\ (Y_i\text{-}1+g_{iy})$$

Where, g_{ix} and g_{iy} correspond to the X and Y co-ordinates of the amino acid at position i in the sequence.

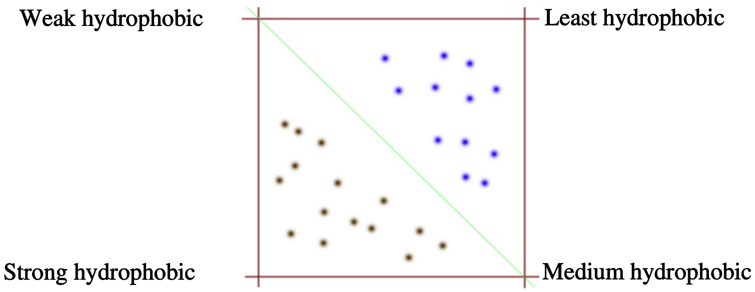
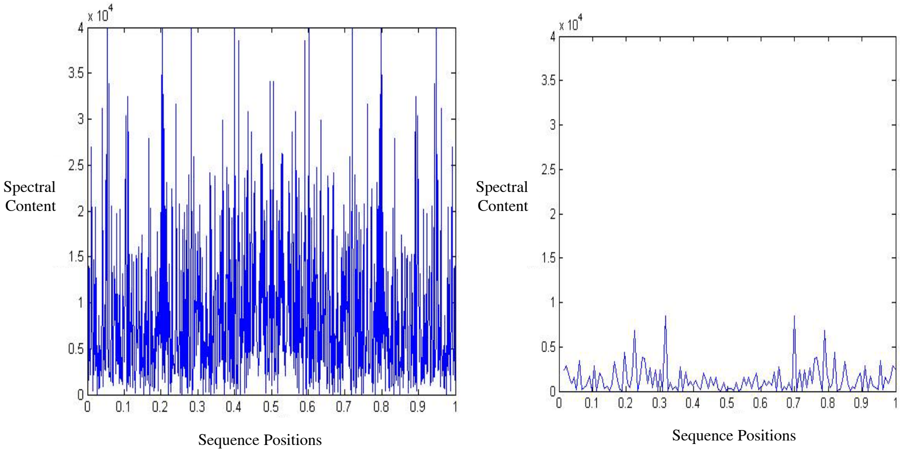


Fig. 1. Hyrophobicity- ratio plot using CGR for any amino acid sequence

Another feature, the spectral areas obtained through Fast Fourier transformation (FFT), were also taken to consideration. FFT was applied for both hydrophobicity and frequency distribution of phosphorylation- prone amino acids, Hystidine, Lysine, Arginine, Serine, Threonine and Tryptophan. The spectra shows remarkable discriminative patterns (Fig. 2).



Hydrophobicity spectra for a highly connected sequence in *H.Sapiens*

Hydrophobicity spectra for a sparsely connected sequence in *H.Sapiens*

Fig. 2. Graphical representations of Hydrophobicity spectral-distribution for a sample hub and non-hub protein sequence

Amino acid Conformational features

The conformational parameters were obtained from the secondary structure information of the amino acid sequences. This includes the percentage of Alpha helices, Beeta sheets and Coils which makes the secondary structure of a protein from its amino acid sequences.

Amino acid Physiochemical features

We took a total of 14 physiochemical properties of amino acids from the amino acid index database AAIndex [16]. According to literature review, most of these features show strong correlation with protein- protein interactions. The chosen physiochemical properties are listed in table1.

Table. 1. Amino acid Physiochemical features compiled from AA index [16]

Sl.No.	Amino acid Properties
1	Free energy of transfer to surface
2	Hydrophobicity index
3	Refractivity
4	Molecular Weight
5	Electron_ion interaction potential
6	Reduced distance
7	Recognition factor
8	Bulkiness
9	Transmembrane Index
10	Flexibility
11	Polarity
12	isoelectric point
13	Absolute entropy
14	Residue Volume

2.4 Feature Pruning

Generally, the performance of any classifier depends on the reliability of the features taken, the size of the training set and the complexity of the classifier [17]. Applying large number of features will increase the computation time which in turn affect the efficiency of classification algorithms [18] over-fitting the training data set [19]. Faster classification models and smallest subset of important and prominent features should be retained, in order to attain maximal classification performance. Feature selection is one of the significant techniques in data preprocessing for machine learning and data mining problems, which trashes out irrelevant, noisy and redundant features and speeds up the data mining algorithm and improves prediction accuracy [17, 20]. For this we adopted two well-known feature selection techniques such as CFS (correlation-based feature selection) [21] and ReliefF feature selection algorithm [22] to prune out the prominent discriminatory set of features. We briefly describe these feature selection algorithms below. In this study 45 features generated from the transformation step explained above and after feature selection a total of 16 features remained.

Feature Selection algorithm: Relief Feature Selection (ReliefF)

This well-known feature selection technique is an extension of Relief algorithm developed to use in classification problems [17, 23]. Based on the strong correlation between the features it evaluates the relevance of these features. An instance i is selected randomly from the dataset and the weight for each feature is rationalized based on the distance of ‘d’ to its NearHit (nearest neighbors from the same class) and NearMiss (nearest neighbors from each of the different classes) at each step of an iterative process.[17]. This process is iterated‘t’ times, where t is a predefined parameter and is equal to the number of samples in dataset. Finally the best subset includes those features with relevance above a chosen cut-off.

Feature trimming algorithm: Correlation-Based Feature Selection (CFS)

This is a powerful technique in filtering uncorrelated and duplicate features. It evaluates the importance of subsets of features by using a best first-search heuristic approach. [17] This heuristic algorithm considers the importance of individual features for predicting the class along with the level of correlation among them. The basic logic in CFS is that good feature subsets include those features that are highly correlated with the target class and uncorrelated with each other.

2.5 Construction of Neural Network Classifiers

Artificial neural network is one of the supervised learning algorithms used commonly to solve classification problems. In this study, we used two types of neural networks configurations, multilayer perceptron trained by the back propagation algorithm and RBF network. For the implementation we used, weka suite, a machine learning workbench developed in java programming language [24]. Since the Back-propagation networks has less memory requirements, it is one of the most common and widely used algorithms for training supervised neural networks [25], [26], [27]. RBF networks are supervised neural networks which are popular substitute to multilayer perceptions which employ reasonably lesser number of locally tuned units and are adaptive in nature. They are widely used for classification and pattern recognition problems. In this study, the training set consisting of 550 hubs and 2010 non-hubs elements was given to the each network in the 10-fold cross-validation scheme. The accuracy of classification using each network was measured. For the comparison of the networks, the time taken by each network to build the model was also noted.

3 Results

3.1 Performance Evaluation

The performance of our proposed classification models were estimated using standard 10-fold cross-validation in which the whole dataset is randomly partitioned into ten evenly-sized subsets. During each test, a neural network is trained on nine subsets and then tested on the tenth one. This method is repeated ten times so that each subset is used for both training and testing on each fold. Several measures were used to evaluate the performance of the neural networks (True positive (TP), True negative (TN), False positive(FP), and False negative (FN), respectively).These measures include, Specificity=TN/ (TN+FP)*100, Sensitivity=TP/ (TP+FN)*100, Precision=TP/ (TP+FP)*100, Matthews correlation coefficient (MCC) = (((TP*TN)-(FP*FN)))/ (√(TP+FP)*(TP+FN)* (TN+FP)*(TN+FN)) and Accuracy= TP+TN/ (TP+TN+FP+TN). Table 2 summarizes the performance of different classifiers.

Table 2. Performance of different Hub prediction algorithms

Classification method	**Sensitivity** (%)	**Specificity** (%)	**Accuracy** (%)	**Precision** (%)	**MCC**
Multilayer perceptron + CFS	92.81	93.17	92.98	96.12	0.93
RBF Network + CFS	91.73	93.48	92.61	98.32	0.87
Multilayer perceptron + Relief-f	92.06	95.31	94.69	98.12	0.91
RBFNetwork + Relief-f	92.62	89.18	90.89	97.56	0.89

Multilayer Perceptron in combination with relief-f algorithm produced highest classification result. Time taken to build the models were 76.42 seconds for multilayer perceptron and 4.21 seconds for RBF network in case of CFS and 78.26 seconds for multilayer perceptron and 5.22 seconds for RBF network in case of relief feature selection algorithm in the same work station. To evaluate the classification model, Self-consistency test and independent test were also done. The results are shown in Table 3. Self-consistency test checks the consistency of the developed model. A classification method can be considered as a good one, if the self-consistency of that method is good. In self-consistency test, observations of training datasets are predicted with decision rules acquired from the same dataset. The accuracy of self-consistency determines the fitting ability of the rules obtained from the features of training sets. Since the prediction system parameters obtained by the self-consistency test are from the training dataset itself, the success rate is high. However poor result of self- consistency test shows the inefficiency of classification method. In independent dataset the training set was composed two equal halves of hub and non-hub proteins. The remaining sequences were used as the testing set.

Table 3. Accuracy of each classifier for self-consistency and independent data test

Classification Method	**Self-consistency** (%)	**Independent Test** (%)
Multilayer perceptron + CFS	95.45	89.47
RBF Network + CFS	94.71	91.83
Multilayer perceptron + Relief-f	98.66	95.67
RBF Network + Relief-f	96.61	88.19

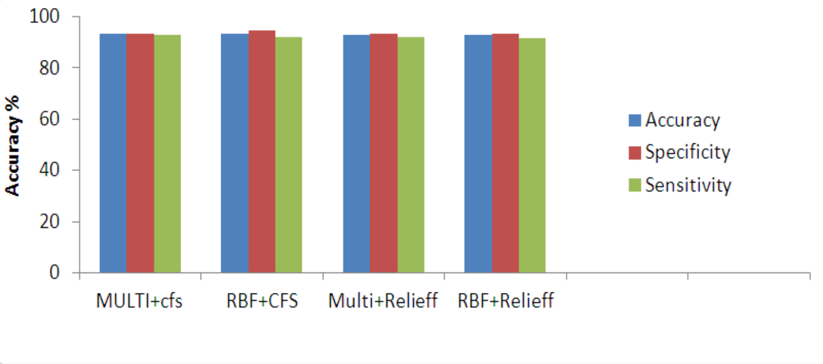


Fig. 3. Average Accuracy, Specificity and Sensitivity for various classification methods

4 Discussion

In this study, a novel hub prediction algorithm which relies only on the use amino acid sequence information was proposed. Analyzing structural and functional phenomena from sequence information is not a novel approach. It has been widely used with the advent of bioinformatics approaches in genomics and proteomics studies. There have been many computational Biology works which applies this approach to various problems including gene finding [28], protein subcellular localization [29] and protein allostery prediction [30].

Our results show that the extracted amino acid features have strong correlation in classifying hub from non- hub proteins. With Correlation based feature selection and the Relief-F algorithm followed by two classification algorithms, multilayer perceptron and RBF Networks, we could effectively trace out useful amino acid features which are significant in the hub protein identification. The biological importance of the chosen amino acid properties in this work are yet to be explained. It would be remarkable to investigate the significance of these properties in the formation of PPINs.

References

1. Aswathi, B.L., Nair, A.S., Sivasankaran, A., Dhar, P.K.: Identification of hub proteins from sequence. Bioinformation 7 (2011)
2. Tun, K., Rao, R.K., Samavedham, L., Tanaka, H., Dhar, P.K.: Rich can get poor: conversion of hub to non-hub proteins. Systems and Synthetic Biology 2, 75–82 (2009)
3. He, X., Zhang, J.: Why do hubs tend to be essential in protein networks? PLoS Genetics 2, e88 (2006)
4. Patil, A., Kinoshita, K., Nakamura, H.: Hub promiscuity in protein-protein interaction networks. International Journal of Molecular Sciences 11, 1930–1943 (2010)
5. Hsing, M., Byler, K.G., Cherkasov, A.: The use of Gene Ontology terms for predicting highly-connected “hub” nodes in protein-protein interaction networks. BMC Systems Biology 2, 80 (2008)

6. Srihari, S.: Detecting hubs and quasi cliques in scale-free networks. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4 (2008)
7. Dandekar, T., Snel, B., Huynen, M., Bork, P.: Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. 23, 324–328 (1998)
8. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N.: The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA 96, 2896–2901 (1999)
9. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D.: Detecting protein function and protein-protein interactions from genome sequences. Science 285, 751–753 (1999)
10. Enright, J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A.: Protein interaction maps for complete genomes based on gene fusion events. Nature 402, 86–90 (1999)
11. Ge, H., Liu, Z., Church, G.M., Vidal, M.: Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat. Genet. 29, 482–486 (2001)
12. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA 96, 4285–4288 (1999)
13. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., et al.: IntAct–open source resource for molecular interaction data. Nucleic Acids Research 35, D561–D565 (2007), `http://www.ebi.ac.uk/intact/main.xhtml`
14. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., et al.: UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 9, D115–D119 (2004), `http://www.uniprot.org`
15. Jeffrey, H.J.: Chaos game representation of gene structure. Nucleic Acids Res. 18, 2163–2170 (1990)
16. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M.: AAindex: amino acid index database, progress report 2008. Nucleic Acids Research 5, D202–D205 (2008), `http://www.genome.jp/aaindex/`
17. Goli, B., Aswathi, B.L., Nair, A.S.: A Novel Algorithm for Prediction of Protein Coding DNA from Non-coding DNA in Microbial Genomes Using Genomic Composition and Dinucleotide Compositional Skew. In: Meghanathan, N., Chaki, N., Nagamalai, D. (eds.) CCSIT 2012, Part II. LNICST, vol. 85, pp. 535–542. Springer, Heidelberg (2012)
18. Hall, M., Holmes, G.: Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Trans. Knowl. Data Eng. 15, 1–16 (2003)
19. Wang, C., Ding, C., Meraz, R.F., Holbrook, S.R.: PSoL.: A positive sample only learn-ing algorithm for finding non-coding RNA genes. Bioinformatics 22, 2590–2596 (2006)
20. Liu, H., Yu, L.: Towards integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering 17(3), 1–12 (2005)
21. Hall, M.A.: Correlation based feature selection for machine learning. Doctoral dissertation, The University of Waikato, Dept. of Comp. Sci. (1999)
22. Marko, R.S., Igor, K.: Theoretical and empirical analysis of relief and rreliefF. Machine Learning Journal 53, 23–69 (2003)
23. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Proceedings of the Ninth International Workshop on Machine Learning, pp. 249–256. Morgan Kaufmann Publishers Inc. (1992)
24. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
25. Werbos, P.J.: Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University (1974)

26. Parker, D.B.: Learning-logic. Technical report, TR-47, Sloan School of Management. MIT, Cambridge (1985)
27. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error-propagation in Parallel distributed processing. In: Explorations in the Microstructure of Cognition, vol. I. Bradford Books, Cambridge (1986)
28. Achuthsankar, S.N., Sreenadhan, S.P.: An improved digital fltering technique using nucleotide frequency indicators for locating exons. Journal of the Computer Society of India 36, 60–66 (2006)
29. Cherian, B.S., Nair, A.S.: Protein location prediction using atomic composition and global features of the amino acid sequence. Biochemical and Biophysical Research Communications 391, 1670–1674 (2010)
30. Namboodiri, S., Verma, C., Dhar, P.K., Giuliani, A., Nair, A.S.: Sequence signatures of allosteric proteins towards rational design. Systems and Synthetic Biology 4, 271–280 (2011)

Rough Set Based Classification on Electronic Nose Data for Black Tea Application

Anil Kumar Bag[1], Bipan Tudu[2],
Nabarun Bhattacharyya[3], and Rajib Bandyopadhyay[2]

[1] Department of Applied Electronics and Instrumentation Engineering,
Future Institute of Engineering and Management, Kolkata-700 150, India
[2] Department of Instrumentation and Electronics Engineering, Jadavpur University,
Salt Lake Campus, Sector III, Block LB, Plot No. 8, Kolkata-700 098, India
[3] Centre for Development of Advanced Computing(C-DAC),
E-2/1, Block – GP, Sector – V, Salt Lake, Kolkata-700 091, West Bengal, India
anilkumarbag@gmail.com, {bt,rb}@iee.jusl.ac.in,
nabarun.bhattacharya@cdac.in

Abstract. The responses generated by a gas sensor array are difficult to classify due to their inherent imprecision, uncertainty and the procedures of computational intelligence are appropriate to deal with such imperfect knowledge. In recent years, rough set theory has attracted more attention of many researchers even though it was proposed in the early 1980's by Z. Pawlak. The rough set based analysis makes it very convenient for classification of data especially with huge volume of information, as the method is very efficient to find the optimal subset of attributes. In this paper, the rough set based algorithm has been applied to generate representative rules using the datasets obtained from a gas sensor array in an electronic nose instrument, capable of sensing aroma of black tea samples and these rules are used to classify the black tea quality.

Keywords: Black tea, Electronic nose, Gas sensor array, Rough set, Reduct, Lower approximation, Upper approximation.

1 Introduction

The electronic nose instrument nowadays finds very useful applications for classification of products based on their odour and intense research in the field of sensors and pattern recognition is advancing the progress of this technology with more and more novel applications [1]-[3]. An extremely useful and necessary application of electronic nose is in the field of tea testing. Till date, tea quality evaluation is based on the verdict of human experts, called tea tasters and they grade different qualities of tea based on their professional acumen and experience. This method of quality assessment is very subjective, and the grades vary from taster to taster. Moreover, the mood and other psychological factors of the tea taster play significant role in the evaluation process. Thus, there is a need in the tea industry for an unbiased and correct procedure for the evaluation of tea quality. But this task is extremely difficult and challenging as

N. Meghanathan et al. (Eds.): Advances in Computing & Inf. Technology, AISC 178, pp. 23–31.
springerlink.com © Springer-Verlag Berlin Heidelberg 2013

the number of volatiles present in tea and contributing to its quality is more than two hundred and an electronic nose can play a significant role in solving this problem.

A few research reports on the applicability of electronic nose for aroma characterization of tea reveal that the instrument, when designed for tea aroma classification, has the potential to be employed regularly as a useful gadget in the tea industry [4]-[6]. The pioneering work has been done by Dutta et al. [4], where the efficacy of electronic nose systems in classifying black tea aroma in different processing stages was demonstrated. Correlation of electronic nose data with the tea taster marks has been successfully carried out in [6]. The electronic nose has demonstrated its usefulness in monitoring the aroma of black tea during the fermentation process [7]. In these systems, the MOS sensors with headspace sampling have been used for aroma characterization of tea, but there are uncertainty and vagueness in the data set generated by these sensors. This vagueness is introduced due to the variation in the amount of volatiles in the samples, sensor drift, and noise. Another important source of uncertainty is the tea taster's score, which is used for training the classifier. As a result, the data set may contain some irrelevant, redundant features, which unnecessarily increase the computational complexity of the classification algorithm. In addition, presence of vagueness in the data set degrades the accuracy of classification. Classification of such data set thus becomes more challenging.

In order to calibrate the electronic nose instrument with such uncertain and vague data, a rough set based classifier has been considered in this paper. So far, to the best of our knowledge, the rough set based approach has not been explored in the field of machine olfaction. The classification algorithms are mostly based on neural networks or fuzzy logic or other computational intelligent methods [6], [8]. Classification accuracy of these algorithms depends upon initialization of different parameters, number of iterations and inconsistency of the data. For consistency of the data set, a separate algorithm is usually employed [9], [10] for feature selection. Compared to these classification methods, the rough set based method has the advantage that the method is capable of handling inconsistent data sets in an efficient manner.

The theory of Rough set was introduced in 1980 by Z. Pawlak [11] as a new intelligent mathematical tool for knowledge discovery and data analysis based on the concept of approximation spaces. The uniqueness of rough set theory based classifications is highlighted by the facts that it does not need any preliminary or additional information about data, i.e., probability in statistics, basic probability assignment in the Dempster-Shafer theory, grade of membership, or the value of possibility in fuzzy set theory. Also, over and above the conventional parametric and non-parametric data classification techniques, the rough set approach is capable of extracting minimal information by data reduction, exploration of hidden patterns efficiently in a data set, evaluating the significance of data, generation of minimal set of decision rules, analysis of conflicts and intelligent pattern classification [12]. These features of rough set theory make it an excellent classifier for electronic nose applications, as it can optimize the sensor array while classifying the patterns. The data analysis algorithm does not create much overhead in the computational system and may easily be embedded in field deployable electronic nose systems for tea quality evaluation.

Essentially, the array of sensors in an electronic nose produces continuous real valued attributes corresponding to different volatiles present in black tea samples. In rough set approach, these real valued attributes are then discretized [13]-[16] based on discernibility matrix [17] to remove superfluous attribute information by unifying values in some intervals and at the same time preserving the necessary information. Then a subset of the attributes is selected which has the same classification capability as with the entire set of attributes. The rules are then extracted using the concept of reduct [18], [19]. This optimum rule set so generated is used finally for classification of the data.

2 Rough Set

In the field of classification of objects which are described by a set of real valued condition attributes and assigned to certain decisions, rough set method is a very efficient tool to find the relative reduct and hence to generate decision rules.

Z. Pawlak introduces the concept of rough set theory in the early 1980s. It is an excellent mathematical tool for the analysis of a vague description of objects. The Information System $IS = \left(U, A \cup \{d\}, V, f\right)$ a tabular form of OBJECT→ATTRIBUTE VALUE relationship, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, V is the union of attribute domains (i.e., $V = \bigcup_{a \in A} V_a$, where V_a denotes the domain of attribute a) and f is a function such that for any $u \in U$ and $a \in A$, $f\left(u, a\right) \in V_a$ while d is called decision attribute. For each possible subset of attributes $B \subseteq A$, a decision table generates an equivalence relation called an indiscernibility relation $IND(B)$, where two objects $\left(u_i, u_j\right)$ are members of the same equivalence class if and only if they cannot be discerned from each other on the basis of the set of attributes B . The equivalence classes of the B -Indiscernibility relation are denoted $\left[u\right]_B$. Indiscernibility relation is defined as

$$IND\left(B\right) = \left\{\left(u_i, u_j\right) \in |U| \times |U| : \forall a \in B, f(u_i, a) = f(u_j, a)\right\}$$

which induces a partitioning of the universe U according to the attribute set B .

The discernibility knowledge of the information system is commonly recorded in a symmetric $|U| \times |U|$ matrix called the discernibility matrix [17]. Thus any set $X \subseteq U$ can be approximated solely on the basis of information in $B \subseteq A$ by constructing a B *-lower approximation* and B *-upper approximation.* The B -lower approximation of X is defined as the unions of all the elementary sets which are certainly in X i.e. $\underline{B}X = \left\{x : [x]_B \subseteq X\right\}$. The B-upper approximation of X is defined as the union of the elementary sets, which have a non-empty intersection with X i.e. $\overline{B}X = \left\{x : [x]_B \cap X \neq \phi\right\}$. Thus the lower approximation consists of objects that definitely belong to X and the upper approximation contains objects that possibly belong to X .

Now the reduct is the minimal subset of attributes that enables the same classification of elements of the universe as the whole set of attributes. In other words, attributes that do not belong to a reduct are superfluous with regard to classification of elements of the universe.

The rough set method deals with discrete attributes. Hence, before attempting to find the reduct set, the real valued attributes are discretized since the discrete features are closer to a knowledge-level representation than the continuous ones. Here we employ the discretization method based on binary discernibility matrix. This method first produces the cut set whose elements are the middle points of the intervals present in each attribute after the real values of each attribute are sorted in either descending or ascending order. Then, reduct finding algorithm is used to find the minimal set of cuts (optimal cut-set) considering cuts as the attributes. The discretized information system is finally presented using the optimal cut set (OCS). Also, the decision rules generated from discrete features are easier to understand for both users and experts. The algorithms for reduct generation, discretization and rule generation can find in the paper published by the authors Bag et al. [20].

3 Electronic Nose for Black Tea Quality Estimation

In this section, a brief description of the electronic nose instrument for tea quality estimation and the experiment with tea samples are presented.

3.1 Customized Electronic Nose Setup for Black Tea

A customized electronic nose setup has been developed for quality evaluation of tea aroma, the details of which are presented in [6]. Five gas sensors from Figaro, Japan – TGS-832, TGS-823, TGS-2600, TGS-2610 and TGS-2611 constitute the sensor array for the setup.

The experimental conditions of the electronic nose for classification of black tea aroma are given as follows:

- Amount of black tea sample = 50 grams,
- Temperature $= 60^0 C \pm 3^0 C$,
- Headspace generation time = 30s,
- Data collection time = 100s,
- Purging time = 100s,
- Airflow rate = 5 ml/s.

Dry tea samples have been used during the experiments in order to avoid the effect of humidity. During each sniffing cycle, all the five sensors are exposed to the tea volatiles, and the maximum response of each sensor is considered for subsequent computation. The above experimental conditions have been optimized for black tea quality evaluation on the basis of repeated trials and sustained experimentation.

3.2 Sample Collection and Tea Taster’s Score

Experiments were carried out for approximately one-month duration each at the tea gardens of the following industries:

- Khongea Tea Estate
- Mateli Tea Estate
- Glenburn Tea Estate
- Fulbari Tea Estate

The industries have multiple tea gardens spread across north and north-east India and the tea produced in their gardens are sent everyday to the tea testing centers for quality assessment. All the companies had expert tea tasters and for our experiments, one expert tea taster was deputed by the respective industries to provide the taster's score to each of the samples. The taster's score were subsequently considered for the correlation study with the computational model. A sample tea taster score sheet is given in Table 1.The scores assigned to "aroma", signify the smell and flavor of the samples and for correlation with electronic nose, only the aroma scores have been considered.

4 Data Analysis and Results

The total number of samples considered for the present study is 194 and their details are presented in Table 1.

Table 1. Sample Details

Tea sample from the garden	Number of data array	Taster's scores (Aroma)
Khongea Tea Estate	104	4, 5, 6, 6.5, 7
Mateli Tea Estate	30	8
Glenburn Tea Estate	30	8
Fulbari Tea Estate	30	7, 7.5

The data arrays produced by electronic nose used as (IS) is shown in Table 2. Each sample is an object and a unique number is assigned to each of the objects in a serial manner. The information for a particular object comprises of the responses of five sensors (the condition attributes - $a1, a2 \ldots a5$) and the corresponding tea taster's mark for aroma (the decision attribute - *d*) and are stored in a row. For the samples under study, there are seven different scores assigned by the tea tasters ranging from 3 to 8. For our convenience, these scores have been replaced by numbers from 1 to 7.

Table 2. Data Arrays Produced by Electronic Nose used as (*IS*)

Objects (*U*)	Attributes (*A*)					
	Sensors response					Tea category
	a1	*a2*	*a3*	*a4*	*a5*	*d*
1	0.0936	0.0583	0.0382	0.1275	0.0008	1
…	…	…	…	…	…	…
194	0.3682	0.1288	0.0681	0.9039	0.0110	7

The (IS) contains the real valued condition attributes. These attributes are discretized as the discrete features are closer to a knowledge-level representation than the continuous ones and also the decision rules generated from discrete features are easier to understand for both users and experts. Then the optimal cut points are obtained for each condition attributes and with respect to these cut points, the real valued condition attributes are discretized. These optimal cut points for our data set are presented in Table 3 and the data set is denoted as the optimal cut set (OCS).

Table 3. Optimum Cut Set (OCS)

Cut points for the condition attributes				
a1	a2	a3	a4	a5
0.0673	0.0538	0.0274	0.1342	0.0076
0.0979	0.1684	---	0.1754	---
0.1568	---	---	0.2212	---
0.7294	---	---	0.2860	---

The real valued condition attributes are then replaced by discrete numerical values using the optimal cut point. The discretized data set (dIS) is shown in the Table 4.

Table 4. The Discretized Data Set (*dIS*)

Objects	Attributes					
	Sensors response					Tea category
	a1	a2	a3	a4	a5	D
1	1	1	1	0	0	1
…	…	…	…	…	…	…
194	3	1	1	4	1	7

The discretized data set (dIS) is then tested to find the presence of any superfluous condition attribute i.e. if there are any redundant sensor in the sensor array of the electronic nose. With our dataset, the algorithm doesn't find presence of such attributes. This implies that all the sensors in the electronic nose sensor array have certain contributions for the decision making.

For classification, we employ the rule based approach. The classification accuracy is validated using the method of 10-fold cross validation [21], [22] where 90% of data constitute the training set and remaining 10% form the test set, and the data-set is folded ten times. In each fold, the training set is used to produce the optimum rule-set using the concept of rule generation These rules are used to classify the corresponding test set. Detailed results of classification using 10-fold cross-validation are presented in the Table 5, where the overall classification accuracy is obtained as 85.09 % with standard deviation as 7.88 %.

Table 5. Detailed Results of 10-Fold Cross-Validation

Fold no.	No of data in training set	No of data in testing set	No of rules generated	No of data classified	No of data misclassified	Classification accuracy (%)
1	175	19	26	15	4	78.94
2	175	19	27	16	3	84.21
3	175	19	25	13	6	68.42
4	175	19	26	16	3	84.21
5	175	19	26	18	1	94.74
6	175	19	26	17	2	89.47
7	175	19	27	18	1	94.74
8	175	19	27	17	2	89.47
9	173	21	26	18	3	85.71
10	173	21	26	17	4	80.95
		Overall classification accuracy				85.09
		Standard deviation				7.88

5 Conclusion

In this paper, an attempt has been made to classify black tea quality from multi-sensor data patterns of an electronic nose using a rough set based classifier. The rough set based method is very useful in handling the vagueness and uncertainty in data, which is very common in machine olfaction. Another uniqueness of this method lies in identifying the redundant attributes or sensors. While all the other methods of sensor array optimization require separate procedures, the rough set based classifier has this feature integrated in it, which effectively increases the accuracy of classification. But as sample collection for tea is difficult, the results presented do not show very good accuracy with a small dataset. There is another important feature of the rough set based method, which is used to filter ambiguous training patterns. Due to small size of the data set, this feature could not be demonstrated. With a large data set, both the features could be utilized and that would result in the increase of classification accuracy. All in all, the method proposed in this paper has very useful features and is likely to be extremely useful for other electronic nose applications.

References

[1] Peris, M., Escuder-Gilabert, L.: A 21st century technique for food control: Electronic noses. Analytica Chimica Acta 638(1), 1–15 (2009)

[2] Guo, D., Zhang, D., Li, N., Zhang, L., Yang, J.: A novel breath analysis system based on electronic olfaction. IEEE Transactions on Biomedical Engineering 57(11), art. no. 5523940, 2753–2763 (2010)

[3] Capua, E., Cao, R., Sukenik, C.N., Naaman, R.: Detection of triacetone triperoxide (TATP) with an array of sensors based on non-specific interactions. Sensors and Actuators, B: Chemical 140(1), 122–127 (2009)

[4] Dutta, R., Hines, E.L., Gardner, J.W., Kashwan, K.R., Bhuyan, M.: Tea quality prediction using a tin oxide-based electronic nose: An artificial intelligence approach. Sens. Actuators B: Chem. 94, 228–237 (2003)

[5] Bhattacharyya, N., Bandyopadhyay, R., Bhyan, M., Ghosh, A., Mudi, R.K.: Correlation of multi-sensor array data with 'tasters' panel evaluation for objective assessment of black tea flavour. In: Proc. ISOEN, Barcelona, Spain (2005)

[6] Bhattacharyya, N., Bandyopadhyay, R., Bhuyan, M., Tudu, B., Ghosh, D., Jana, A.: Electronic nose for black tea classification and correlation of measurements with "Tea Taster" marks. IEEE Trans. Instrum. Meas. 57, 1313–1321 (2008)

[7] Bhattacharyya, N., Seth, S., Tudu, B., Tamuly, P., Jana, A., Ghosh, D., Bandyopadhyay, R., Bhuyan, M., Sabhapandit, S.: Detection of optimum fermentation time for black tea manufacturing using electronic nose. Sens. Actuators B, Chem. 122(2), 627–634 (2007)

[8] Tudu, B., Metla, A., Das, B., Bhattacharyya, N., Jana, A., Ghosh, D., Bandyopadhyay, R.: Towards Versatile Electronic Nose Pattern Classifier for Black Tea Quality Evaluation: An Incremental Fuzzy Approach. IEEE Trans. Instrum. Meas. 58(9), 3069–3078 (2009)

[9] Kermani, B.G., Schiffman, S.S., Nagle, H.T.: A novel method for reducing the dimensionality in a sensor array. IEEE Trans. Instrum. Meas. 47(3), 728–741 (1998)

[10] Elkov, T., Martensson, P., Lundstrom, I.: Selection of variables for interpreting multivariate gas sensor data. Anal. Chim. Acta 381, 221–232 (1999)

[11] Pawlak, Z.: Rough set theory and its applications to data analysis. Cybernetics and Systems: An Int. J. 29, 661–688 (1998)

[12] Pawlak, Z.: Some Issues on Rough Sets. In: Peters, J.F., Skowron, A., Grzymała-Busse, J.W., Kostek, B., Świniarski, R.W., Szczuka, M.S. (eds.) Transactions on Rough Sets I. LNCS, vol. 3100, pp. 1–58. Springer, Heidelberg (2004)

[13] Komorowski, J., Polkowski, L., Skowron, A.: Rough Sets: A Tutorial, Rough Fuzzy Hybridization, pp. 3–98. Springer (1999)

[14] Nguyen, S.H., Nguyen, H.S.: Pattern extraction from data. Fundamental Informaticae 34, 129–144 (1998)

[15] Hussain, F., Liu, H., Tan, C.L., Dash, M.: Discretization: An enabling technique. Data Min. Knowl. Dis. 6, 393–423 (2002)

[16] Dai, J.-H., Li, Y.-X.: Study on discretization based on rough set theory. In: Proc. of the First International Conference on Machine Learning and Cybernetics, Beijing, pp. 1371–1373 (November 2002)

[17] Yang, P., Li, J., Huang, Y.: An attribute reduction algorithm by rough set based on binary discernibility matrix. In: Proc. of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery, pp. 276–280 (2008)

[18] Li, J., Pattaraintakorn, P., Cercone, N.: Rule Evaluations, Attributes, and Rough Sets: Extension and a Case Study. In: Peters, J.F., Skowron, A., Düntsch, I., Grzymała-Busse, J.W., Orłowska, E., Polkowski, L. (eds.) Transactions on Rough Sets VI. LNCS, vol. 4374, pp. 152–171. Springer, Heidelberg (2007)

[19] Kovacs, E., Ignat, I.: Reduct equivalent rule induction based on rough set theory. In: Proc. IEEE 3rd International Conference on Intelligent Computer Communication and Processing, pp. 9–15 (2007)

[20] Bag, A.K., Tudu, B., Roy, J., Bhattacharyya, N., Bandyopadhyay, R.: Optimization of sensor array in electronic nose: a rough set-based approach. IEEE Sensors Journal 11, 3000–3008 (2011)

[21] Rodriguez, J.D., Perez, A., Lozano, J.A.: Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Trans. Pattern Anal. Mach. Intel. 32(3), 569–575 (2010)

[22] Singh, S., Hines, E.L., Gardner, J.W.: Fuzzy neural computing of coffee and tainted-water data from an electronic nose. Sens. Actuators B 30(3), 185–190 (1996)

Hybrid Technique for Effective Knowledge Representation

Poonam Tanwar[1], T.V. Prasad[2], and Kamlesh Datta[3]

[1] Dept. of CSE, Lingaya's University,
Faridabad, Haryana, India & PhD Scholar,
Uttarakhand Technical University, Dehradun, Uttarakhand, India
poonam.tanwar@rediffmail.com
[2] Lingaya's University, Faridabad, Haryana, India
tvprasad2002@yahoo.com
[3] National Institute of Technology,
Hamirpur, Himachal Pradesh, India
kdnith@gmail.com

Abstract. Knowledge representation and inference mechanism are most desirable thing to make the system intelligent. System is known to an intelligent if its intelligence is equivalent to the intelligence of human being for a particular domain or general. Because of incomplete ambiguous and uncertain information the task of making intelligent system is very difficult. The objective of this paper is to present the knowledge base system architecture integrated with hybrid knowledge representation technique for making the system effective.

Keywords: Knowledge Representation (KR), Semantic Net, Script.

1 Introduction

1.1 Knowledge Representation

In AI system implementation, efficiency, speed and maintenance are the major things affected by the knowledge representation. A KB structure must be capable of representing the broad spectrum of knowledge types categorized by Feigenbaum include [5].

- Objects - information on physical objects and concepts
- Events - time-dependent actions and events that may indicate cause and effect relationships.
- Performance – procedure or process of performing tasks
- Meta-knowledge – knowledge about knowledge including its reliability, importance, performance evaluation of cognitive processors.

Many of the problems in AI require extensive knowledge about the world. Objects, properties, categories and relations between objects, situations, events, states and

N. Meghanathan et al. (Eds.): Advances in Computing & Inf. Technology, AISC 178, pp. 33–43.
springerlink.com

© Springer-Verlag Berlin Heidelberg 2013

time, causes and effects are the things that AI needs to represents. Knowledge representation provides the way to represent all the above defined things [38].An overview of various types of knowledge representation techniques are given below.

1.2 Semantic Net

A semantic network is widely used knowledge representation technique. Semantic network is a KR technique in which the relationship between class and objects are represented by the connection/link between objects or class of objects.

The nodes / vertices in semantic net are used to represent the Generic class or a particular class or an instance of a class (object).Relation between them is represented by the link, which shows the activation comes from where .The links are unidirectional .these links represents the semantic relationship between the objects. Semantic network are generally used to represent the inheritable knowledge. Inheritance is most useful form of inference. Inheritance is the belongings in which element of some class inherit the attribute and values from some other class shown in Fig.1 [38].

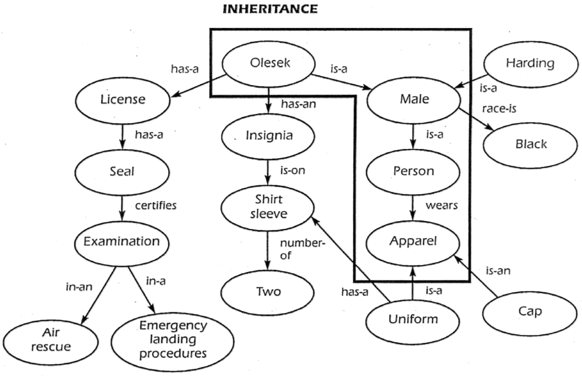


Fig. 1. Represents the inheritance relation [35][38].

Because there is an association between two or more nodes the Semantic nets are also known as associative nets. These associations are proved to be useful for inferring some knowledge from the existing one. If user wants to get any knowledge from the knowledge base they need not to put any query. The activated association or relation provides the result directly or indirectly only need to follow the links in the semantic net. IS-A, and A-KIND-OF are generally used to represent the value of a link in semantic net shown in fig 2.

KR techniques are divided in to two main categories one is declarative and other is procedural. Semantic net is a declarative KR technique that can be used either to represent knowledge or to support automated systems for reasoning about knowledge. Semantic net can be used in variety of ways, as per the requirement following are six of the most common kinds of semantic networks.

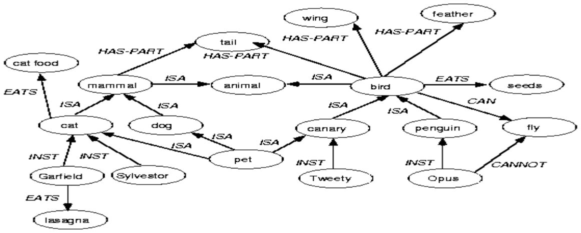


Fig. 2. Represents of IS-A, HAS, INSTANCE [17], [38]

1. Definitional networks
2. Assertional networks
3. Implicational networks
4. Executable networks
5. Learning networks
6. Hybrid networks

During 1975 (See Walker) Partitioned semantic net came in picture for speech understanding system. Then after that in 1977 Hendrix explained how we can expend the utility of semantic net using partitioned semantic net [8].In case of a huge network semantic net can be divided in to two more net. The semantic net is to be partitioned to separate the various nodes and arcs in to units and each unit is known as spaces. Using partitioned semantic net user can define the existence of the entity. One space is assigned to every node and arc and all nodes and arcs lying in the same space are distinguishable from those of other spaces. Nodes and arcs of different spaces may be linked, but the linkage must pass through the boundaries which separate one space from another [38].

Partitioning semantic nets can be used to delimit the scopes of quantified variables. While working with quantified statements, it will be help full to represent the pieces of information consist some event .For ex "Poonam believes that earth is round " is represented by the fig 3. Nodes<POONAM>' is an agent of Event node.<EARTH>' and <ROUND> represent the objects of space1.

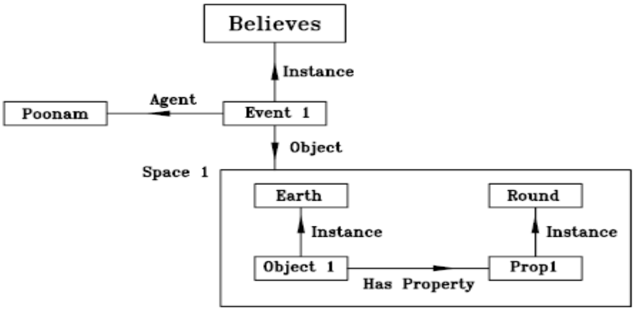


Fig. 3. Partitioned Semantic Net [38]

Universal and existential quantifier can also be represent by the Partitioning semantic net. For ex, “Every sister knots the rakhee to her brother" in predicate logic. In predicate logic the sister S and rakhee R are represented as objects while the knot event is expressed by a predicate where as in case of semantic net the event is represented as an object of some complex object, i.e., the bite event is a situation which could be the object of some more complex event. Partitioning semantic net can also be used to represent universal quantifier. For ex “Every sister knots the rakhee to her brother" is represented in fig 4 [38]. Partitioning semantic net can also be used for complex quantifleations which involve nested scopes by using nesting space.

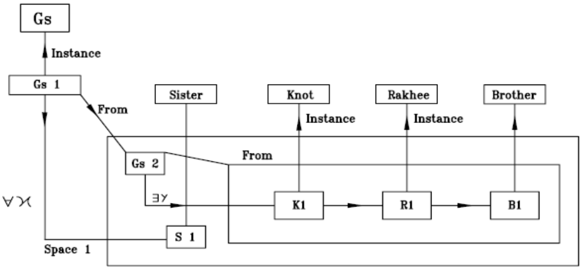


Fig. 4. Represents Partitioned Semantic Net for Quantifiers [38]

1.3 Frames

Frame can be considered as an extension to the semantic net. As we know that Semantic net is a graphical representation of knowledge as the knowledge increases the graph becomes complicated ie complexity of the system is directly proportional to the knowledge required for the problem domain. Then the Frame (KR) is the best way of representing the knowledge. A frame is a collection of attributes or slots and their associated values which describe the real world entity. An example of a Class frame is given in Fig 5 [38]. The frame is used to represent the following:

- a class which represents a set,
- an instance which represents an element of a class.
- Frame has three main components
- frame name
- Attributes (slots)
- Values (fillers: list of values, range, string, etc.)

There are two different naming system for frame first is its true name that uniquely describe the frame and second it can have any number of public names. Public names are values stored in the name slot of the frame. For instance, Frame frame-30 will look as:

name: ("women")
sex: (frame-3)
spouse: (frame-31)

child: (frame-29 frame-31) here frame 30 is the true name that refer it uniquely. True names are the pointers from one frame to another that actually represent the structure of the knowledge base. Public names are for communication with other agents [38][8].

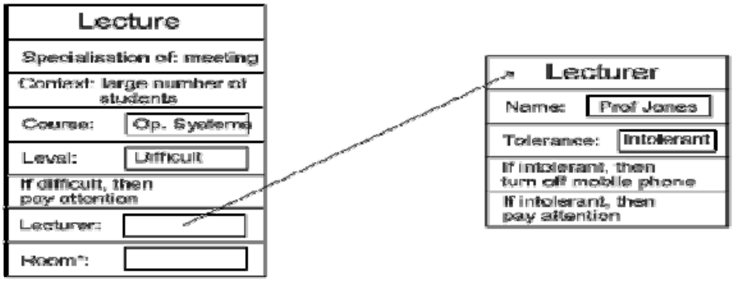

Lecture
Specialisation of: meeting
Context: large number of students
Course:
Op. Systems
Level:
Difficult
If difficult, then pay attention
Lecturer:
Room*:
Lecturer
Name:
Prof Jones
Tolerance:
Intolerant
If intolerant, then turn off mobile phone
If intolerant, then pay attention

Fig. 5. Frame Knowledge representation technique [27]

The advantage of a frame based knowledge representation is that there is no need to search the entire knowledge-base because the objects related to a frame can be easily accessed directly looking in a slot of the frame. In 1993 Christian Rathke presented the language Frame Talk for developing the frame [12].

1.4 Conceptual Dependency (CD)

Conceptual Dependency (CD) Theory was developed by Roger Schank in 1973 to represent the knowledge acquired from natural language input. CD KR technique is used to represents the Sentences (knowledge) in sequential diagram which represents the actual action using the real situation and concept. CD representation provides the sets of primitive actions, different types of states, and different theories of inference. The agent and the objects are represented. Basically CD is a theory of how to represent sentences as shown in Fig.6.1and Fig 6.2. It may have two axioms [40]:

- Sentences that have similar sense/meaning could be represented by single representation.
- Implicit information can be made explicit in the representation.
- CD provides [36]:
- a structure into which nodes representing information can be placed
- a specific set of primitives from which meaning is built

ACT	action
PP	objects (picture producers)
AA	modifiers of actions (action aiders)
PA	modifiers of objects (picture aiders)

- at a given level of granularity.

Examples of Primitive Actions are:

ATRANS (Transfer of an abstract relationship. e.g. take.)
PTRANS (Transfer of the physical location of an object. e.g. jamp.)
PROPEL (Application of a physical force to an object. e.g. pull).
MTRANS (Transfer of mental information. e.g. ask).
MBUILD (Construct new information from old. e.g. decide).
SPEAK (Utter a sound. e.g. say).
ATTEND (Focus a sense on a stimulus. e.g. listen, watch).
MOVE (Physical Movement of a body part e.g. hit, throw.
GRASP (Actor grasping an object. e.g. clutch).
INGEST (Actor ingesting an object. e.g. eat).
EXPEL (Actor getting relieve of an object from body eg. Ram , Shayam).

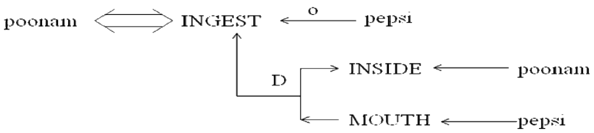


Fig. 6.1. CD Representation for" Poonam drink the pepsi"

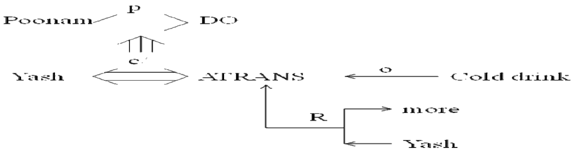


Fig. 6.2. CD Representation for" Poonam prohibited Yash to drinking more cold drink"

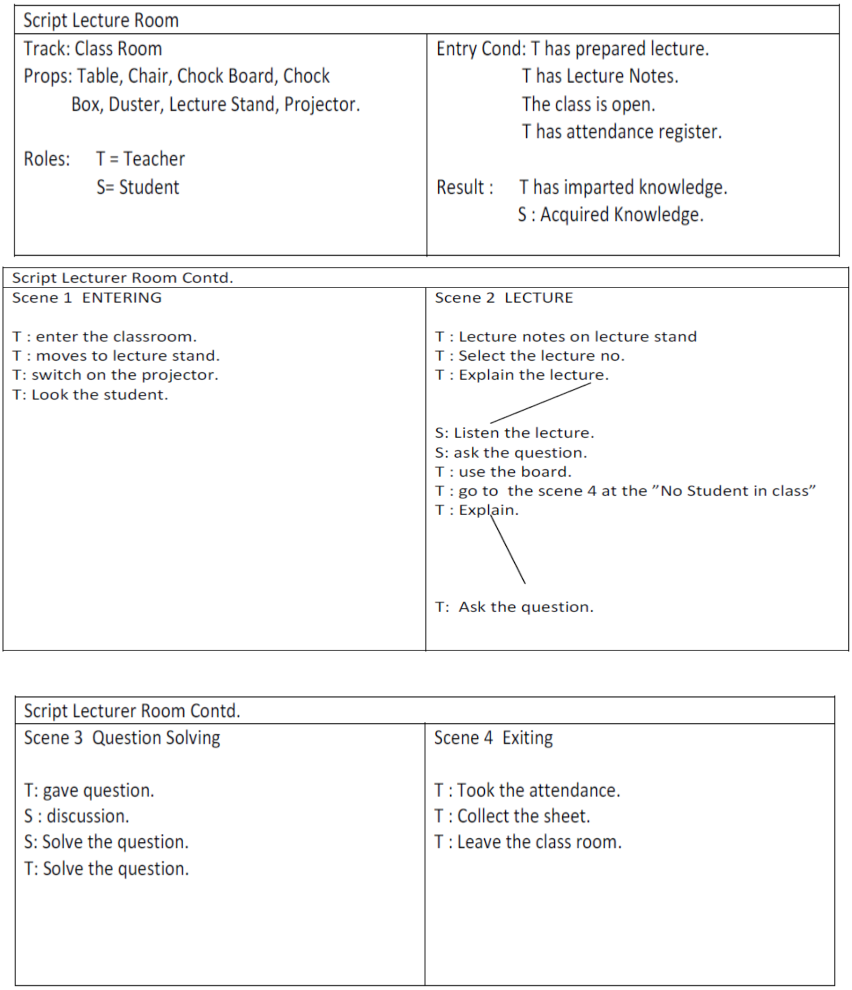
1.5 Scripts

A variation in the theme of structured objects called scripts was devised by Roger Schank and his associates in 1973[3].It is an active type information which contain class of events in terms of contexts, participants and sub-events represented in the form of collection of slots or series of frames which uses inheritance and slots . Scripts predict unobserved events and can build coherent account from disjointed observations. Scripts basically describes the stereotypical knowledge i.e if the system in not given the information dynamically then it assumes the default information to be true Scripts are beneficial because real world events do follow stereotyped patterns as human beings use previous experiences to understand verbal accounts. A script is used for organizing the knowledge as it directs the attention and recalls the inference. They provide knowledge and expectations about specific events or experiences and can be applied to new situations. For example: "Rohan went to the restaurant and had some pastries". it was good now meaning derived from the above text one gets to know he got the pastries from the restaurant and that for eating and that was good.

Script defines an episode with the known behavior and describes the sequence of events. The script consist the following.

- Current plans (Entry condition, Result)
- Social link(Track)
- Played roles,
- Scene.
- Probs.
- Anything indicating the behavior of the script in a given situation.

An example of script for class room is shown in fig.7.



Script Lecture Room	
Track: Class Room Props: Table, Chair, Chock Board, Chock Box, Duster, Lecture Stand, Projector. Roles: T = Teacher S= Student	Entry Cond: T has prepared lecture. T has Lecture Notes. The class is open. T has attendance register. Result : T has imparted knowledge. S : Acquired Knowledge.

Script Lecturer Room Contd.	
Scene 1 ENTERING T : enter the classroom. T : moves to lecture stand. T: switch on the projector. T: Look the student.	Scene 2 LECTURE T : Lecture notes on lecture stand T : Select the lecture no. T : Explain the lecture. S: Listen the lecture. S: ask the question. T : use the board. T : go to the scene 4 at the "No Student in class" T : Explain. T: Ask the question.

Script Lecturer Room Contd.	
Scene 3 Question Solving T: gave question. S : discussion. S: Solve the question. T: Solve the question.	Scene 4 Exiting T : Took the attendance. T : Collect the sheet. T : Leave the class room.

Fig. 7. Script structure for class room

Advantages of using scripts:

- Details for a particular object remain open and
- Reduces the search space.

Disadvantages

- Less general than Frames
- It may not be suitable for all kind of Knowledge

2 Hybrid Knowledge Representation Technique

The KR system must be able to represent any type of knowledge, "Syntactic, Semantic, logical, Presupposition, Understanding ill formed input, Ellipsis, Case Constraints, Vagueness". In our previous paper we have proposed the model for effective knowledge representation technique that consist five different parts the K Box, Knowledge Base, Query applier, reasoning and user interface as shown in fig 8. This time the total emphasis is on knowledge representation. This section used to describe the new hybrid knowledge representation technique which is the integration of script and semantic net KR technique.

Every knowledge representation technique has their own merits and demerits that depend on which type of knowledge we want to represent. To navigate the problem associated with single knowledge representation technique the hybrid knowledge representation came in picture.

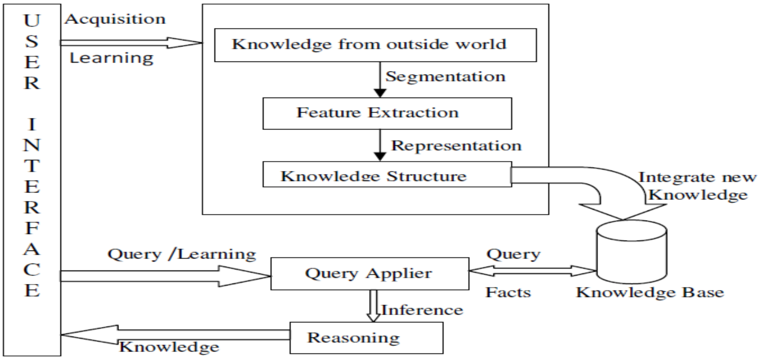
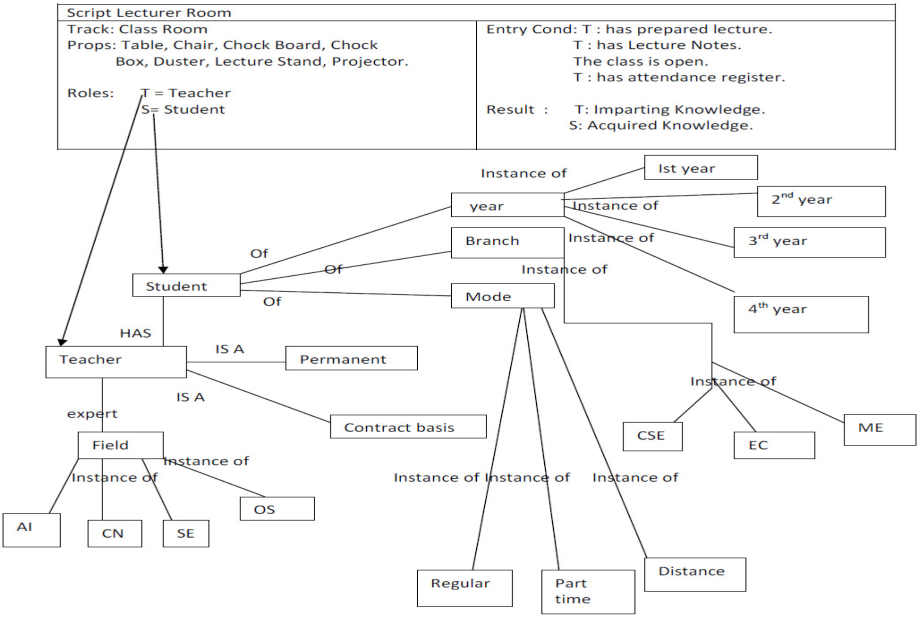


Fig. 8. Knowledge Base System Model /Architecture [39]

The script and semantic net alone is a strong representation technique but still they have some disadvantages. The previous section consist the example of script for lecture room using that we are unable to get the detail like the teacher can teach one or more subject, Is a permanent or on contract basis ,student is a regular student or part time. Student opted one or many subject. Whereas using semantic net we can't represent the knowledge scene wise. Semantic net can't be use to represent the knowledge event by event. So to get all the knowledge from the system, integrated knowledge representation technique is used. The hybrid structure is shown in fig 9. From script to semantic net two different directional link coming out that shows the link between the roles of script with the two different classes of semantic net. In the same way we can make the link between other roles and objects involve in scripts (scene wise) with the class and object in the semantic net. The unnamed link in semantic net shows the generalization for eg. Mode can be part time, full time and regular.



Script Lecturer Room
Track: Class Room
Props: Table, Chair, Chock Board, Chock Box, Duster, Lecture Stand, Projector.
Roles: T = Teacher
S= Student
Entry Cond: T : has prepared lecture.
T : has Lecture Notes.
The class is open.
T : has attendance register.
Result : T: Imparting Knowledge.
S: Acquired Knowledge.

Fig. 9. Hybrid Knowledge Representation technique

2.1 Strength of Hybrid Knowledge Representation Technique

Human beings use past/previous learning & senses to understand verbal communication and in actual real world events do follow stereotyped patters. Communication style of each one is different from other and it is quite often when relating events, do leave large amount of blanks/gaps or assumed details out of their communication. This may lead to miscommunication. In real life it is not easy to deal with a system that are not able to fill up the missing conversational features. Whereas scripts can predict/ assume unobserved events. Scripts can fill the gaps created from incomplete/disjoined observations and can build a sequential information. Semantic net is best knowledge representation technique for representing non event based knowledge with its technical simplicity. Even non technology savvy can also extract information/ knowledge from the semantic net.

3 Conclusion

There are various knowledge representation schemes in AI. All KR techniques have their own semantics, structure as well as different control mechanism and power. Combination of two or more representation scheme may be used for making the system more efficient and improving the knowledge representation. We are trying to build the intelligent system that can learn itself by the query and have a power full mechanism for representation and inference. The semantic net and script are very powerful techniques in some respects so the aim is to take the advantage of these techniques under one umbrella.

References

[1] Sowa, J.F.: Encyclopedia of Artificial Intelligence, 2nd edn. Wiley (1992)

[2] Rich, E., Knight, K.: Artificial Intelligence, 2nd edn. McGraw-Hill (1991)

[3] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edn. Prentice Hall (2009)

[4] Davis, R., Shrobe, H., Szolovits, P.: What is a Knowledge Representation? AI Magazine 14(1), 17–33 (1993)

[5] Brachman, R., Levesque, H. (eds.): Readings in Knowledge Representation. Morgan Kaufman (1985)

[6] Stillings, L.: Knowledge Representation, Ch. 4 and 5 (1994), http://www.acm.org/crossroads/.www.hbcse.tifr.res.in/jrmcont/notespart1/node28.html

[7] Houben, G.J.P.M.: Knowledge representation and reasoning. Dutch Research Database (Period 01/2002)

[8] Frost, R.A.: A Method of Facilitating the Interface of Knowledge Base System Components. Computer Journal 28(2), 112–116 (1985)

[9] Sharif, A.M.: Knowledge representation within information systems in manufacturing environments. Brunel University Research Archive (2004)

[10] Brewster, C., O'Hara, K., Fuller, S., Wilks, Y., Franconi, E., Musen, M.A., Ellman, J., Buckingham Shum, S.: Knowledge representation with ontologies: the present and future. IEEE Intelligent Systems, 72–81 (2004) ISSN 1541-1672

[11] Allen, J., Ferguson, G., Gildea, D., Kautz, H., Schubert, L.: Artificial Intelligence, Natural Language Understanding, and Knowledge Representation and Reasoning, 2nd edn. Benjamin Cummings (1994)

[12] Ali, S.S., Iwanska, L.: Knowledge representation for natural language processing in implemented system. Natural Language Engineering 3, 97–101(1997)

[13] Morgenstern, L.: Knowledge Representation. Columbia University (1999), http://wwwformal.stanford.edu/leora/krcourse/

[14] Reichgelt, H.: Knowledge Representation: An AI Perspective, Chapter 5 (Semantic Networks) and Chapter 6 (Frames)

[15] van Harmelen, F.: Knowledge Representation and Reasoning. Vrije Universitetit Amsterdam, http://www.cs.vu.nl/en/sec/ai/kr

[16] Kuechler Jr., W.L., Lim, N., Vaishnavi, V.K.: A smart object approach to hybrid knowledge representation and reasoning strategies. In: Hawaii International Conference on System Sciences, HICSS 1995 (1995)

[17] Shetty, R.T.N., Riccio, P.-M., Quinqueton, J.: Hybrid Model for Knowledge Representation. In: 2006 International Conference on, vol. 1, pp. 355–361 (2006)

[18] Chi, X., Haojun, M., Zhen, Z., Yinghong, P.: Research on hybrid expert system application to blanking technology, National Die and Mold CAD Engineering Research Center. Shanghai Jiao Tong University, Shanghai 200030, PR China (1999)

[19] Quesgen, W., Junker, U., Voss, A.: Constraints in Hybrid Knowledge Representation System. Expert Systems Research Group, F.R.G., http://dli.iiit.ac.in/ijcai/IJCAI-87-VOL1/PDF/006.pdf

[20] Rathke, C.: Object-oriented programming and frame-based knowledge representation. In: 5th International Conference, Boston (1993)

[21] Hendrix, G.G.: Expanding the Utility of Semantic Networks through Partitioning. In: Artificial Intelligence Center, Stanford Research institute Menlo Park, California 94025

[22] Lehmann, F.: Semantic networks, Parsons Avenue, Webster Groves, Missouri, U.S.A.

[23] Gow, J.: Lecture notes, Imperial College, London, http://www.doc.ic.ac.uk/~sgc/teaching/v231/lecture4.ppt

[24] Lee, T.B.: Chapter on "Semantic web road map" (1998), http://www.w3.org

[25] Khatib, W.: Semantic modeling and knowledge representation in Multimedia (1999), http://ieeeexlore.ieee.org

[26] Lecture notes, http://www.cs.odu.edu/~toida/nerzic/content/logic/pred_logic/inference/infer_intro

[27] Presentation on "Knowledge representation", http://www.doc.ic.ac.uk/~sgc/teaching/v231/lecture4.ppt

[28] Presentation on "Knowledge representation techniques", http://www.scribd.com/doc/6141974/semantic-networks-standardisation

[29] Web document on "Predicate logic history", http://www.cs.bham.ac.uk/research/projects/poplog/thought/chap6/node5.html

[30] Web document on "Introduction to Universal semantic net", http://sempl.net/

[31] Lecture notes on "knowledge representation misc psychology and languages for knowledge representation", http://misc.thefull-wiki.org/Knowledge_representation

[32] Lecture notes on frame knowledge representation technique, http://userweb.cs.utexas.edu/users/qr/algy/algy-expsys/node6.html

[33] Presentation on "Knowledge representation using structured objects", http://www.freshtea.files.wordpress.com/2009/../5-knowledge-representation.ppt

[34] Jeng, S.-K.: Lecture notes on "Knowledge representation", http://www.cc.ee.ntu.edu.tw/~skjeng/Representation.ppt

[35] Presentation on "Knowledge representation and rule based systems", http://www.arun555mahara.files.wordpress.com/2010/02/knowledge-representation.ppt

[36] Presentation on "Various knowledge representation techniques", http://www.ee.pdx.edu/~mperkows/CLASS_ROBOTICS/FEBR,19/019.representa.ppt

[37] PPT and lecture notes, http://people.dbmi.columbia.edu/homepages/wandong/KR/krglossary.html

[38] Tanwar, P., Prasad, T.V., Aswal, M.S.: Comparative Study of Three Declarative Knowledge Representation Techniques. International Journal on Computer Science and Engineering 02(07), 2274–2281 (2010)

[39] Tanwar, P., Prasad, T.V., Datta, K.: An Effective Knowledge base system Architecture and issues in representation techniques. International Journal of Advancements in Technology, http://ijict.org/, ISSN 0976-4860

[40] Lecturer notes on Knowledge Representation, http://www.scribd.com/doc/13599253/Knowledge-Representation

A Language Independent Approach to Develop Urdu Stemmer

Mohd. Shahid Husain[1], Faiyaz Ahamad[2], and Saba Khalid[3]

[1] Department of Information Technology, Integral University, Lucknow, India
siddiquisahil@gmail.com
[2] Department of Computer Science & Engineering, Integral University, Lucknow, India
faiyaz.ahamad@yahoo.com
[3] Department of Computer Science & Engineering, Integral University, Lucknow, India
sksabask@gmail.com

Abstract. Especially, during last few years, a wide range of information in Indian regional languages like Hindi, Urdu, Bengali, Tamil and Telugu has been made available on web in the form of e-data. But the access to these data repositories is very low because the efficient search engines/retrieval systems supporting these languages are very limited. Hence automatic information processing and retrieval is become an urgent requirement. This paper presents an unsupervised approach for the development of an Urdu stemmer. To train the system a training dataset, taken from CRULP [22], consists of 111,887 words is used. For generating suffix rules two different approaches, namely, frequency based stripping and length based stripping have been proposed. The evaluation has been made on 1200 words extracted from the Emille corpus. The experiment results shows that these are very efficient algorithms having accuracy of 85.36% and 79.76%.

Keywords: Stemmer, Morphological Analysis, Information Retrieval, Unsupervised Stemming.

1 Introduction

The use of digital technologies and growth in technological developments for storing, manipulating and accessing of information has led to development of valuable information repositories on the internet. The rapid growth of electronic data has attracted the attention in the research and industry communities for efficient methods for indexing, analysis and retrieval of information from this high volume of data repositories for a vast domain of applications.

Stemming is the backbone process of any IR system. Stemmers are used for getting base or root form (i.e. stems) from inflected (or sometimes derived) words. Unlike morphological analyzer, where the root words have some lexical meaning, it's not necessary with the case of a stemmer. Stemming is used to reduce the overhead of indexing and to improve the performance of an IR system. Stemming is the basic process of any query system, because a user who needs some information on آخری may also be interested in documents that contain the word آخر (without the ی).

N. Meghanathan et al. (Eds.): Advances in Computing & Inf. Technology, AISC 178, pp. 45–53.
springerlink.com

© Springer-Verlag Berlin Heidelberg 2013

The approaches used for developing a stemmer can be broadly classified as Rule-based (knowledge-based) and machine learning (supervised and unsupervised) approaches. A rule-based stemmer makes use of linguistic knowledge to develop rules for stemming. Besides being language specific it is very difficult and time consuming to obtain such rules. Specifically, for languages like Urdu, which is a very highly inflectional language, the task becomes quite cumbersome. Supervised learning is an alternative approach to frame stemming rules. In order to learn suffixes this approach uses set of inflection-root pair of words which are manually segmented. But this algorithm is also not produce very effective results for Urdu language as it is highly inflectional language and this becomes a complex task. Manually segmenting the Urdu words is a very time-consuming task and is not feasible because in Urdu for a root word there are many inflections. It also requires a very good linguistic knowledge to segment words and get the root and the inflections. For designing stemmer for Urdu language we have used unsupervised stemming approach. This approach does not require any specific knowledge of the language in case. It uses a set of words (training dataset) to learn suffixes. As the approach used in this work is language independent, it can be easily used for the development of the stemmers of other languages as well. For suffix rule generation two different approaches have been discussed. First is the Length based approach which is very simple suffix stripping approach. The second is Frequency based approach. The experiment results shows that the second approach used, gives the more accurate results. The rest of the paper is organized as follows:

Section 2 reviews the earlier work done in morphological analysis and stemming for Indian languages. Section 3 gives a brief idea about the proposed approach. Section 4 presents the detail of experimental setup. Section 5 discusses the important results and observations and finally conclusion have been made in section 6.

2 Related Works

The most basic component of any Information Retrieval system is Stemmers. Among all the morphological systems, stemmers are the simplest system.

Earlier stemmers were designed on rule-based approach. Julie Lovins published the first paper on rule-based stemming in the year 1968. The approach used by Lovins was Iterative Longest Match heuristic. The most noteworthy work in the field of rule based stemmer was presented by Martin Porter in 1980 [9]. He simplified the rules of Lovin to about 60 rules. To access information available in English or some other European languages there are number of efficient IR systems. Work involving development of IR systems for Asian languages is only of recent interests. Development of such systems is constraint by the lack of the availability of linguistic resources and tools in these languages. Until recently, For Indian regional languages the work done by IR community involves languages like Hindi, Bengali, Marathi, Tamil and Oriya. But there is no reported work done for Urdu language. Although, as per our knowledge there is no reported works done by the IR community to efficiently retrieve the information available on net in Urdu language, however, a lot of research has been done towards computational morphological analysis and stemming of Urdu. Computational analysis of different parts of speech in Urdu is described by Rizvi [1] and Butt [2]. To stem French words in a corpus a dictionary-based approach is used [3].

Various researches have been done on Arabic and Farsi stemmers, most of them uses statistical and heuristics based approaches [4, 5]. Although the writing script of Urdu is similar (not the same) to Farsi and Arabic, stemmers used for those languages (Arabic and Farsi) are not adequate for stemming Urdu words because of these reasons:

- Stemmers used for Farsi language accurately stems only the Farsi loan words and produce a number of errors (incorrect stems) on native Urdu and Arabic loan words.
- Arabic language has high inflection and complex grammar. So stemmers used for Arabic language produces a large number of over-stemming and mis-stemming errors for Urdu.

Stemmers may be developed by using either rule based or statistical approaches. Rule-based stemmers require prior morphological knowledge of the language, while statistical stemmers use corpus to calculate the occurrences of stems and affixes. A rule-based stemmer is developed for English by Krovetz, using machine-readable dictionaries. Due to high dependency on dictionary the systems lacks consistency [8]. In Porter Stemmer, the algorithm enforces some terminating conditions of a stem. Until any of the conditions is achieved, it keeps on removing endings of the word iteratively [9]. To perform stemming of Arabic an approach using stop word list is proposed by Thabet. This algorithm gives accuracy of 99.6% for prefix stemming and 97% for postfix stemming [10]. Paik and Parui [11] have proposed an interesting stemming approach based on the general analysis of Indian languages. This technique is used for Bengali, Hindi and Marathi languages. For Persian language a rule based algorithm was proposed by Sharifloo and Shamsfard for stemming. The accuracy of this algorithm is 90.1 % [12]. Besides rule-based stemmers there are a number of statistical stemmers for different languages. These stemmers use some statistical analysis of the training data and then rules are derived from these analyses for stripping the inflected words to get the root word. Croft and Xu provide two methods for stemming i.e. Corpus-Specific Stemming and Query-Specific Stemming [13]. Kumar and Siddiqui propose an algorithm for Hindi stemmer. The algorithm achieves 89.9% accuracy [14]. An Urdu stemmer called Assas-Band, has been developed by Qurat-ul-Ain Akram, Asma Naseer, Sarmad Hussain using affix based exception lists, which increases accuracy up to 91.2% [16].

2.1 Language Challenges

The Indian regional languages are different from each other in orthography, morphology and character encoding aspects. Designing a stemmer for such languages is quite tough and hence designing a standard stemmer to support Indian regional languages is a quite complex job. For stemming purpose Urdu is a challenging language because of the following two reasons:

- Its Perso-Arabic script and second,
- Its morphological system having inherent grammatical forms and vocabulary of Arabic, Persian and the native languages of South Asia.

It is estimated there are about there are around 490 million speakers of Urdu around the world [18]. According to George Weber's article Top Languages: The World's 10 Most Influential Languages in Language Today, Hindi/Urdu is the fourth most spoken language in the world, with 4.7 percent of the world's population [19]. Urdu is a composition of many languages and adopts words from other languages with ease. Although it has its own morphology, Urdu morphology is strongly influenced by Farsi (Persian), Arabic, and Turkish. Therefore, Urdu vocabulary is composed of the above mentioned languages along with many Sanskrit-based and English words. For example, the word pachim (Hindi) and Maghrib (Arabic) both mean the direction west in English and are both Urdu words as well. Urdu is rich in both inflectional and derivational morphology. Urdu verbs inflect to show agreement for number, gender, respect and case. In addition to these factors, verbs in Urdu also have different inflections for infinitive, past, non-past, habitual and imperative forms. All these forms (twenty in total) for a regular verb are duplicated for transitive and causative (ditransitive) forms, thus giving a total of more than sixty inflected variations. Urdu nouns also show agreement for number, gender and case. In addition, they show diminutive and vocative affixation. Moreover, the nouns show derivational changes into adjectives and nouns. Adjectives show similar agreement changes for number, gender and case. Urdu is a bi-directional language with an Arabic-based orthography. Bi-directional means that it is very common in Urdu to see an English word written in Latin-based characters. Sometimes an English word is written phonetically with Urdu characters (e.g. executive is written as ایگزیکٹو). Although Urdu has Arabic orthography, its grammar is based on Sanskrit and Persian. Urdu has gender marking on its parts of speech (e.g. paharh (mountain) and paharhi (hill). Therefore, stemming Urdu words will increase recall and also conserve on space usage of the indices.

Hindi and Urdu are considered one language for linguistic purposes. As Urdu is closely related to Hindi and it shares morphology, syntax and almost all phonology. Urdu shares its grammar with Hindi with only some differences in vocabulary, and writing style. Urdu is quite complex language because its morphology is a combination of many languages: Sanskrit, Arabic, Farsi, English and Turkish to name a few. This aspect of Urdu becomes quite a challenge while doing morphological analysis to build a stemmer. Urdu's descriptive power is quite high. This means that there could be many different ways a concept can be mentioned in Urdu and in many different forms. Urdu has a property of accepting lexical features and vocabulary from other languages, most notably English. This is called code switching in linguistics e.g. it is not uncommon to see a right to left flow interrupted by a word written in English (left to right) and then continuation of the flow right to left. For example, ہے laptop میرا وہ [That is my laptop].

3 Our Approach

Our proposed approach is based on n-gram splitting model. For learning purpose of the stemmer, documents from the Urdu Corpus available at CRULP are used. The words taken from these documents are split to get n split suffixes, using n gram model. Where n n=1, 2, 3…l, for word length l.

Then the frequency count of the split words is calculated to get the probability of the stem - suffixes pair extracted from the n-gram splitting.

Then we have calculated the optimal split probability, which is the multiplication of the stem probability and suffix probability. By observing the results, a particular frequency threshold was taken. The splits whose frequency count lies above this threshold value were considered as valid candidates and were used for suffix generation rules. Also the maximum split probability corresponds to the optimal split segments which are considered to be the valid candidate for framing suffix generation rule.

Table 1. Algorithmic steps

- Split words into n gram
- Generate stem and suffix list
- Sort suffixes on decreasing order of their frequency
- Generate suffix stripping rules
  i. using Frequency based stripping
  ii. using length based stripping

3.1 Word Splitting and Stem Classes Generation

In this step n-gram model is used to obtain corresponding stems and suffixes of a word Wy by splitting it into n-grams as given below

Wy: = {(stem1y|suffix1y); (stem2y|suffix2y); …… (stemxy|suffixxy)}

Where x, y=1, 2, 3… l (where l denotes the length of the word) and stemxy is the xth stem of yth word and suffixxy is the xth suffix of yth word.

For example, the word آئزلینڈ gives the following stem-suffix pairs after n-gram splitting:

آئزلینڈ := { (NULL -- آئزلینڈ); (آ-- ئزلینڈ); (آئ--زلینڈ); (آئز-- لینڈ); (آئزل-- ینڈ); (آئزلی-- نڈ); (آئزلین-- ڈ); (آئزلینڈ --NULL) }

Next a common stem class is used to group the words having common stems. To find common stems, maximum common prefix method is used.

For example the stem equivalence class for the words آخری and آخرکار Can be given as: آخری = :{آخر, آخرکار}

3.2 Generation of Stem and Suffixes

The longest common prefix method is used to obtain the correct stems and suffixes from the inflected words. We have used the stem equivalence class, generated in the first phase of the algorithm to find out the longest common prefixes. These prefixes are then stored as the stems and the remaining part of the word as the valid suffix along with its corresponding frequency count. This information is then used to frame rules for suffix stripping. The suffixes in the generated list having higher frequency are considered as valid suffixes for generating suffix stripping rule.

For example the common root word of different inflected words with their suffixes is stored as;

آخر: = {ی, کار}

3.3 Frequency Counting

In this step the frequency count of the suffixes generated in step 2 is calculated. This list of suffixes is then arranged in order of their count. By manual analyses of the system a frequency count is taken as the threshold. The suffixes having there frequency count below this threshold value are discarded and not considered for suffix rule generation while those lying above the preset threshold value are considered as the valid candidates for framing the suffix stripping rules.

3.4 Generation of Suffix Rules

In this step, two different approaches are used for the purpose of suffix stripping rule generation.

3.4.1 Length Based Suffix Stripping

This is the crudest method for suffix rule generation. In this approach, the suffix list obtained from step 2 is sorted according to their lengths in decreasing order. This approach is quite valid as it removes the suffix in a word which is of max length. The drawback of this approach is that in many cases over-stemming occurs.

3.4.2 Frequency Based Suffix Stripping

This is the simplest method for generating suffix stripping rule. The suffixes obtained in the second step, are sorted in descending order of their corresponding frequency counts. By manual observation a threshold value is being set. The suffixes having there frequency count below this threshold value are discarded and not considered for suffix rule generation while those lying above the preset threshold value are considered as the valid candidates for framing the suffix stripping rules. This method is quite effective for Urdu and other very highly inflectional languages because as they have very large number of suffixes.

4 Experiment

For the evaluation purpose of the proposed stemmer, following experiment was conducted. The parameter used to measure the performance of the stemmer is accuracy. The accuracy can be defined as the fraction of words stemmed correctly. Mathematically it can be stated as:

$$\text{Accuracy} = \frac{\text{Number of Correctly stemmed Words}}{\text{Total Number of Words}} \text{ x } 100$$

For testing of the stemmer a list of 1200 words, taken from Emille corpus, with their suffixes and stems is created manually. Then the developed system is used to get the stem of these words and cross checked with the list of manually stemmed words. The following table gives a summary about the statistics used for the evaluation of the stemmer.

Table 2. Data Set Specification

Training Dataset	D1	D2	D3
Count of words	50495	50836	10559
Count of Unique words	6428	6178	2492
Testing words	1200	1200	1200

To perform the evaluation of the proposed stemmer, the experiment is conducted in three runs. In Run1 Dataset D1 have been used for training, in Run2 Dataset D2 have been used for training and in Run3 Dataset D3 have been used for training. The statistics used for evaluation are shown in the following table.

Table 3. Experiment Specification

Run	R1	R2	R3
Training Dataset	D1	D2	D3
Testing Dataset	Test Dataset	Test Dataset	Test Dataset

Table 4. Results of the Experiments

Run	Accuracy of Implemented approach	
	Frequency based	Length based
R1	82.78	81.28
R2	85.36	77.85
R3	84.67	79.76

Table 4 shows the comparison between results obtained by using different methods.

5 Results and Discussions

It is clear from table 4, that the frequency based suffix generation approach gives the maximum accuracy of 84.27% whereas Length based suffix stripping algorithm gives maximum accuracy of 79.63%.

- The first approach that is length based approach is affected by over-stemming. For example the word گاڑیاں (automobiles) should be stemmed in گاڑی (automobile) but it stemmed it to گاڑ. Because یاں and اں are both suffixes and suffix of maximum length is removed so a part of the word is also removed as suffix.
- The second approach that is frequency based approach is affected by under-stemming. For example the word بیچنا should be stemmed into بیچ but the stemmer stemmed it into بیچن. Because ا and نا both are suffixes but as the frequency of ا is more the system removes this as suffix and return the remaining word as stem.

Moreover both the approaches discussed above are free from any language specific inputs and linguistic constraints. So these approaches can be used for other languages also.

Effect of stop words on stemming: when we have removed the stop words from the training dataset then there is some effect on the suffix list generated (the number of suffixes decreases by 2%), but there is no effect on stemming i.e. the result of stemmer is same after the stop word removal as it was before the stop word removal.

The stemmer is also very efficient for stemming English words transliterated in Urdu. For example اتھارٹیز, پرپوزل, سپلائرs

6 Conclusion and Future Work

The approach used in this work gives promising results for Urdu language. As the approach used is language independent it can be tested and implemented for other languages in near future.

As there is some problem of under stemming and over stemming in the used approaches. In future one can attempt to reduce these effects to improve the efficiency of the system.

As we know that stemmers have tremendous use in the Information Retrieval. We plan to make use of the designed stemmer for other related work of Information retrieval in case of Urdu language.

References

[1] Rizvi, J., et al.: Modeling case marking system of Urdu-Hindi languages by using semantic information. In: Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE 2005 (2005)

[2] Butt, M., King, T.: Non-Nominative Subjects in Urdu: A Computational Analysis. In: Proceedings of the International Symposium on Non-nominative Subjects, Tokyo, pp. 525–548 (December 2001)

[3] Savoy, J.: Stemming of French words based on grammatical categories. Journal of the American Society for Information Science 44(1), 1–9 (1993)

[4] Chen, A., Gey, F.: Building and Arabic Stemmer for Information Retrieval. In: Proceedings of the Text Retrieval Conference, p. 47 (2002)

[5] Mokhtaripour, A., Jahanpour, S.: Introduction to a New Farsi Stemmer. In: Proceedings of CIKM, Arlington, VA, USA, pp. 826–827 (2006)

[6] Wicentowski, R.: Multilingual Noise-Robust Supervised Morphological Analysis using the Word Frame Model. In: Proceedings of Seventh Meeting of the ACL Special Interest Group on Computational Phonology (SIGPHON), pp. 70–77 (2004)

[7] Rizvi, Hussain, M.: Analysis, Design and Implementation of Urdu Morphological Analyzer. In: SCONEST, pp. 1–7 (2005)

[8] Krovetz, R.: View Morphology as an Inference Process. In: The Proceedings of 5th International Conference on Research and Development in Information Retrieval (1993)

[9] Porter, M.: An Algorithm for Suffix Stripping. Program 14(3), 130–137 (1980)

[10] Thabet, N.: Stemming the Qur'an. In: The Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (2004)

[11] Paik, Pauri: A Simple Stemmer for Inflectional Languages. In: FIRE 2008 (2008)

[12] Sharifloo, A.A., Shamsfard, M.: A Bottom up Approach to Persian Stemming. In: IJCNLP (2008)

[13] Croft, Xu: Corpus-Based Stemming Using Co occurrence of Word Variants. ACM Transactions on Information Systems, 61–81 (1998)

[14] Kumar, A., Siddiqui, T.: An Unsupervised Hindi Stemmer with Heuristics Improvements. In: Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data (2008)

[15] Kumar, M.S., Murthy, K.N.: Corpus Based Statistical Approach for Stemming Telugu. In: Creation of Lexical Resources for Indian Language Computing and Processing (LRIL), C-DAC, Mumbai, India (2007)

[16] Akram, Q.-U.-A., Naseer, A., Hussain, S.: Assas-Band, an Affix-Exception-List Based Urdu Stemmer. In: Proceedings of ACL-IJCNLP 2009 (2009)

[17] http://en.wikipedia.org/wiki/Urdu

[18] http://www.bbc.co.uk/languages/other/guide/urdu/steps.shtml

[19] http://www.andaman.org/BOOK/reprints/weber/rep-weber.html

[20] Siddiqui, T.: Natural Language processing and Information Retrieval, U S Tiwary

[21] Frakes, W.B., Baeza-Yates, R.: Information retrieval: data structure and algorithms

[22] http://www.crulp.org/software/ling_resources.html