ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING 206

Álvaro Rocha
Ana Maria Correia
Tom Wilson
Karl A. Stroetmann (Eds.)

Advances in Information Systems and Technologies

Springer

Advances in Intelligent Systems and Computing 206

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

Álvaro Rocha, Ana Maria Correia, Tom Wilson,
and Karl A. Stroetmann (Eds.)

Advances in Information Systems and Technologies

Springer

Editors
Dr. Álvaro Rocha
LIACC
University of Porto
Porto
Portugal

Dr. Ana Maria Correia
Instituto Superior de Estatística e
Gestão de Informação
Campus de Campolide
Universidade Nova de Lisboa
Lisboa
Portugal

Dr. Tom Wilson
Sheffield
United Kingdom

Dr. Karl A. Stroetmann
Empirica GmbH
Bonn
Germany

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-642-36980-3 ISBN 978-3-642-36981-0 (eBook)
DOI 10.1007/978-3-642-36981-0
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013932785

Printed on acid-free paper

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

Preface

These proceedings contain all papers accepted for presentation and discussion at The 2013 World Conference on Information Systems and Technologies (WorldCIST'13). This Conference is organised by AISTI (Associação Ibérica de Sistemas e Tecnologias de Informação /Asociación Ibérica de Sistemas y Tecnologías de Información / Iberian Association for Information Systems and Technologies) and took place on 27th-30th of March in Olhão, Algarve, Portugal.

The World Conference on Information Systems and Technologies (WorldCIST) is a global forum for researchers and practitioners to discuss their most recent research, innovations, trends, results, experiences and concerns, also in view of recent and foreseeable technical trends as well as national and global policies. The meeting covers all relevant domains of Information Systems and Technologies, including Knowledge Management, Organisational Models, Decision Support Systems, Software Systems – Architectures - Applications – Tools, Computer Networks - Mobility - Pervasive Systems, Radar Technologies, Human-Computer Interaction. One of its main aims is to strengthen the drive towards symbiosis between academy, society and industry in all these fields and their application domains.

The Program Committee of WorldCIST'13 is composed of a multidisciplinary group of researchers and experts who are intimately concerned with Information Systems and Technologies research and application. They have had the responsibility for evaluating, in a 'blind review' process, the papers received for each of the main themes proposed for the Conference: A) Information and Knowledge Management (IKM); B) Organizational Models and Information Systems (OMIS); C) Intelligent and Decision Support Systems (IDSS); D) Software Systems, Architectures, Applications and Tools (SSAAT); E) Computer Networks, Mobility and Pervasive Systems (CNMPS); F) Radar Technologies (RAT); G) Human-Computer Interaction (HCI).

WorldCIST'13 received contributions from thirty-three countries. The papers accepted for its presentation and discussion at the Conference are published by Springer and will be indexed by ISI, EI, SCOPUS, DBLP and EBSCO, among others. Selected papers will be published in relevant journals and in a SCI (Studies in Computational Intelligence) series book.

We acknowledge all those contributing to the staging of WorldCIST13 (authors, committees and sponsors); their involvement is very much appreciated. It is our aim that WorldCIST becomes the global forum for discussing both latest developments in information systems and technologies RTD as well as in its varied application fields.

January 2013

Álvaro Rocha
Ana Maria Correia
Tom Wilson
Karl A. Stroetmann

Organization

Co-chairs

Ana Maria Ramalho Correia, Full Visiting Professor, University Nova de Lisboa, Rtd, PT; Associate Professor, Department of Information Sciences, University of Sheffield, UK

Tom Wilson, Emeritus Professor, University of Sheffield, UK; Visiting Professor, University of Leeds, UK; Visiting Professor, University of Boras, SE

Karl Stroetmann, Senior Research Fellow, Empirica Communication & Technology Research, DE

Álvaro Rocha, President of AISTI; Senior Research Fellow, LIACC, University of Porto, PT; Invited Professor, University of Santiago de Compostela, ES

Advisory Committee

Bezalel Gavish	Southern Methodist University, Dallas, US
Carlos Vaz de Carvalho	IEEE; Polytechnic of Porto, PT
Ezendu Ariwa	London Metropolitan University, UK
Gustavo Rossi	La Plata National University, Lifia, AR
Ivan Lukovic	University of Novi Sad, RS
Janusz Kacprzyk	Systems Research Institute, Polish Academy of Sciences, PL
Lei Li	Hosei University, JP
Luis Paulo Reis	University of Minho, PT
Marcelo Sampaio Alencar	Federal University of Campina Grande, BR

Program Committee

A. Aladwani	Kuwait University, KW
Abdallah AL Sabbagh	University of Technology, Sydney (UTS), AU
Adrian Florea	'Lucian Blaga' University of Sibiu, RO
Alexandre Balbinot	Federal University of Rio Grande do Sul, BR
Ali Elrashidi	University of Bridgeport, US
Alisson Brito	Federal University of Paraíba, BR
Amnart Pohthong	Prince of Songkla University, TH
Andrey Koptyug	Mid Sweden University, SE
Ankit Chaudhary	Bits Pilani, IN
Arthur Taylor	Rider University, US
Arun Gupta	M.S.College, IN
Avireni Srinivasulu	Vignan University, IN
Berk Canberk	Istanbul Technical University, TR
Bernard Cousin	University of Rennes 1 - IRISA, FR
Bharati Rajan	Loyola College, Chennai, IN
Bin Zhou	University of Maryland, Baltimore County, US
Bridget Kane	Norwegian University of Science and Technology, NO
Carla Pinto	Polytechnic of Porto, PT
Carlos Costa	ISCTE - IUL, PT
Cathryn Peoples	University of Ulster, UK
Cesar Alexandre de Sousa	University of São Paulo, BR
Chien-Ta Bruce Ho	National Chung Hsing University, TW
Ching-Hsien Hsu	Chung Hua University, TW
Christopher Martinez	University of New Haven, US
Christos Bouras	University of Patras and CTI&P-Diophantus, GR
Christos Douligeris	University of Piraeus, GR
Concha Soler	Radiotelevisión Valenciana, ES
Constantinos Hilas	Technological Educational Institute of Serres, GR
Cristina Alcaraz	University of Malaga, ES
Cristina Olaverri Monreal	Technische Universität München, DE
Damon Shing-Min	National Chung Cheng University, TW
David Mason	Victoria University of Wellington, NZ
Dinko Vukadinovic	University in Split, HR
Dirk Thorleuchter	Fraunhofer INT, DE
Dragan Milivojevic	Mining and Metallurgy Institute - Informatics Department, RS
Duggirala Venkata	TJPS PG College, IN
Eda Marchetti	CNR-ISTI, IT
Eli Upfal	Brown University, US
Elisa Francomano	University of Palermo, IT
Enrique Herrera-viedma	University of Granada, ES
Ernesto Exposito	LAAS-CNRS, FR
Erol Egrioglu	Ondokuz Mayıs University, TR

Fabrizio Montesi	IT University of Copenhagen, DK
Farhan Siddiqui	Walden University, CA
Fernando Almeida	University of Porto, PT
Fernando Bobillo	University of Zaragoza, ES
Fernando J Aguilar	Universiity of Almería, ES
Fernando Morgado Dias	University of Madeira, PT
Francesco Bianconi	Università degli Studi di Perugia, IT
Filipe Portela	Universidade do Minho, PT
Filippo Neri	University of Malta, MT
Fu-Chien Kao	Da-Yeh University, TW
Gabriele Oliva	University Campus Biomedico, IT
Garyfallos Arabatzis	Democritus University of Thrace, GR
Gitesh Raikundalia	Victoria University, AU
Gonçalo Paiva Dias	University of Aveiro, PT
Guenther Ruhe	University of Calgary, CA
Habiba Drias	University of Sciences and Technology Houari Boumediene, DZ
Hamdi Tolga Kahraman	Karadeniz Technical University, TR
Hamid Alinejad-Rokny	University of Newcastle, AU
Hari Om	Indian School of Mines Dhanbad, IN
Hartwig Hochmair	University of Florida, US
Hernani Costa	Universidade de Coimbra, PT
Hing Kai Chan	Norwich Business School, University of East Anglia, UK
Hironori Washizaki	Waseda University, JP
I-Hsien Ting	National University of Kaohsiung, TW
Isabel Pedrosa	Polytechnic Institute of Coimbra, PT
Ilaim Costa Junior	Fluminense Federal University, BR
J. Octavio Gutierrez-Garcia	ITAM, MX
Jaflah University of Bahrain	University of Bahrain, BH
Jan Nolin	Swedish School of Library and Information Science, SE
Jannica Heinstrom	Abo Akademi University, FI
Jari Makinen	Tampere University of Technology, FI
Jérôme Darmont	University of Lyon, FR
Jezreel Mejia	Centro de Investigacion en Matematicas Unidad Zacatecas, MX
Jianhua Chen	Louisiana State University, US
Jian Li	Tulane University, US
Jingfeng Xia	Indiana University, US
José Braga de Vasconcelos	University Fernando Pessoa, PT
Jose C. Valverde	University of Castilla-La Mancha, ES
Jose Luis Herrero Agustin	University of Extremadura, ES
José Luis Reis	ISMAI, PT
José Machado	University of Minho, PT

José Vicente Rodríguez Muñoz	University of Murcia, ES
Juan Carlos Torres	University of Granada, ES
Judith Broady-Preston	Aberystwyth University, UK
Jyh-Horng Chou	National Kaohsiung First University of Science and Technology, TW
Kaoru Hirota	Tokyo Institute of Technology, JP
Kewei Sha	Oklahoma City University, US
Kevin Ho	University of Guam, GU
Khaled Sayed	Modern University for Technology and Information, EG
Konstantinos Baltzis	Aristotle University of Thessaloniki, GR
Konstantinos Lakakis	Aristotle University of Thessalonikim GR
Kuan Yew Wong	Universiti Teknologi Malaysia, MY
Kun Chang Lee	Sungkyunkwan University, KR
Lori McCay-Peet	Dalhousie University, CA
Luis Velez Lapão	University Nova de Lisboa, PT
Lynne Conventry	Northumbria University, UK
Manfred Thaller	University at Cologne, DE
M. Rosario Fernández Falero	University of Extremadura, ES
Maciej Dabrowski	Digital Enterprise Research Institute, NUI Galway, IE
Manolis Vavalis	University of Thessaly, GR
Mansaf Alam	Jamia Millia Islamia - A Central University, IN
Manuel Cota	University of Vigo, ES
Manuel Mazzara	United Nations University, CN
Manuel Silva	Polytechnic of Porto, PT
Marco Scaioni	Tongji University, Shanghai, CN
Maria Bermudez-Edo	University of Granada, ES
Maria Lee	Shih Chien University, TW
Mário Pinto	Polytechnic Institute of Porto, PT
Martin Zelm	INTEROP VLab, BE
Masoud Abbaszadeh	Maplesoft, CA
Matthias Galster	University of Groningen, NL
Mazdak Zamani	Universiti Teknologi Malaysia, MY
Michal Strzelecki	Technical University of Lodz, PL
Mirjana Kljajic Borstnar	University of Maribor, SI
Mirna Ariadna Muñoz Mata	Centro de Investigación en Matemáticas- Unidad Zacatecas, MX
Miroslav Koncar	Oracle, HR
Mohammad Reza Daliri	Iran University of Science and Technology, IR
Mu-Song Chen	Da-Yeh University, TW
Natheer Gharaibeh	Al-Balqa Applied University, JO
Newlin Rajkumar Manokaran	Ann University of Technology, Rajkumar, IN
Niels Lohmann	Universität Rostock, DE

Noemi Emanuela Cazzaniga	Politecnico di Milano, IT
Nuno Fernandes	Polytechnic Institute of Castelo Branco, PT
Patrick Wang	Northeastern University, US
Panos Balatsoukas	University of Manchester, UK
Paolo Bientinesi	RWTH Aachen, DE
Pascal Lorenz	University of Haute Alsace, FR
Paweł Karczmarek	The John Paul II Catholic University of Lublin, PL
Pedro Peris	Computer Security Lab (COSEC), Carlos III University of Madrid, ES
Pedro Sousa	Universidade do Minho, PT
Peng-Sheng Chen	National Chung Cheng University, TW
Persephone Doupi	National Institute for Health and Welfare - THL, FI
Pierpaolo D'Urso	Sapienza - University of Rome, IT
Pierre L'Ecuyer	Université de Montréal, CA
Pierre Robillard	Polytechnique Montréal, CA
Preben Hansen	Swedish Institute of Computer Science, SE
Radu Prodan	University of Innsbruck, AT
Rainer Schmidt	HTW - Aalen University, DE
Ramiro Gonçalves	University of Trás-os-Montes e Alto Douro, PT
Rashid Ali	Taif University, SA
Rébecca Deneckère	University of Paris 1 Panthéon-Sorbonne, FR
Ren-Song Ko	National Chung Cheng University, TW
Rimvydas Skyrius	Vilnius University, LT
Robert Walker	University of Calgary, CA
Roberto Montemanni	Dalle Molle Institute for Artificial Intelligence (IDSIA), CH
Rubén González Crespo	Pontifical University of Salamanca, ES
Saeed Tavakoli	University of Sistan and Baluchestan, IR
Saleem Abuleil	Chicago State University, US
Salim Bitam	University of Biskra, DZ
Sandra Costanzo	University of Calabria, IT
Sangkyun Kim	Kangwon National University, KR
Sarang Thombre	Tampere University of Technology, FI
Santosh Kumar Nanda	Eastern Academy of Science and Technology, IN
Sergio Escalera	University of Barcelona, ES
Shaikh Abdul Hannan	Vivekanand College (LPP), Maharashtra, IN
Shamim Khan	Columbus State University, US
Sherali Zeadally	University of the District of Columbia, US
Sirje Virkus	Tallinn University, EE
Stasa Milojevic	Indiana University Bloomington, US
Stefan Jablonski	University of Bayreuth, DE
Stefan Wagner	University of Stuttgart, DE
Suksant Sae Lor	HP Labs, UK
Thomas Blaschke	University of Salzburg, AT
Tossapon Boongoen	Royal Thai Air Force Academy, TH

Tzung-Pei Hong	National University of Kaohsiung, TW
Valentina Emilia Balas	Aurel Vlaicu University of Arad, RO
Vitalyi Igorevich Talanin	Zaporozhye Institute of Economics and Information Technologies, UA
Waqas Bangyal	Iqra University Islamabad, PK
Wojciech Cellary	Poznan University of Economics, PL
Wolf Zimmermann	Martin-Luther-University Halle-Wittenberg, DE
WY Szeto	The University of Hong Kong, HK
Xiangmin Zhang	Wayne State University, US
Xiaoli Li	Institute for Infocomm Research, SG
Yafia Radouane	Ibn Zohr University, MA
Yaohang Li	Old Dominion University, US
Yair Wiseman	Bar-Ilan University, IL
Yi Gu	University of Tennessee at Martin, US
Yo-Ping Huang	National Taipei University of Technology, TW
Yong-Hyuk Kim	Kwangwoon University, KR
Yuhua Li	University of Ulster, UK
Yunqing Xia	Tsinghua University, CN
Zehua Chen	Taiyuan University of Technology, CN
Zhefu Shi	University of Missouri, Kansas City, US
Zhisheng Huang	Vrije University Amsterdam, NL
Zonghua Gu	Zhejiang University, CN

Contents

Information and Knowledge Management

Knowledge Acquisition Activity in Software Development 1
Olivier Gendreau, Pierre N. Robillard

An Electronic Learning System for Integrating Knowledge Management and Alumni Systems 11
Amnart Pohthong, Panumporn Trakooldit

Knowledge Management Systems and Intellectual Capital Measurement in Portuguese Organizations: A Case Study 23
Mário Pinto

Semantic Patent Information Retrieval and Management with OWL 33
Maria Bermudez-Edo, Manuel Noguera, José Luis Garrido, María V. Hurtado

Multilevel Clustering of Induction Rules for Web Meta-knowledge 43
Amine Chemchem, Habiba Drias, Youcef Djenouri

Knowledge-Based Risk Management: Survey on Brazilian Software Development Enterprises 55
Sandra Miranda Neves, Carlos Eduardo Sanches da Silva, Valério Antonio Pamplona Salomon, André Leonardo Almeida Santos

Leveraging Knowledge from Different Communities Using Ontologies 67
Herlina Jayadianti, Carlos Sousa Pinto, Lukito Edi Nugroho, Paulus Insap Santosa, Wahyu Widayat

An Approach for Deriving Semantically Related Category Hierarchies from Wikipedia Category Graphs 77
Khaled A. Hejazy, Samhaa R. El-Beltagy

Privacy Policies in Web Sites of Portuguese Municipalities: An Empirical Study 87
Gonçalo Paiva Dias, Hélder Gomes, André Zúquete

Knowledge Integration in Problem Solving Processes 97
Maria José Sousa

Collaborative Elicitation of Conceptual Representations: A Corpus-Based Approach 111
Cristóvão Sousa, Carla Pereira, António Soares

Effect of Demography on Mobile Commerce Frequency of Actual Use in Saudi Arabia 125
AbdulMohsin Alkhunaizan, Steve Love

Specialized Knowledge Systems – A Model for Intelligent Learning Management within Organizations 133
Isabel Mendes, Henrique Santos, Celina Pinto Leão

How Small and Medium Enterprises Are Using Social Networks? Evidence from the Algarve Region 143
Ana Belo, Guilherme Castela, Silvia Fernandes

Temporal Visualization of a Multidimensional Network of News Clips 157
Filipe Gomes, José Devezas, Álvaro Figueira

User Documentation: The Cinderella of Information Systems 167
Brigit van Loggem

Task Topic Knowledge vs. Background Domain Knowledge: Impact of Two Types of Knowledge on User Search Performance 179
Xiangmin Zhang, Jingjing Liu, Michael Cole

Community Detection by Local Influence 193
Nuno Cravino, Álvaro Figueira

Predict Sepsis Level in Intensive Medicine – Data Mining Approach 201
João M.C. Gonçalves, Filipe Portela, Manuel Filipe Santos, Álvaro Silva, José Machado, António Abelha

Constructing Conceptual Model for Security Culture in Health Information Systems Security Effectiveness 213
Ahmad Bakhtiyari Shahri, Zuraini Ismail, Nor Zairah Ab. Rahim

Using Domain-Specific Term Frequencies to Identify and Classify Health Queries 221
Carla Teixeira Lopes, Daniela Dias, Cristina Ribeiro

Dealing with Constraint-Based Processes: Declare and Supervisory Control Theory 227
Sauro Schaidt, Agnelo Denis Vieira, Eduardo de Freitas Rocha Loures, Eduardo Alves Portela Santos

Adopting Standards in Nursing Health Record – A Case Study in a Portuguese Hospital 237
Bruno Rocha, Álvaro Rocha

The Relationship between Portal Quality and Citizens' Acceptance: The Case of the Kuwaiti e-Government 249
Adel M. Aladwani

Knowledge Management Framework for Six Sigma Performance Level Assessment 255
Jevgeni Sahno, Eduard Sevtsenko, Tatjana Karaulova

Android, GIS and Web Base Project, Emergency Management System (EMS) Which Overcomes Quick Emergency Response Challenges 269
Atif Saeed, Muhamamd Shahid Bhatti, Muhammad Ajmal, Adil Waseem, Arsalan Akbar, Adnan Mahmood

Pervasive Intelligent Decision Support System – Technology Acceptance in Intensive Care Units 279
Filipe Portela, Jorge Aguiar, Manuel Filipe Santos, Álvaro Silva, Fernado Rua

GLORIA: The First Free Access e-Infrastructure of Robotic Telescopes for Citizen Science 293
Carlos Jesús Pérez-del-Pulgar, Raquel Cedazo, Juan Cabello, Esteban González, Víctor F. Muñoz, Fernando Serena, María C. López, Fernando Ibáñez, Francisco M. Sánchez, Alberto Castro, Ronan Cunniffe

Evaluating Web Site Structure Based on Navigation Profiles and Site Topology 305
Alberto Simões, Anália Lourenço, José João Almeida

The Problems of the Insolvency Register in the Czech Republic from the Perspective of Information Technology 313
Luboš Smrčka

Improving Public Transport Management: A Simulation Based on the Context of Software Multi-agents 323
Marcia Pasin, Thiago Lopes Trugillo da Silveira

Social Networks Mining Based on Information Retrieval Technologies and Bees Swarm Optimization: Application to DBLP 331
Drias Yassine, Drias Habiba

Organizational Models and Information Systems

Information and Information Systems Project for a Strategic Digital City: A Brazilian Case 345
Denis Alcides Rezende, Frederico de Carvalho Figueiredo, Leana Carolina Ferreira Setim, Luciane Maria Gonçalves Franco, Gilberto dos Santos Madeira

Linking Benefits to Balanced Scorecard Strategy Map 357
Jorge Gomes, Mário Romão, Mário Caldeira

Modeling e-Government for Emergent Countries: Case of S.Tome and Príncipe 371
Artur Celestino Vera Cruz

An Analysis of the Disclosure of Social Responsibility in Australian Universities 383
Raquel Garde Sánchez, Manuel Pedro Rodríguez-Bolívar, Laura Alcaide-Muñoz, Antonio M. López-Hernández

Information Architectures Definition – A Case Study in a Portuguese Local Public Administration Organization 399
Filipe Sá, Álvaro Rocha

TSPi to Manage Software Projects in Outsourcing Environments 411
Jezreel Mejia, Andrés Garcia, Mirna A. Muñoz

Aspects That Contribute to the Success of Personalized Web Applications 421
José Luís Reis, João Álvaro Carvalho

Implementing eHealth Services for Enhanced Pharmaceutical Care Provision: Opportunities and Challenges 433
Luís Velez Lapão, João Gregório, Afonso Cavaco, Miguel Mira da Silva, Christian Lovis

Establishing Multi-model Environments to Improve Organizational Software Processes 445
Muñoz Mirna, Mejia Jezreel

Standardization of Processes Applying CMMI Best Practices 455
Vítor Serrano, Anabela Tereso, Pedro Ribeiro, Miguel Brito

Developing and Validating a Scale for Perceived Usefulness for the Mobile Wallet 469
Debby Ho, Milena Head, Khaled Hassanein

Term Proximity and Data Mining Techniques for Information Retrieval Systems 477
Ilyes Khennak, Habiba Drias

A Set of Requirements for Business Process Management Suite (BPMS) 487
Nemésio Freitas Duarte Filho, Norben P.O. Costa

High Level Architecture for Trading Agents in Betting Exchange Markets 497
Rui Gonçalves, Ana Paula Rocha, Fernando Lobo Pereira

An MDA Approach to Develop Web Components 511
José Luis Herrero Agustin, Pablo Carmona, Fabiola Lucio

Towards a Conceptual Framework for Early Warning Information Systems (EWIS) for Crisis Preparedness 523
Mohamed Saad, Sherif Mazen, Ehab Ezzat, Hegazy Zaher

Intelligent and Decision Support Systems

Aggregation Operators and Interval-Valued Fuzzy Numbers in Decision Making 535
József Mezei, Robin Wikström

Deriving Weights from Group Fuzzy Pairwise Comparison Judgement Matrices 545
Tarifa S. Almulhim, Ludmil Mikhailov, Dong-Ling Xu

An Economic Production Quantity Problem with Fuzzy Backorder and Fuzzy Demand 557
József Mezei, Kaj-Mikael Björk

Analyzing Website Content for Improved R&T Collaboration Planning 567
Dirk Thorleuchter, Dirk Van den Poel

Building Accountability for Decision-Making into Cognitive Systems 575
Victoria L. Lemieux, Thomas Dang

The Relationship between Management Decision Support and Business Intelligence: Developing Awareness 587
Rimvydas Skyrius, Gėlytė Kazakevičienė, Vytautas Bujauskas

Multi-Agent System for Teaching Service Distribution with Coalition Formation 599
José Joaquim Moreira, Luís Paulo Reis

A Comprehensive Study of Crime Detection with PCA and Different Neural Network Approach 611
Ahmad Kadri Junoh, Muhammad Naufal Mansor

A Conceptual Model of Layered Adjustable Autonomy 619
Salama A. Mostafa, Mohd Sharifuddin Ahmad, Muthukkaruppan Annamalai, Azhana Ahmad, Saraswathy Shamini Gunasekaran

A Dynamically Adjustable Autonomic Agent Framework 631
Salama A. Mostafa, Mohd Sharifuddin Ahmad, Muthukkaruppan Annamalai, Azhana Ahmad, Saraswathy Shamini Gunasekaran

High-Level Language to Build Poker Agents 643
Luís Paulo Reis, Pedro Mendes, Luís Filipe Teófilo, Henrique Lopes Cardoso

New Crime Detection with LBP and GANN 655
Ahmad Kadri Junoh, Muhammad Naufal Mansor

Cooperative Scheduling System with Emergent Swarm Based Behavior 661
Ana Madureira, Ivo Pereira, Diamantino Falcão

Software Systems, Architectures, Applications and Tools

Finding the Suitable Number of Resources to Maximize System Throughput 673
M. Carmen Ruiz, Diego Pérez, Juan José Pardo, Diego Cazorla

Step towards Paper Free Hospital through Electronic Health Record 685
Maria Salazar, Júlio Duarte, Rui Pereira, Filipe Portela, Manuel Filipe Santos, António Abelha, José Machado

Remote Scientific Computing over the Internet: An Example in Geometry 695
Miguel Ferreira, Miguel Casquilho

Achieving Multiple Dispatch in Hybrid Statically and Dynamically Typed Languages 703
Francisco Ortin, Miguel Garcia, Jose M. Redondo, Jose Quiroga

Open Source Technologies Involved in Constructing a Web-Based Football Information System 715
Pedro Rodrigues, António Belguinha, Carlos Gomes, Pedro Cardoso, Tiago Vilas, Renato Mestre, J.M.F. Rodrigues

A Collaborative Tourist System Using Serious Games 725
Rui Pedro Araújo Fernandes, João Emilio Almeida, Rosaldo J.F. Rosseti

SOA – Based Authentication System for Dynamic Handwritten Signature 735
Andreea Salinca, Sorin Mircea Rusu, Ana-Maria Pricochi

Extending the Groovy Language Using AST Transformations to Monitor Variables and Methods 745
Carlos Cortinhas, Fernando Barros

Network Based Analysis of Intertextual Relations 753
Ioan Cristian Ghiban, Ştefan Trăuşan-Matu

Building an Integrated System for the Management of Scientific Nature Events through Open Source Software Integration 763
Carlos Serrão, Miguel Gamito

A Tool for Fractal Component Based Applications Performance Modelling Using Stochastic Well Formed Nets 773
Nabila Salmi, Malika Ioualalen, Smail Lallali, Hamza Zerguine

Open Source Software Documentation Mining for Quality Assessment 785
Nuno Ramos Carvalho, Alberto Simões, José João Almeida

A Networked Application to Support the Learning of Electronic Marketing Based on e-Learning and a Portfolio of Mediating Tools 795
Luis Vaz, Nuno David

A Platform-as-a-Service API Aggregator 807
David Cunha, Pedro Neves, Pedro Sousa

MC64-Cluster: A Many-Core CPU Cluster for Bioinformatics Applications 819
Francisco J. Esteban, David Díaz, Pilar Hernández, Juan A. Caballero, Gabriel Dorado, Sergio Gálvez

Self-Portrait Images for Mobile Application 827
Teh Phoey Lee, Nael Kabbany, Chan Kei Jun

Lessons Learned from Creating a General Purpose Tool for Experience Sampling Methods 839
André Coelho, Rui José

ManPro: Framework for the Generation and Assessment of Documentation for Nuclear Facilities 849
Cristina Olaverri-Monreal, Carsten Dlugosch, Klaus Bengler

Text Mining Indicators of Affect and Interaction: A Case Study of Students' Postings in a Blended-Learning Course of English for Specific Purposes 861
Helvia P.P. Bastos, Magda Bercht, Leandro K. Wives, Júlia Kambara-Silva, Yasmmin Martins

GEMINI: A Generic Multi-Modal Natural Interface Framework for Videogames 873
Luís Filipe Teófilo, Pedro Alves Nogueira, Pedro Brandão Silva

A Semi-automatic Negotiation Strategy for Multi-attribute and Multiple Participants 885
Rharon Maia, Carlos Dias, Marcus R. Laurentino, Alisson Vasconcelos Brito

Smart Land Record Application Using Web GIS and GPS 893
Muhammad Shahid Bhatti, Muhammad Ajmal, Atif Saeed, Maqsood Ahmed, Rizwan Khalid, Nouman Arshad

Computer Networks, Mobility and Pervasive Systems

Polynomial Approximation of the Battery Discharge Function in IEEE 802.15.4 Nodes: Case Study of MicaZ 901
Odilson T. Valle, A. Milack, C. Montez, Paulo Portugal, Francisco Vasques

Enhancing PTN to Improve the QoS Provided by the IP Mobility Management 911
David Cortés-Polo, José-Luis González-Sánchez, Javier Carmona-Murillo, Fco. Javier Rodríguez-Pérez, Javier Corral-García

User's Requirements in Internet Access Sharing 923
Conceição Tavares, Henrique Santos

A Multidimensional Model for Monitoring Cloud Services 931
Nuno Palhares, Solange Rito Lima, Paulo Carvalho

Effects of NGNs on Market Definition 939
João Paulo Ribeiro Pereira

Ambient Assisted Living and the Integration and Personalization of Care Services 951
Alexandra Queirós, Nelson Pacheco da Rocha

A Mobile and Web Indoor Navigation System: A Case Study in a University Environment 959
Sara Paiva

Towards a Security Solution for Mobile Agents 969
Djamel Eddine Menacer, Habiba Drias, Christophe Sibertin-Blanc

Radar Technologies

X-Band Radar Sensor for the Landslide Risk Mitigation 981
Sandra Costanzo, Giuseppe Di Massa, Marco Salzano

Compact Slotted Antenna for Wideband Radar Applications 989
Sandra Costanzo, Antonio Costanzo

High Resolution Software Defined Radar System for Target Detection 997
Sandra Costanzo, Francesco Spadafora, Antonio Borgia, Oswaldo Hugo Moreno, Antonio Costanzo, Giuseppe Di Massa

Design of a Reconfigurable Reflectarray Unit Cell for Wide Angle Beam-Steering Radar Applications 1007
Francesca Venneri, Sandra Costanzo, Giuseppe Di Massa

Human-Computer Interaction

Age Differences in Computer Input Device Use: A Comparison of Touchscreen, Trackball, and Mouse 1015
Ho-chuen Ng, Da Tao, Calvin K.L. Or

Color-Concept Associations among Chinese Steel Workers and Managerial Staff 1025
Heller H.L. Wang, Calvin K.L. Or

Cycle of Information Retrieval through Color in e-Commerce: Store Choice 1033
M. Rosario Fernández Falero, Libertad Sánchez Gil, Luis V. Gordillo Tapia

Visualization and Manipulation of Information in 3D Immersive Environments 1041
Filipe Costa, João Paulo Pereira, António Castro

Construction Processes Using Mobile Augmented Reality: A Study Case in Building Engineering Degree. 1053
Albert Sánchez Riera, Ernesrt Redondo, David Fonseca, Isidro Navarro

A User-Centered Interface for Scheduling Problem Definition 1063
Jesus Piairo, Ana Madureira, João Paulo Pereira, Ivo Pereira

Perceived Site Security as a Second Order Construct and Its Relationship to e-Commerce Site Usage 1075
Edward Hartono, Ki-Yoon Kim, Kwan-Sik Na, James T. Simpson, David Berkowitz

Accessibility Study in Sites of Public Higher Education Institution in Brazil 1087
Mariana Angelo Lopes Sanches, Lílian Simão Oliveira

Exploring the Design Space of Mobile Payment Systems 1095
Rui José, Nuno Otero, Helena Rodrigues, Filipe Meneses, Odete Coelho

SketchyDynamics: A Sketch-Based Library for the Development of Physics Simulation Applications 1105
Abílio Costa, João Paulo Pereira

Designing User Learning Experience in Virtual Worlds: The Young Europeans for Democracy Serious Application 1117
Gonçalo Cruz, Ana Maia, Leonel Morgado, Benjamim Fonseca, Hugo Paredes, Fernando Bessa, Clara Rodrigues, Paulo Martins

The Behaviour Assessment Model for the Analysis and Evaluation of Pervasive Services 1129
Bernhard Klein, Ivan Pretel, Ulf-Dietrich Reips, Ana B. Lago, Diego Lopez-de-Ipiña

A Study of Biometric Authentication Adoption in Health Services 1141
Paulo Rodrigues, Henrique Santos

Cell Life: A Biology Game to Support Biology Classrooms 1149
Teresa Futscher de Deus, Pedro Faria Lopes

Author Index 1157

Knowledge Acquisition Activity in Software Development

Olivier Gendreau and Pierre N. Robillard

Department of Computer and Software Engineering
Polytechnique Montréal,
C.P. 6079, Succ Centre-Ville
Montréal, Qc, Canada, H3C 3A7
{olivier.gendreau,pierre.robillard}@polymtl.ca

Abstract. Data from four field studies are analyzed to find the patterns of knowledge acquisition activity in software development projects with respect to other cognitive activities such as documentation, coding and V&V. The data are obtained from self-recorded activity time slips approach. Data are codified based on an information source model, which is related to Nonaka and Takeuchi's knowledge creation model. It shows that knowledge acquisition activities account for almost 15% of the total effort. We also find out that this effort, in most cases, cannot be restricted to the first phase of the project during requirement and architectural design, which is expected from waterfall or disciplined processes. About half of the learning is done during the code implementation even within a disciplined process. This finding is in line with one value of the Agile philosophy that promotes team interactions and users involvement for the whole project duration.

Keywords: Knowledge acquisition, software development, cognitive activities, knowledge flow, cognitive factors, empirical studies, field studies.

1 Introduction

Software engineering is a knowledge-intensive activity [1,2,3,4]. Software development requires programmers to gather and absorb large amounts of knowledge distributed over several domains, such as application and programming, and to encode that knowledge in the software [3].

In order to better understand the complexity of software development, Ko et al. [5] suggest analyzing software activities from a knowledge perspective. However, the nature of knowledge poses a methodological challenge. Since knowledge is the product of various cognitive activities and mostly resides in a software developer's mind, it might be better described by the developers themselves. Therefore, this study is based on a data acquisition approach in which software developers record their activities from a knowledge viewpoint. This approach is used to gain an understanding of how knowledge acquisition needs evolve throughout the development of a software project.

Á. Rocha et al. (Eds.): *Advances in Information Systems and Technologies*, AISC 206, pp. 1–10.
DOI: 10.1007/978-3-642-36981-0_1

In the cognitive sciences, four knowledge models are referred to widely: Kolb's model of experiential learning [6], Argyris and Schön's double-loop learning theory [7], Wenger's theory of communities of practice [8], and Nonaka and Takeuchi's theory of knowledge creation [9], which is the model used for this study. We describe six cognitive factors involved in software development projects. Data were collected from four selected industrial capstone projects, which were based on requirements supplied by a single avionics industrial partner.

Section 2 presents the knowledge model used to describe the cognitive activities. Section 3 describes the four field studies that provide the data. Section 4 presents the self recording Activity Time Slip (ATS) approach. Section 5 presents the knowledge acquisition patterns observed in the field studies.

2 Information Source Model

We used an information source model (see Fig. 1) which is related to Nonaka and Takeuchi's knowledge creation model [10] to analyze software development from a knowledge acquisition perspective.

The three round-cornered boxes in Fig. 1 represent information sources from which developers can built their knowledge. External information may come from various sources, such as the Web, a paper, or a book or technical documentation related to the product to be developed. Development artifact information comes from any of the project's documentation. Source code strictly includes executable statements and comments. Tacit knowledge is individual knowledge built from interaction with information sources. The arrows in Fig. 1 represent the various cognitive factors that constitute the knowledge built up from the various information sources. The acquisition cognitive factor is involved when a developer needs to increase his tacit knowledge by taking in external information. The crystallization cognitive factor is the translation of a developer's mental representation of a concept (tacit knowledge) into an artifact (explicit knowledge), such as a use case diagram or an architectural plan. The realization cognitive factor involves the translation of tacit knowledge into explicit information or documentation artifacts, but requires, in addition, technical know-how, which is related to source code production. The validation cognitive factor involves bidirectional information flow between tacit knowledge and development artifacts (explicit knowledge), in order to validate the consistency between the knowledge and the information source. The verification cognitive factor is like validation, except that source code is the information source, and so it involves technical know-how. The work planning cognitive factor mostly involves developers' synchronization of the project's planning and progress information.

This information source model is limited to software development activities. The management activities related to the software project are not taken into consideration in it, because they are not specific to software development, and also because they frequently involve several projects.

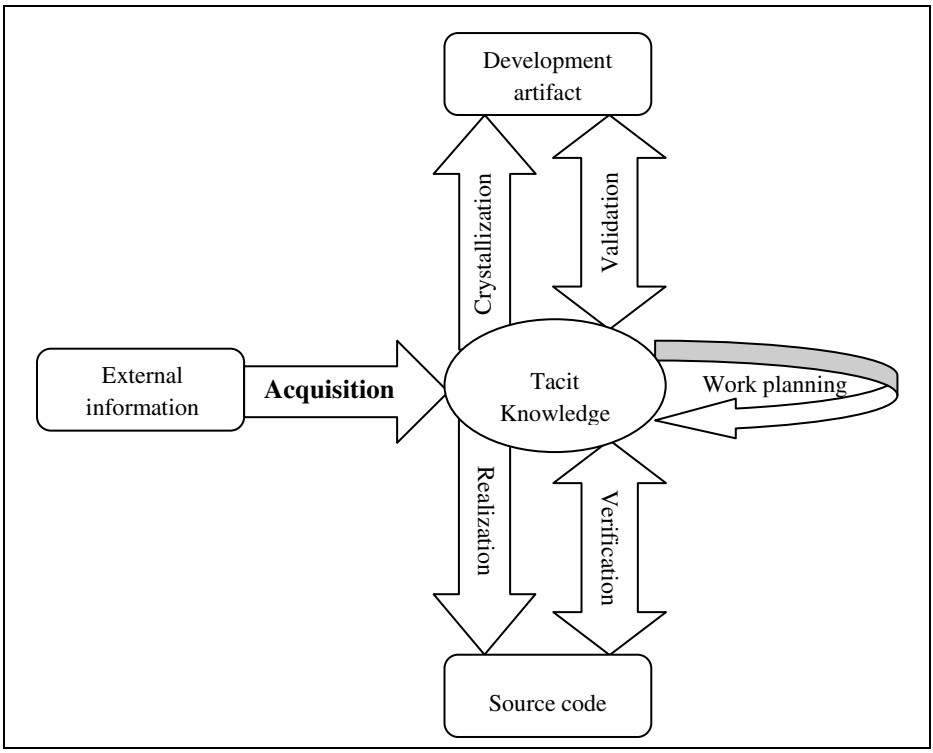

Fig. 1. Information source model

3 Field Studies

Data were collected from four selected industrial capstone projects conducted at our engineering school. The projects labeled P06 to P09 were successful in terms of teamwork and functional deliverables. The four projects were based on requirements supplied by a single avionics industrial partner.

The teams of five students were formed based on four criteria: number of cumulative total credits, past internship experience in industry, current grade point average, as well as software design and process course grades. They were chosen with the objective of balancing the know-how and experience on the teams.

The capstone projects were conducted over one semester (14 weeks) on a fixed schedule of three half-day team working sessions per week, and a flexible schedule of up to three extra half-days per week. The teams had access to an equipped dedicated room on campus for the duration of the project. All the projects, which are briefly described, are related to avionic applications and required adding some functionality to existing software systems.

The P06 project required to add a graphical interface to a design and configuration system used to build avionic model and enable editing of the various system model components and messages.

The P07 project required to build a graphical interface to create, delete or modify the hierarchy of the groups in Doxygen, which is a documentation system (under GNU General Public License) that generates documentation from source code [11].

The P08 project required to build a tool that will automatically extract data from a video of a plane cockpit dashboard. This tool is based on OCR (Optical character recognition) software. Once the video is loaded, the user can define the areas from which the data must be extracted. The extracted data are saved on a CVS file for further processing.

The P09 project required to translate a proprietary file format into a Microsoft Windows Presentation Foundation Format (WPF) [12]. The new files must have the same functionality as the old ones statically and dynamically. This project involved writing from scratch a home-made parser.

The external validity of empirical studies with students is a commonly raised concern. According to Carver et al. [13], more and more students are employed for either a summer internship or a full internship in an industrial environment. Höst et al. [14] conclude that only minor differences exist between students and professionals, and their research does not challenge the assumption that final year software engineering students are qualified to participate in empirical software engineering research. Similar results were obtained in a study on detection methodologies for software requirement inspection conducted by Porter et al. [15] among students, and then replicated among professionals [16]. Consequently, the external validity of our study is increased because it was conducted among senior students who have some internship experience in industry.

4 Data Acquisition Approach

The ATS (Activity Time Slip) approach, which is used in these field studies, is an augmented work diary approach focusing on the activity instead of the task being performed [17, 18]. The meaning of *work* (as in "work diary") is different from that of *activity* (as in the ATS). On the one hand, work is related to a task, and is often part of a schedule and is related to project resources. On the other hand, an activity is a personal endeavor undertaken while a developer is executing a task. Examples of activities reported on an ATS are: browsing the Web, reading about an API, talking to teammate about a concern, etc.

The ATS approach requires that the developer log, in an ATS token, the details of every activity performed. Table 1 gives an example of ATS token fields. Each ATS entry takes into account activities that may last more than an hour or only few minutes and records the teammates who are involved.

Each developer uses a preformatted spreadsheet to detail activities on an ongoing basis, at the rate of roughly one entry per hour. The ATS approach was applied throughout the entire duration of the projects.

Table 1. Activity Time Slip (ATS) token content

Field	Description	Example
ID	Unique token identifier	75
Date	Activity date	2012-05-05
Start time	Activity start time	9:20
End time	Activity end time	10:15
Effort	Activity duration (computed from the start/end time fields)	55
P_1 .. P_n	P_1 to P_n participants involved in executing the activity	GL, PN
Input artifact	Main input artifact of the activity	SRS
Output artifact	Main output artifact of the activity	CPA
Activity description	Detailed description of the activity	Use-Case Realization A,B,C,S
Process	Process discipline related to the activity	REQ
Task	Prescribed task	Modeling interface

Table 2 presents the total software development effort in hours, the number of tokens, and the tokens-per-hour ratio for each of the four projects.

Table 2. Software development effort and tokens

Project	Effort (Hrs)	Number of tokens	Tokens/hour
P06	997	1426	1.4
P07	750	1408	1.9
P08	810	887	1.1
P09	628	621	1.0

In order to extract knowledge behavior from self-reported developer activities, a coding scheme, based on the information source model, has been designed. An ATS token is codified according to the cognitive factor concerned. However, some tokens involve more than one cognitive factor. In this case, the coder needs to determine the dominant cognitive factor, mainly based on the description of the token and its context (input artifact, process, etc.). For instance, fixing a code defect involves both the verification and realization cognitive factors. First, it requires locating the defect in the code, which is related to the verification cognitive factor. Then, the actual fixing of the code involves the realization cognitive factor. In this situation, the dominant cognitive factor remains verification.

All the tokens of the four projects were codified by two independent coders, who had to decide which cognitive factor was dominant. However, tokens related to academic and technical activities were not accounted for in the codification, since they were not specific to project development. Academic activities are related to the lectures given by the instructors or presentations by the students, such as teamwork

training and project presentation. Technical activities are related to tasks which can be performed by technicians, such as configuring the network or setting up and maintaining the development environment.

5 Knowledge Acquisition in Software Projects

Every project has three phases, ending with a milestone. Many development artifacts are produced during the first phase, such as SRS, use case documentation, architecture and design documentation. The first milestone requires the development team to present their system architecture to the industrial partner. This occurs between 20% and 35% of project completion, depending on the project. Most of the system is coded during the second phase. The second milestone requires the team to package its application for acceptance testing. This occurs between 85% and 90% of project completion. Integration and acceptance testing are performed during this third phase. The third milestone occurs at the end of the semester, when the product is delivered to the client.

We recall that the four projects had the same industrial client, used the same disciplined process, and developed similar but different avionics applications. All participants had similar background and experience, but the teams were different for each project.

Fig. 2 to 5 show the total effort expended on each cognitive factor in relation to the four project completion. Each of the 6 curves of the graphs represents the relative total effort expended (Y-axis) for a given cognitive factor with respect to the percentage of project completion (X-axis). For example, in P06 (Fig. 2), at 30% of project completion (X-axis), 14% of the total effort (Y-axis) had been expended on crystallization.

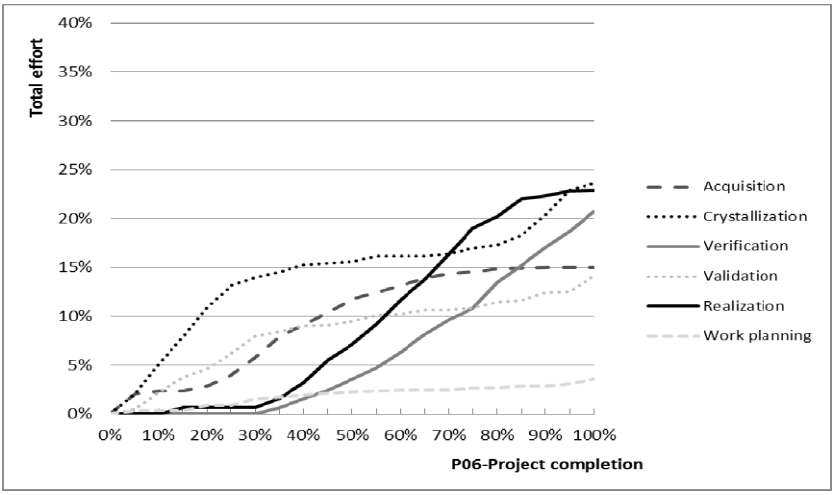

Fig. 2. Project P06 total effort distribution (Add graphical interface)

Validation represented 9% of the total effort, acquisition 6%, and realization and work planning both accounted for 1%. No verification effort had been expended to that point. Analysis of the slopes of the 6 curves in Fig. 2 provides a better understanding of the relationships between cognitive factors throughout the project. At project completion (X-axis 100%) we see that acquisition accounts for 15% of the total effort. In project P06 (Fig. 2), the learning activity (acquisition) is following the realization activity and the new knowledge is documented (crystallization) at the end of the project. Half of the learning effort (acquisition) was expended during the coding phase. Team members learned as they develop the product. There is as much total effort in realization activity as in crystallization activity.

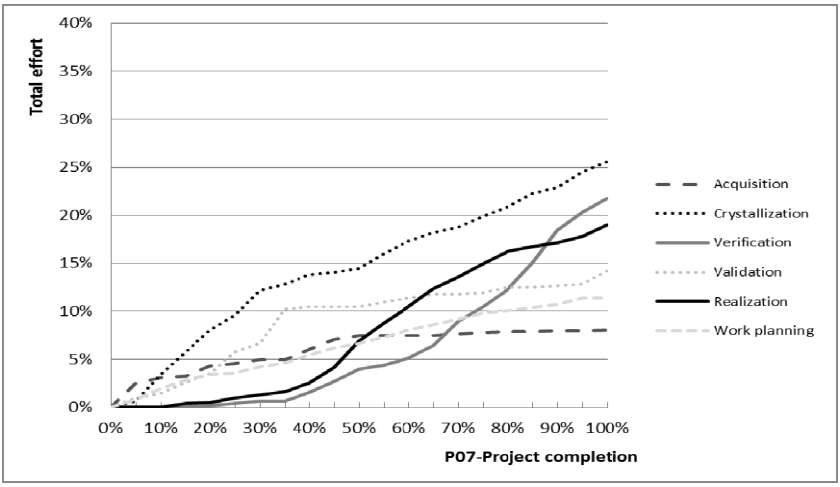

Fig. 3. Project P07 total effort distribution (Doxygen grouping)

In project P07 (Fig. 3) most of the acquisition occurred at the beginning of the project, during the first phase, and there is almost no acquisition during the coding activities. The learning behavior of this team was different in many respects. It is the project that required the least total learning activity (8%) and there is no learning during the coding phase. The cognitive factor requiring the most total effort was crystallization, followed by verification and realization. This cognitive behavior is related to the low level of technical difficulty of the project.

In project P08 (Fig. 4), the team expended as much total effort in documenting (crystallization) as in testing (verification). As in P06, half of the learning effort (acquisition) was expended during the coding phase. This new learning was documented as it is shown by the crystallization curve that followed the acquisition curve.

In the project P09 (Fig. 5), which is clearly coding oriented (realization), a minimum of acquisition activity occurred at the beginning of the project, and this produced documentation that would be revisited only during the third phase. The realization phase began early in the project (at 20% of project completion). However,

the acquisition activity was maintained throughout the coding activities. At 60% of project completion, the team needed to learn more about the project. However, part of this new knowledge was updated (crystallization) only at the very end of the project.

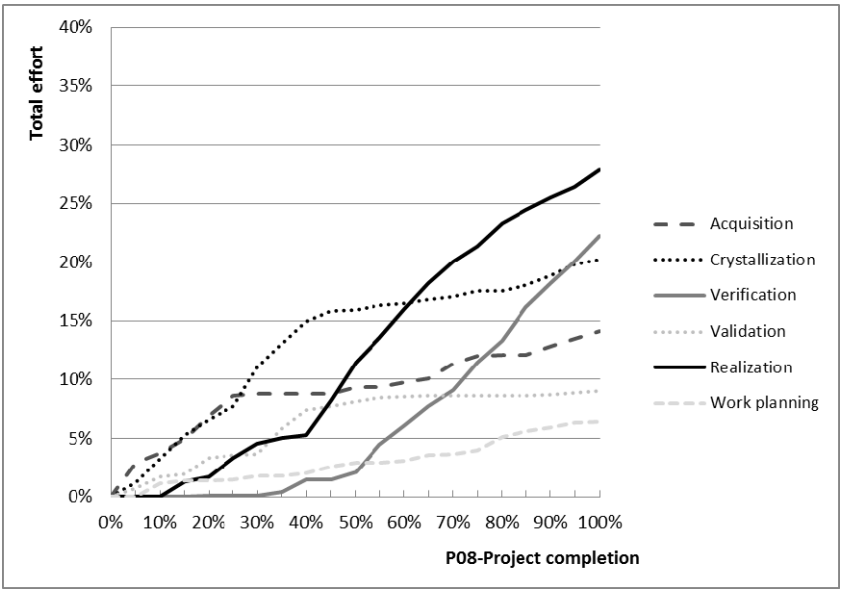

Fig. 4. Project P08 total effort distribution (OCR from video)

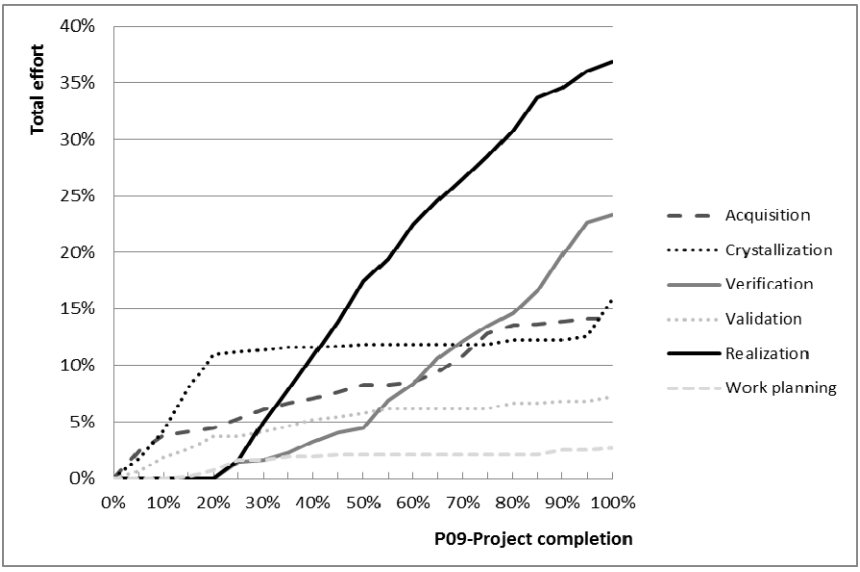

Fig. 5. Project P09 total effort distribution (Home-made parser)

Preliminary results from these field studies seem to indicate that the acquisition activity has an impact on the coding effort. For example, we observed that all the projects but one (P07) required around 15% of the acquisition effort, and that the later in the project the acquisition activity occurred, the greater the coding effort (see P08 and P09).

It seems that the need for more information or more learning emerged from activities performed during the coding phase. This could have an impact on the efficiency of Agile philosophy development, for example. Both projects P06 and P08 involved reusing software components. In P06, the team had to enhance a system by adding a graphical interface, while in P08, OCR components were used to build the required system. From a knowledge acquisition standpoint, it is noteworthy that in both cases, half of the acquisition was expended during the coding phase, even though a disciplined software process was used. This tends to confirm that each cognitive factor (acquisition, crystallization, validation, realization, and verification) is important throughout to whole project duration, as promoted in the Agile philosophy.

6 Concluding Remarks

There is a growing need to consider the knowledge perspective in software development, since developers' activities are mostly cognitive. Such knowledge cannot be measured directly, since it is mostly tacit, that is, it resides in the developer's mind. However we can measure the activities that lead to learning.

The ATS approach presented in this paper is a compromise between the acquisition of very accurate (think-aloud) data on participant cognitive activities in a short time and that of self-reported data on these activities over the duration of a project. The level of accuracy obtained with the ATS approach is sufficient to explore various knowledge acquisition perspectives in software development.

We find that the acquisition of information, which is learning, leads to improved knowledge and requires almost 15% of the total team effort. We also find that this effort, in most cases, cannot be restricted to the first phase of the project during requirement and architectural design, which is expected from waterfall or disciplined processes. About half of the learning is done during the code implementation even within a disciplined process. This finding is in line with one value of the Agile philosophy that promotes team interactions and users involvement for the whole project duration.

Future works will involve repeating these field studies with similar projects but with Agile teams. It could be interesting to see how the information acquisition patterns are modified with these new software development approaches.

Project managers should be aware that learning is an important component of software development and they should provide the social and physical environments to facilitate these learning activities.

Acknowledgments. This research was supported in part by NSERC grant A-0141.

References

1. Henninger, S.: Tools Supporting the Creation and Evolution of Software Development Knowledge. In: International Conference on Automated Software Engineering (ASE 1997), pp. 46–53 (1997)
2. Robillard, P.N.: The Role of Knowledge in Software Development. Commun. ACM 42(1), 87–92 (1999)
3. Xu, S., Rajlich, V., Marcus, A.: An Empirical Study of Programmer Learning During Incremental Software Development. In: 4th IEEE International Conference on Cognitive Informatics, pp. 340–349 (2005)
4. Bjornson, F.O., Dingsoyr, T.: Knowledge Management in Software Engineering: A Systematic Review of Studied Concepts, Findings and Research Methods Used. Inf. Softw. Technol. 50(11), 1055–1068 (2008)
5. Ko, A.J., DeLine, R., Venolia, G.: Information Needs in Collocated Software Development Teams. In: 29th International Conference on Software Engineering, pp. 344–353 (2007)
6. Kolb, D.: Experiential Learning: Experience as the Source of Learning and Development. Prentice Hall, Englewood Cliffs (1984)
7. Argyris, C., Schon, D.A.: Organizational Learning: A Theory of Action Perspective. Addison-Wesley, Reading (1978)
8. Wenger, E.: Communities of Practice: Learning, Meaning and Identity. Cambridge University Press, Cambridge (1998)
9. Nonaka, I., Takeuchi, H.: The Knowledge-Creating Company – How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, Oxford (1995)
10. Gendreau, O., Robillard, P.N.: Exploring Knowledge Flow in Software Project Development. In: International Conference on Information, Process, and Knowledge Management (eKNOW 2009), pp. 99–104 (2009)
11. Doxygen Manual, http://www.doxygen.org
12. WPF Tutorial.net, http://www.wpftutorial.net
13. Carver, J., Jaccheri, L., Morasca, S., Shull, F.: Issues in Using Students in Empirical Studies in Software Engineering Education. In: Ninth International Software Metrics Symposium (METRICS 2003), pp. 239–249 (2003)
14. Höst, M., Regnell, B., Wohlin, C.: Using Students as Subjects – A Comparative Study of Students and Professionals in Lead-Time Impact Assessment. Empir. Softw. Eng. 5(3), 201–214 (2000)
15. Porter, A., Votta, L., Basili, V.R.: Comparing Detection Methods for Software Requirements Inspection: A replicated experiment. IEEE Trans. Softw. Eng. 21(6), 563–575 (1995)
16. Porter, A., Votta, L.: Comparing detection methods for software requirements inspection: A replication using professional subjects. Empir. Softw. Eng. 3(4), 355–380 (1998)
17. Germain, E., Robillard, P.N.: Engineering-Based Processes and Agile Methodologies for Software Development: A comparative case study. J. Syst. Softw. 75(1-2), 17–27 (2005)
18. Gendreau, O., Robillard, P.N.: Knowledge Conversion in Software Development. In: Nineteenth International Conference on Software Engineering and Knowledge Engineering (SEKE 2007), pp. 392–395 (2007)

An Electronic Learning System for Integrating Knowledge Management and Alumni Systems

Amnart Pohthong[1] and Panumporn Trakooldit[2]

[1] Information and Communication Technology Programme
[2] Software Engineering and Applications(SEA) Group, Department of Computer Science, Faculty of Science, Prince of Songkla University, Hat Yai, Thailand
amnart.p@psu.ac, pnon6ster@gmail.com

Abstract. Nowadays, most people worldwide accept the fact that knowledge is a very valuable asset for their success. Educational institutions such as universities, whose main objective is to educate students to gain more knowledge and skills for their careers as well as how to adapt and live in their community. Over the past decades, teaching and learning paradigms have rapidly changed from traditional styles to computer-based styles. Electronic learning (e-learning) has been recognized as an effective computer-based technology for knowledge management, life long learning, and distant learning. Although most universities have created electronic alumni networks and systems, those systems lack the concern of knowledge management. Therefore, in this paper, we propose an electronic learning system for integrating knowledge management and alumni systems.

Keywords: E-learning, Knowledge management, Alumni system.

1 Introduction

Nowadays, universities and colleges have become knowledge-intensive organizations. Teaching and learning processes in educational institutes are expected to educate their students to meet all requirements stated in their curricula. When these students complete their studies and become graduates, they bring their knowledge to apply in their careers or further studies, as well as in their lives in their community. M. Shih, J. Feng, and C-C Tsai also addressed learning and teaching perspectives from *South Africa's Draft white paper on education: transforming learning and teaching through ICT (2003)* and from the British government-commissioned report on *Teaching and Learning 2020* in [1] about the importance of developing learners' critical thinking, decision-making, problem-solving skills through collaborative learning environments and the focus on the differences of each individual learner.

Self development and individual learning would increase graduates' knowledge. However, the results of their knowledge improvement and competency depend on each individual potential and opportunity. Some alumni still need to be supported by their previous institutions. Therefore, knowledge management (KM) becomes a mechanism for knowledge sharing, especially in subject areas where knowledge

Á. Rocha et al. (Eds.): *Advances in Information Systems and Technologies*, AISC 206, pp. 11–21.
DOI: 10.1007/978-3-642-36981-0_2

changes rapidly such as Computer Science (CS) and Information Technology (IT). Knowledge management has also emerged from IT supports [2, 3].

Although most universities have developed their on-line alumni systems in order to link and communicate with their alumni, it still lacks the KM integration to those alumni systems. Therefore, the research reported in this paper proposes the framework for integrating knowledge management into the alumni system. The prototyped system can be used as an electronic learning system for knowledge sharing among alumni, current students, and institutional staff.

2 Electronic Learning and Knowledge Management

2.1 Electronic Learning

E-learning technologies have been widely used to deliver information, including learning materials, electronically from instructors to learners, especially for education and training. The results of the study reported in [4] confirm that instructors are willing to use e-learning environments to aid their teaching activities while learners also respond favorably to e-learning environments for complementing to their learning activities. Some critical factors influencing learner satisfaction for e-learning are reported in [5]. These factors can be grouped into six dimensions: learner, instructor, course, technology, design, and environmental dimensions. Among thirteen factors, seven factors were identified to be critical factors: learner computer anxiety, instructor attitude toward e-learning, e-learning course flexibility, e-learning course quality, perceived usefulness, perceived ease of use, and diversity in assessments. J. Andrade, J. Ares, R. Garcia, and S. Rodriguez suggested the three main blocks for organizing the elaboration process of e-learning actions: didactical material, follow-up and tutoring, and alternative learning [6]. J. Ismail also suggested the critical components for designing an e-learning system in [7]: learning management system, learning content design system, learning content management system, and learning support system.

2.2 Knowledge Management

Over the past decades, most universities worldwide have focused on data, information and knowledge as the important factors for their businesses. R.D. Corbin, C.B. Dunbar, and Q. Zhu highlighted in [8] that information results from the collection and assembly of “facts (data)” while knowledge involves the human intelligence traits. The terms of data and information management seem to gain better understanding than knowledge management (KM). KM is a systematic approach for capturing and creating, storing and accessing, validating, disseminating and applying knowledge to accomplish organizational goals and objectives [3]. Hence, most organizations expect that KM can lead to competitive advantages. KM in organizations relies on many systems and processes. In [9], Debowski suggests three types of organizational infrastructure for KM: managerial, technological, and social infrastructures.

KM activities can be addressed in various perspectives such as design, IT, management, artificial intelligence, and ontology perspectives [10]. Among several activities suggested in these perspectives, knowledge sharing has gained public attention in order to value and make use of both tacit and explicit knowledge as well as individual and organizational knowledge. Tacit knowledge can be transformed to be explicit knowledge and vice versa. However, knowledge sharing becomes more difficult than information sharing. Community of Practice (CoP) is one effective mechanism used to share knowledge among practitioners within the same organization or across different organizations.

Nowadays, IT tools have emerged to support knowledge sharing more efficiently than in the past, especially within and between CoP[11,12]. Many innovative technologies have been introduced and adopted to leverage KM such as knowledge portals [13,14] and e-learning [15,16].

3 System Framework and Implementation

3.1 Case Study

The Department of Computer Science at Prince of Songkla University in Thailand was selected as a case study for employing the proposed system. The department firstly recruited graduate students in 1986 and undergrad students in 1991. At present, there are around 1400 alumni. The current alumni system has only news and photo galleries.

3.2 Alumni System Components

The requirements for an alumni system are common to many universities, for example, the following alumni systems appear on their universities' web sites.

- Harvard medical school alumni system consists of news, events, alumni council, community, alumni benefits, and giving [17].
- Harvard business school alumni system consists of career development, reunions, travel, events, boards & volunteers, giving to HBS, FAQs, Clubs, Bulletin, and tools [18].
- MIT alumni system consists of alumni association, networks, benefits & services, volunteering, learn, travel, news & views, MIT students, parents association, and giving to MIT [19].
- University of Cambridge, UK., the Cambridge alumni relations office (CARO)'s main page consists of news, Olympics 2012, alumni events, alumni groups, alumni benefits, travel programme, and contact us [20].

3.3 Knowledge Management Components

The components of web-based knowledge management system suggested by Debowski in [9] are as follows:

(1) Business process management
(2) Content management
(3) Web content management
(4) Knowledge applications management

3.4 System Analysis and Proposed Framework

From the basic requirements for an alumni system and the components of web-based knowledge management, more requirements were gathered during the system analysis phase from the following groups of related users at the case study site.

(1) Alumni
Alumni can access provided courses, share learning media and knowledge. However, these media and knowledge will be evaluated by the departmental executive or committee before being deposited into a repository. Alumni can also manage some attributes of their personal data such as a degree earned and work experience.
(2) Students
Students can access courses provided on the departmental virtual class room (VCR) or those provided on the university learning management system called LMS@PSU. They can also access the knowledge repository via the knowledge sharing space.
(3) Support staff
Support staff can manage alumni profiles, news and activities, the departmental web-board and social networks.
(4) Teachers/Lecturers
In addition to manage their courses on the VCR or LMS@PSU, teachers or lecturers can manage courses for alumni, evaluate shared media and knowledge.
(5) Executive/Committee
The departmental executives and student affair committee can do the same activities similar to teachers and lecturers. They can also access some reports or some graphs of students and alumni statistics.

The main activities for the proposed system are specified by using the use-case diagram as shown in Fig.1. From information analysis, the proposed framework for an integrating KM and alumni system via an e-learning system was designed into six main parts as shown in Fig.2.

(1) Virtual class room (VCR)/Learning Management System (LMS)
(2) Course Materials for Alumni
(3) Alumni Competency and Learning Material Evaluation
(4) Knowledge Sharing Space
(5) Knowledge Repository
(6) Alumni Profile / Personal Profile

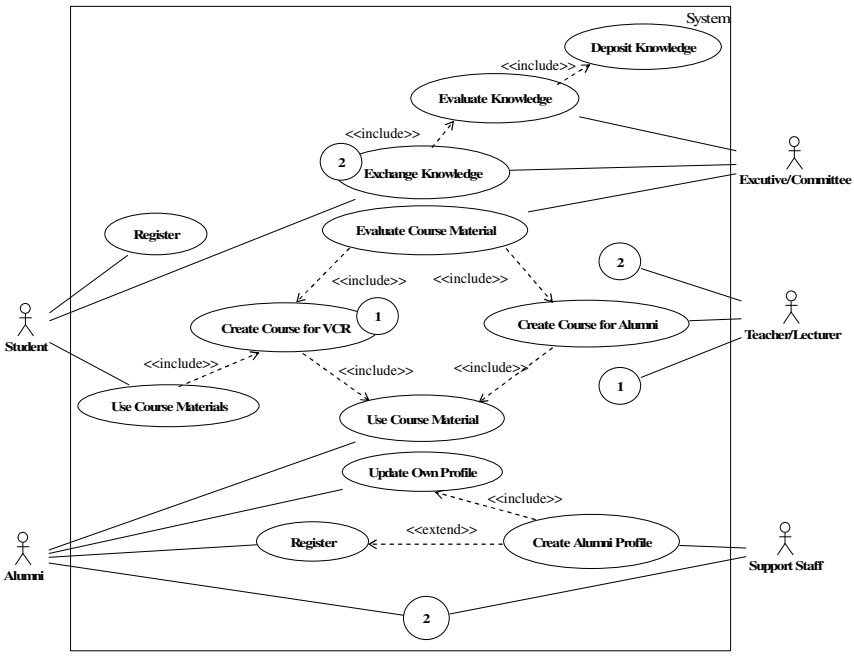

Fig. 1. The use-case diagram for the proposed system

3.5 System Design and Implementation

From the proposed framework, the system architecture was designed as shown in Fig.3. It consists of thirteen components: (1) *Membership* for managing the system users, (2) *AlumniData* for managing alumni data and profiles, (3) *Associate* for managing alumni groups, (4) *CourseforAlumni* for managing courses, (5) *Materials* for managing shared materials, (6) *Knowledge* for managing shared knowledge, (7) *KnowledgeEvaluation* for managing knowledge evaluation, (8) *Repository* for managing knowledge repository, (9) *YearStat* for managing statistical reports, (10) *News* for managing news, (11) *Event* for managing events, (12) *Webboard* for managing alumni webboard, and (13) *SocialNetwork* for managing social networks. The system users were managed into two main groups: members and non-members. The system members consist of alumni, current students, support staff, teachers/lecturers, executive/committee, and system administrators while non-members are end-users. The system database and its relational schemas were designed using Entity-Relationship (E-R) technique as shown in Fig.4.

The proposed system was prototyped as a web-based system. Some examples of user interfaces in the prototyped system such as the main menu, and the menu of knowledge submission by alumni are shown as Fig.5-6.

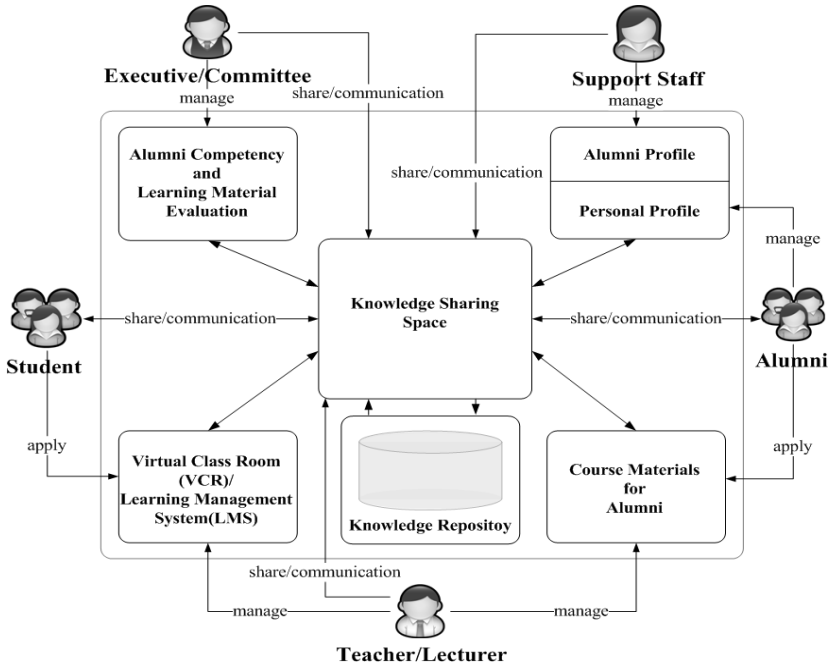

Fig. 2. The proposed framework for the prototyped system

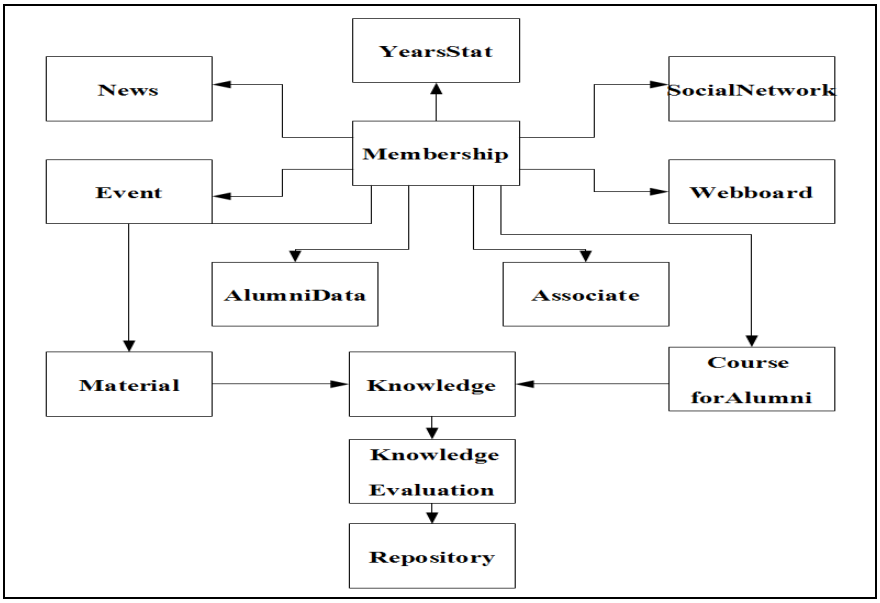

Fig. 3. The system architecture

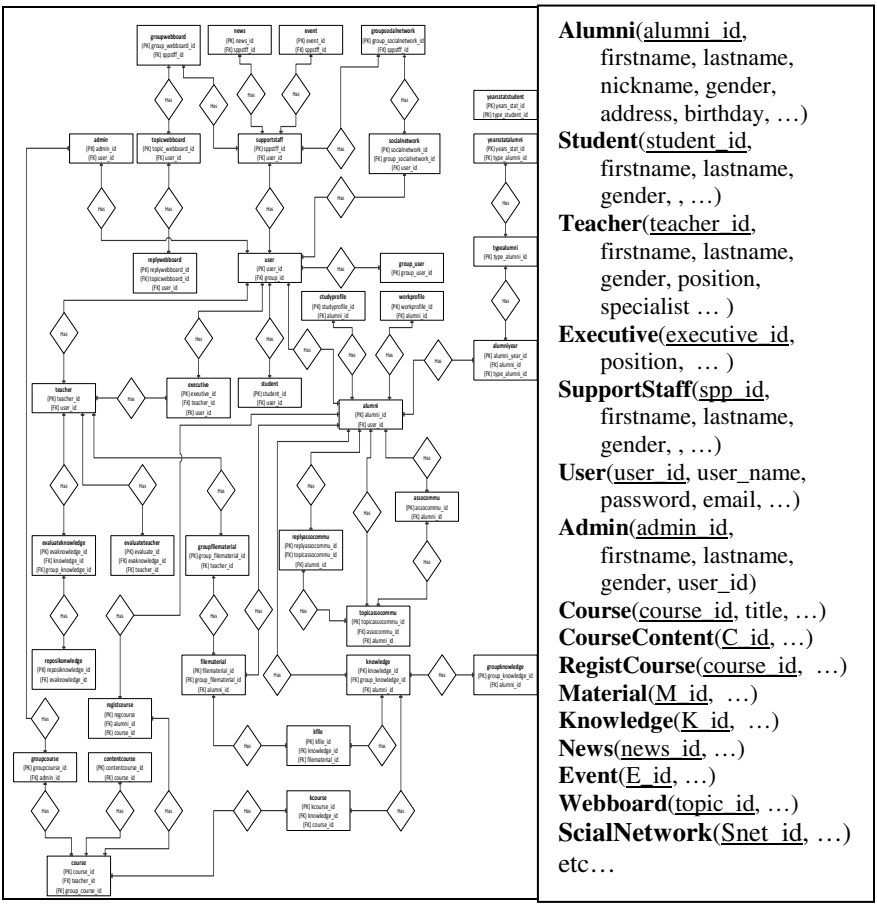

Fig. 4. The database structure for the system

4 System Evaluation

The system was evaluated in our laboratory for users' satisfaction. Thirty-five subjects were voluntarily selected. They consisted of one departmental executive who was in charge with the deputy head of department for students' affair, three lecturers, thirty alumni, and one support staff who was in charge of the departmental alumni system.

The evaluation tools were (1) the prototyped system (2) an introductory tutorial for the system (3) the given problems and (4) the questionnaire. The given problems corresponded to the type of users and consisted of several questions, allowed the subjects to explore and use the most related features of the system. The questionnaire consisted of three parts: personal data, questions for rating satisfaction level using Likert scale (from 1=very low to 5=very good), and open suggestions. Content validity in the questionnaire was evaluated by three experts and then calculated the

index of item-objective congruence (IOC). Questions that have the IOC value from 0.6 to 1.00 were included in the questionnaire. The revised questionnaire was used with five dry-run subjects in order to evaluate its reliability, using the Alpha-Coefficient method. The Alpha-Coefficient value was 0.91, showing that the questionnaire has good reliability.

The evaluation procedures started with giving a brief introduction about the system and its evaluation method as in the introductory tutorial for the system, taking around 10 minutes. Then, the subjects performed the given problems. Finally, all subjects completed the questionnaires.

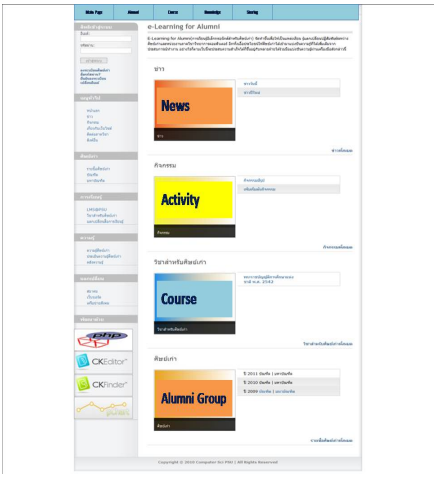

Fig. 5. The system main menu

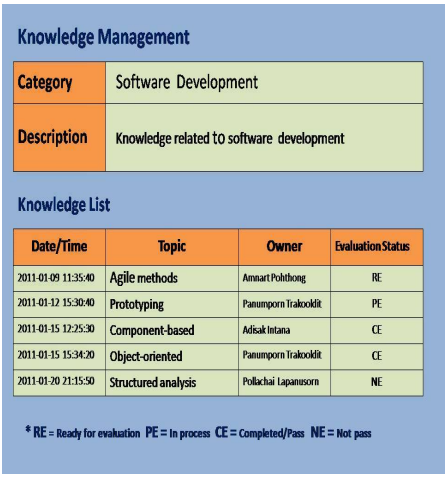

Fig. 6. The menu of knowledge submission by alumni

The results of system evaluation performed by two main groups of subjects: (1) members of staff (consisting of teachers/lectures, executive/committee, and support staff), and (2) alumni, are described as shown in Table 1 and Table 2.

Table 1. The result of each evaluation item performed by the departmental members of staff

Evaluation Items	Average	SD	Satisfaction Level
1. System Usage			
1.1 Convenience of data entry	3.2	0.55	Moderate
1.2 Convenience of data modification	3.2	0.45	Moderate
1.3 Automatic prevention of input errors	3.0	0	Moderate
1.4 Search support	3.4	0.55	Moderate
1.5 Appropriation of presentation sequence	3.4	0.55	Moderate
1.6 Completeness of requirements	3.0	0.70	Moderate
1.7 Appropriation of screen design	3.8	0.84	Good
1.8 Clarity of communication	3.4	0.89	Moderate

Table 1. (*continued*)

2. Efficiency			
2.1 Accuracy of data processing	3.2	0.84	Moderate
2.2 Accuracy of data retrieval	3.4	0.89	Moderate
2.3 System sustainability	3.0	0.71	Moderate
2.4 Efficiency of data storage	3.6	0.55	Good
3. Access ability			
3.1 Coverage of users	4.0	0.71	Good
3.2 Security	3.8	0.84	Good
4. System Value			
4.1 Learning support	3.8	0.84	Good
4.2 Organizational development	3.4	0.55	Moderate
4.3 Knowledge resource	3.6	0.89	Good
4.4 Knowledge sharing	3.6	0.89	Good
5. Overall Satisfaction	3.4	0.55	Moderate

Table 2. The result of each evaluation item performed by the departmental alumni

Evaluation Items	Average	SD	Satisfaction Level
1. System Usage			
1.1 Convenience of data entry	4.5	0.51	Very Good
1.2 Convenience of data modification	3.6	0.55	Good
1.3 Automatic prevention of input errors	3.4	0.72	Moderate
1.4 Search support	3.6	0.57	Good
1.5 Appropriation of presentation sequence	3.5	0.51	Good
1.6 Completeness of requirements	3.6	0.63	Good
1.7 Appropriation of screen design	3.6	0.57	Good
1.8 Clarity of communication	3.5	0.58	Good
2. Efficiency			
2.1 Accuracy of data processing	4.5	0.51	Very Good
2.2 Accuracy of data retrieval	3.5	0.57	Good
2.3 System sustainability	3.6	0.57	Good
2.4 Efficiency of data storage	3.6	0.57	Good
3. Access ability			
3.1 Coverage of users	3.5	0.57	Good
3.2 Security	3.6	0.57	Good
4. System Value			
4.1 Learning support	3.4	0.57	Moderate
4.2 Organizational development	4.5	0.51	Very Good
4.3 Knowledge resource	4.5	0.51	Very Good
4.4 Knowledge sharing	3.8	0.68	Good
5. Overall Satisfaction	3.5	0.51	Good

5 Conclusion

An electronic learning system for alumni was proposed. This system allows alumni to learn new knowledge and share their knowledge with other groups of users, especially with the departmental executives and lecturers. The system was prototyped for the case study at the Department of Computer Science, Prince of Songkla University, Thailand. The system was evaluated by thirty-five subjects for users' satisfaction in four categories: system usage, efficiency, access ability, and system value. The overall results of users' satisfaction were rated as moderate quality by the members of staff and as good quality by alumni. In future, the proposed system should share students' data with the university database, since current students will eventually become alumni in order to evaluate alumni competency after their graduation.

Acknowledgments. We wish to express our sincere thanks to all experts and all who acted as our subjects in both the dry run and the experiment.

References

1. Shih, M., Feng, J., Tsai, C.-C.: Research and trends in the field of e-learning from 2001 to 2005: a content analysis of cognitive studies in selected journals. Computers & Education 51, 955–967 (2008)
2. Liao, S.-H.: Knowledge management technologies and applications – literature review from 1995 to 2002. Expert Systems with Applications 25, 155–164 (2003)
3. Carnes, W.B.: Knowledge mangement: a stste of the practice summary. In: IEEE 7th Human Factors Meeting, Scottsdale Arizona (2002)
4. Liaw, S.-S., Huang, H.-M., Chen, G.-D.: Surveying instructor and learner attitudes toward e-learning. Computers & Education 49, 1066–1080 (2007)
5. Sun, P.-C., Tsai, R.J., Finger, G., Chen, Y.-Y.: What drives a successful e-learning? An empirical investigation of the critical factors influencing learn satisfaction. Computers & Education 50, 1183–1202 (2008)
6. Andrade, J., Ares, J., Garcia, R., Rodriguez, S.: Guidelines for the development of e-learning systems by means of proactive questions. Computers & Education 51, 1510–1522 (2008)
7. Ismail, J.: The design of e-learning system beyond the hype. The Internet and Higher Education 4, 329–336 (2002)
8. Corbin, R.D., Dunbar, C.B., Zhu, Q.: A three-tier knowledge management scheme for software engineering support and innovation. The Journal of Systems and Software 80, 1494–1505 (2007)
9. Debowski, S.: Knowledge Management, John Wiley & Sons Australia, Ltd. (2006)
10. Lai, L.F.: A knowledge engineering approach to knowledge management. Information Sciences 177, 4072–4094 (2007)
11. Pan, S.L., Leidner, D.E.: Bridging communities of practice with information technology in pursuit of global knowledge sharing". Strategic Information Systems 12, 71–88 (2003)
12. Chua, A.: Knowledge management system architecture: a bridge between KM consultants and tecnologists. International Journal of Information Management 24, 87–98 (2004)

13. Baalen, P.V., Bloemhof-Ruwaard, J., Heck, E.V.: Knowledge sharing in an emerging network of practice: the role of a knowledge portal. European Management Journal 23(3), 300–314 (2005)
14. Fernandes, K.J., Raja, V., Austin, S.: Portals as a knowledge repository and transfer tool – VIZCon case study. Technovation 25, 1281–1289 (2005)
15. Chen, R.-S., Hsiang, C.-H.: A study on the critical success factors for corporations embarking on knowledge community-base e-learning. Information Sciences 177, 570–586 (2007)
16. Weller, M., Pegler, C., Mason, R.: Use of innovative technologies on an e-learning course. The Internet and Higher Education 8, 61–71 (2005)
17. Harvard Medical Alumni Association, http://alumni.hms.harvard.edu/
18. Harvard Business School Alumni, http://www.alumni.hbs.edu/
19. MIT Alumni Association, http://alum.mit.edu/
20. University of Cambridge, UK., alumni relations office (CARO), http://my.alumni.cam.ac.uk/

Knowledge Management Systems and Intellectual Capital Measurement in Portuguese Organizations: A Case Study

Mário Pinto

ESEIG, Polytechnic Institute of Porto, Portugal,
KMILT Research Group
mariopinto@eu.ipp.pt

Abstract. This paper presents the results of an exploratory study on knowledge management in Portuguese organizations. The study was based on a survey sent to one hundred of the main Portuguese organizations, in order to know their current practices relating knowledge management systems (KMS) usage and intellectual capital (IC) measurement. With this study, we attempted to understand what are the main tools used to support KM processes and activities in the organizations, and what metrics are pointed by organizations to measure their knowledge assets.

Keywords: knowledge management systems, intellectual capital, intangible assets, metrics.

1 Introduction

Knowledge and Knowledge management (KM) are increasingly recognized as a key driver to innovation, competitive advantage and future sustainability [1], [2], [3]. In the new economy, knowledge based resources can be considered the main source of value creation [4], [5]. The competitiveness of organizations as well as their ability to develop distinctive capabilities of its competitors, is closely related with their capacity to create, store, share and apply their knowledge assets [4], [6].

In this context, KMS play a role of increasing importance. These systems contribute to support organizational processes and activities, which enable the knowledge sharing and knowledge application across organizations [7]. KMS also increase communication and collaboration, promoting a culture of knowledge sharing, and managing knowledge as a crucial asset for the organization [23]. Despite its importance in the modern economy, these intangible assets are not yet clearly measured and reported. Measuring these intangible assets shows their impact in value creation and its benefits for organization [5], [7], [8]. According some authors, evaluating the economic impact of knowledge in organizations, i.e., the Intellectual Capital (IC) measurement is a key issue in KM [5], [9], [11].

The aim of this paper is to know what Portuguese organizations are doing in terms of KM practices, namely KMS usage and IC measurement. A survey was made with two main purposes: i) identify the KMS used by Portuguese organizations; ii) identify the metrics specified to measure the main components of IC. A brief literature review

Á. Rocha et al. (Eds.): *Advances in Information Systems and Technologies*, AISC 206, pp. 23–32.
DOI: 10.1007/978-3-642-36981-0_3

about IC measurement and KMS is made in the second section of this paper, as the background of this study. Section three describes the research methodology used in this study, while section four shows the results obtained and presents a brief discussion of them. The fifth section provides some conclusions and draws some directions for future research.

2 Background

2.1 Intellectual Capital Measurement Models

Numerous definitions of IC have been proposed, focusing IC as knowledge that can be converted into value [9], intellectual material [10] or combined intangible assets which enable the company to function [11]. However, almost all definitions have three common elements [12]: i) intangibility; ii) knowledge that can create value; iii) effect of collective practice. From these perspectives it is possible to describe IC as intangible assets that may be used as a source of sustainable competitive advantage, creating wealth in organizations.

There is a general agreement that intangible assets may be decomposed in a set of components. Almost all authors refer to IC as consisting on a set of human, relationship and structural capital [9], [12], [13]:

- Human capital is concerned with individual capabilities, knowledge, skills, experience and abilities to solve problems. It represents the employee's competence, attitude and intellectual agility [14], [15]. Competences include skills and education, while attitude covers the behaviour of the employees. Intellectual agility enables to think on innovative solutions and to change practices in order to solve problems [12].
- Structural capital is concerned with systems, organizational processes, technologies, concepts and models of how business operate, databases, documents, patents, copyrights and other codified knowledge. According to Roos [16], structural capital is what remains in the company when employees go home at night.
- Relationship capital is concerned with alliances and relationships with customers, partners, suppliers, investors and communities. It also includes brand recognition, organization image and market position. The relationship capital represents the knowledge embedded and the value added from the relationships with other external entities [17].

While IC represents the intangible assets that brings competitive advantage and value creation to the organization, its measurement reflects the influence and the impact of these assets in the organization [9], [12], [14]. Measuring the knowledge value and their impact in the organizations is a growing area of interest in the KM field, which reflects the value added by knowledge to the organizations and enables to monitor the performance of the knowledge resources and KM activities [18].

According to Luthy and Williams [2], [8], [19] there are two general approaches for measuring IC and its main components:

- Direct Intellectual Capital Methods, which estimate the monetary value of IC by identifying its various components. Once these components are identified, they can be evaluated, either individually or as an aggregated coefficient. It represents an attempt to fill the gap between market and book value.
- Scorecard Methods, which identify the knowledge resources that bring value added. Metrics for measuring these knowledge resources are reported in scorecards or graphs, giving a more detailed picture of the value of knowledge in organization. No estimates are made of monetary value of IC.

Tables 1 summarizes a review of the IC measurement models grouping them according Williams classification [20]:

Table 1. IC measurement models review

IC Approaches	IC Measurement Model	Authors
Scorecard Methods	Skandia Navigator	Edvinsson e Malone
	Balanced Scorecard	Kaplan & Norton
	Intangible Assets Monitor	Sveiby
	Intelect Model	Euroforum
	Intellectual Capital Index	Roos & Edvinsson
	Nova Model	Camisón, Palácios et al.
	Intangible Value Framework	Allee
	IC Rating	Edvinsson
	Intellectual Capital Rating	Joia
	Heng Model	Heng
	Meritum Guidelines	Meritum Guidelines
	Danish Guidelines	Mouritzen & Bukh
	Value Chain Scoreboard	Lev
	Chen, Zhu & Xie Model	Chen, Zhu & Xie
	VAIC	Pullic
	Intellectus	IADE & CIC
Direct Intellectual Capital Methods	Technology Broker	Brooking
	Citation-Weighted Patents	Bontis
	Inclusive Valuation Methodology	M´Pherson & Pike
	Total Value Creation	Anderson & McLean
	The Value Explorer	Andriessen & Tissen
	The 4-Leaf Model	Leliaert, Candries et al.
	Value Added Intellectual Coefficient	Pullic

2.2 Knowledge Management Systems

KMS are systems developed with the purpose of supporting KM processes, namely knowledge creation, storage and retrieval, knowledge transfer and application, as well as the flows between them [7], [21]. These systems enable an environment that facilitates the creation of knowledge, its sharing and application, and also the

communication and collaboration among the organization employees. More than technological tools, the KMS could be viewed as virtual spaces that promote knowledge conversion between explicit and tacit dimensions of knowledge [24]. According to Nonaka [24], Knowledge conversion from one form to another occurs frequently and leads to the creation of new knowledge.

Not all KMS are based on technologies. A non-virtual community of practice or a face meeting are ways to create and share knowledge. However, nowadays almost all KMS are based in information technologies. The amount of knowledge that needs to be captured, stored and shared, the geographic distribution of people and the dynamic knowledge evolution make the use of technology a necessity [21].

Many authors have written about the use of different types of KMS [21], [22], [23], [25], [26]. The variety of classifications referred by these authors takes us to develop a systematization of KMS categories [15], regarding their addressed issues, capabilities and functionalities [4], [27]. Table 2 summarizes this categorization, presenting the KMS categories considered and their main functionalities:

Table 2. Knowledge management systems categorization

Categories	Main functionalities
Document management systems (knowledge repositories)	Document management; edition collaboration; versions control; documents sharing; support for all content types (text, audio, video, graphs, xml, web, etc.); searching and retrieval advanced mechanisms.
Knowledge maps	Categorizing and indexing knowledge in taxonomies; creating knowledge maps; pointing to organizational knowledge; inserting tags and labels in documents; alerting to relevant information.
Collaboration systems (groupware)	Synchronous or asynchronous communication; process and people collaboration; virtual meetings; instant messenger, videoconference; real-time conversation; grouping calendar and scheduling, etc.
Workflow systems	Business processes automation; support automated flows of activities, tasks and information; support documental flows.
Business intelligence and Data mining tools	Statistical, OLAP analysis; reveal patterns and hidden relationships between data; generate new knowledge from existing one; query and reporting tools; data mining and data warehousing tools.
Expert systems	Expert identification; connect users with experts to solve certain problems; ask questions, provide recommendations and explain logical processes; capture and store new questions and rules in a knowledge base.
Competence management	Employees profiles; experts, customers, vendors or others profiles in some systems; competence maps; individual competence analysis; training programs recommendation based on employees skills; recruitment and selection support.

Table 2. (*continued*)

E-learning systems	Environment personalization; evaluation and progress tracking: exercises quiz and tests; collaboration tools; reusable learning and object libraries; support different types of contents: text, audio, video, etc.; classes' workgroups; authoring, scheduling and reporting tools; searching and matching tutorials.
Help-desk systems	Self-desk and help-desk functionalities; FAQs access and maintenance; on-line customer support; expert help; customer profiles; customers queries.
Corporative portals	Environment personalization; filtering relevant information; search and retrieval advanced mechanisms; news, activities, tasks and calendar management; unified access environment to other tools: documents management, workflow, knowledge maps, groupware, etc.; integration with other applications.
Web 2.0 tools	Interaction, collaboration, participation of people: blogs, wikis, social bookmarking, tagging, platforms for content sharing.

3 Research Methodology

A survey was made with the purpose of knowing the current practices of the Portuguese organizations, regarding KMS usage and IC measurement. With this study we seek to understand which type of tools are most used by Portuguese organizations in supporting KM processes, and what metrics they generally use to measure knowledge resources.

The survey was based on a questionnaire, sent to one hundred of the main Portuguese organizations. The organizations were selected from a publication that produces an annual ranking of companies, based on their value creation for the Portuguese economy. The questionnaire was sent to the director of the knowledge management department or information systems department (when the first did not exist in the company). With the questionnaire one letter was also sent, explaining the concepts of KMS, IC and their main components: human, structural and relationship capital; also explaining the aims of the study, assuring confidentiality and requesting collaboration from the most suitable person in the organization. The questionnaire was structured in three main sections:

- Organization identification: it includes the organization name and business area.
- Knowledge management systems identification: It comprises the identification of KMS categories used in the organization. The questionnaire presents the eleven KMS categories described in table 2 and the organizations could select the adequate categories or add new ones.
- Intellectual capital metrics identification: It comprises the identification of the metrics used in the organization to measure IC and their components: human, structural and relationship capital. The questionnaire contains a comprehensive list of qualitative and quantitative metrics, resulting from an extensive review of IC measurement models [15]. However, the respondents could also complete this list, adding the metrics used in their organizations.

An extensive review of survey studies shows that some forms of follow-ups can increase response rates [12], [29]. According some authors [28] the resistance response rate is continuously increasing during the survey: it is relatively high at first, drops for a short period after the follow-ups and then starts to increase. Thus, two follow-ups were carried out using letters, telephone calls and e-mails. The delay between these two follow-ups was six weeks. Twenty-one valid questionnaires answers were received, corresponding to a response rate of 21%.

4 Results and Discussion

From the questionnaires received, six were from service organizations (e.g. telecommunications, energy) and fifteen were from industry organizations (e.g. automobile industry, electronic, pneumatics). All of the respondent organizations have identified a set of KMS, used to support knowledge processes, but only twelve organizations have specified a set of metrics to measure IC assets. The remaining organizations said that they didn't make a regular management and measurement of intangible assets.

4.1 Knowledge Management Systems

One issue addressed in the survey was the use of KMS in Portuguese organizations. Table 3 summarizes the several categories mentioned by respondents in the questionnaire, presenting the respective occurrence rate.

Table 3. Knowledge management systems usage in Portuguese organizations

KMS Categories	Rate
Business Intelligence and Data mining tools	67%
Knowledge Maps	25%
Document Management Systems (repositories)	75%
Collaboration Systems (groupware)	43%
Workflow Systems	50%
Expert Networks	13%
Competence Management Systems	55%
E-learning Systems	25%
Help-desk tools	75%
Corporative Portals	67%
Web 2.0 tools	75%

Based on the results presented above, some important conclusions can already be drawn:

- The findings presented in table 3 point that KMS supporting mainly explicit knowledge were most mentioned by respondents. Document management systems, business intelligence, competence management and help-desk systems are examples of tools that lead mainly with explicit to explicit conversion of knowledge, according to Nonaka knowledge conversion model [24].

- Expert systems, knowledge maps and e-learning systems were tools with a reduced response rate. According the questionnaire answers, these tools that mainly support personal knowledge and tacit to explicit conversion of knowledge, they have a low dissemination in Portuguese organizations.
- Help-desk tools are also strongly referred in the survey. The relationship with customers and other external entities is crucial to obtain competitive advantages in a global economy. Thus, the result obtained shows the importance recognized to relationship capital and the need to satisfy the customer's needs.

It is interesting to note that none of the respondents have identified other categories of tools, beyond those mentioned in the questionnaire.

4.2 Intellectual Capital Measurement

Another issue addressed in this study was the metrics used by Portuguese organizations to measure IC. Table 4 summarizes the metrics mentioned by respondents, grouping them by IC component: human, relationship and structural capital.

Table 4. Summary of IC metrics survey

	Metrics	Metrics
Human Capital	▪ Training programs (days per year)	▪ Investment in training (per capita)
	▪ Employees in training plans (%)	▪ Employee turnover
	▪ Execution rate of annual training plan	▪ Value added per capita
	▪ Duration of training plans (average)	▪ Full-time employees (%)
	▪ Average level of academic degree	▪ Part-time employees (%)
	▪ Employees satisfaction (index)	▪ Specialized/expert employees (%)
	▪ Average duration of employees relationship	▪ Innovative employees (new ideas)
	▪ Age distribution of employees	▪ Employees with initiative (new ideas)
	▪ Absenteeism rate	▪ IT literacy skills (average)
	▪ Internships (number)	▪ Profits by employee
Relationship Capital	▪ Grow rate of customer's portfolio	▪ Investment Information Technologies
	▪ % of small, medium and large customers	▪ Investment in marketing
	▪ Profitability by costumer (average)	▪ New customers/customers lost (rate)
	▪ Bill per customer (average)	▪ Annual sales per customer
	▪ Customer's satisfaction index	▪ Market share in segment
	▪ Customers claims (number)	▪ Business alliances and partnerships
	▪ New contracts / proposals (rate)	▪ Average duration of customer relationship
	▪ Delay in delivery orders (average)	▪ Customers contacts (number)

Table 4. (*continued*)

	Metrics	Metrics
Structural Capital	▪ Quality certifications (number)	▪ New business generated
	▪ Processes in non-conformity	▪ Innovation and creativity capabilities (new products/services, upgrades)
	▪ Certified products (number)	▪ New products launched
	▪ Quality tests performed	▪ Revenue generated by new products / total revenue
	▪ Key-process documented	▪ Number of customers per employee
	▪ Continuous improvement projects	▪ Business partners (number)
	▪ Investment in developing new skills	▪ Computers per employee
	▪ Investment in training programs	▪ Knowledge management initiatives (#)
	▪ Investment in new products/services	▪ Protocols with Innovation entities (#)
	▪ Investment in new processes	▪ Suggestions from employees accepted by administration
	▪ Investment in Information Technologies	▪ Productivity index
	▪ Investment in R&D	▪ Time response to customers' requests
	▪ Administrative expense/employee	▪ Time to processing payments
	▪ Administrative expense/total revenues	

The results presented in the previous tables, allow us to draw the following considerations:

- The metrics focused on the human capital component are, mostly, related with the characterization of the employee's profile, and their effort in training programs. Few metrics are focused on measuring the value added by employees and their contribution to the organizational knowledge. Most of the metrics translate the effort invested in activities (training plans, for example) rather than the results obtained.
- The structural capital includes a few number of metrics for measuring research and development activities. The questionnaire answers do not include metrics for measure the innovative capability, for instance, the number of new products/services developed, the number of new ideas that generate new products or services and the number of new patents registered.
- Although Portuguese organizations define relationship capital as valuable relations with external entities, including customers, suppliers, partners, investors and other entities, they measure basically the customer capital. The metrics employed are almost all related with customers, ignoring the value of external relationships with, for instance, innovation entities, governmental departments, investors or business partners.

5 Conclusions

The KMS categories more mentioned by the respondents, namely business intelligence, document management systems (knowledge repositories), competence management and workflow systems, are focused on supporting explicit knowledge. Help-desk systems could offer support to both explicit and tacit knowledge. These results are in compliance with the IC metrics addressed in the survey questionnaire. Almost all metrics are aligned with the measurement of explicit knowledge, and they can be provided from the tools above mentioned. Almost all metrics referred by organizations to evaluate human capital, for example, could be found on competence management systems, which manage knowledge related with human competences and skills. On the other hand, a significant number of metrics pointed by organizations to measure relationship capital, could be obtained from explicit knowledge wrapped in help-desk or business intelligence systems. The results of the survey show that current practices relating KMS usage and IC measurement are in compliance, focusing mainly the management and the evaluation of the explicit knowledge, easier to represent, codify and share, than tacit knowledge.

The findings of this study do not allow us to obtain sustainable results. Unfortunately, a considerable number of organizations, in Portugal, still do not have a culture of manage their knowledge assets and haven't a knowledge management infrastructure based on KM tools. Future work comprises a more deeply study in some organizations that have answered to the questionnaire.

References

1. Bontis, N., Curado, C.: managing Intellectual capital: the MIC matrix. International Journal of Knowledge and Learning 3(2/3) (2007)
2. Leliaert, C., et al.: Identifying and managing IC: a new classification. Journal of Intellectual Capital 4(2), 202–214 (2003)
3. Ammann, E.: A Hierarchical Modelling Approach to Intellectual capital development. Electronic Journal of Knowledge Management 8(2), 181–192 (2010)
4. Pinto, M., Lopes, A., et al.: A Framework for Characterizing Knowledge Management Systems. In: 6th European Conference on Knowledge Management, Limerick, Ireland, pp. 442–450 (2005)
5. Cabrita, M., Vaz, J., Bontis, N.: Modelling the creation of value from intellectual capital: A Portuguese banking perspective. International Journal of Knowledge and Learning 3(2/3), 266–280 (2008)
6. Martí, J.: In Search of an Intellectual Capital General Theory. Electronic Journal of Knowledge Management 1(2), 213–226 (2003)
7. Alavi, M., Leidner, D.: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. MIS Quarterly 25(1), 107–136 (2001)
8. Fritzsche, A., Germany, C.: Implicit Evaluations of Intellectual Capital. Practical Decision Making 10(3), 236–243 (2012)
9. Edvinsson, L., Malone, M.: Intellectual Capital: Realizing your Company's True Value by finding its Hidden Brainpower. Harper Business, New York (1997)

10. Stewart, T.: Intellectual Capital - The New Wealth of Organization. Doubleday, New York (1997)
11. Brooking, A.: Intellectual Capital: Core Assets for the Third Millennium Enterprise. Thomson Business Press, London (1996)
12. Cabrita, M., Vaz, J.: Intellectual Capital and Value Creation: Evidence from the Portuguese Banking Industry. The Electronic Journal of Knowledge Management 4(1), 11–20 (2006)
13. Suciu, M., Piciorus, L., Imbriscã, I.: Intellectual Capital, trust, cultural traits and reputation in the Romanian education system. Electronic Journal of Knowledge Management 3, 223–235 (2012)
14. Leliaert, C., et al.: Identifying and managing IC: a new classification. Journal of Intellectual Capital 4(2), 202–214 (2004)
15. Pinto, M., Lopes, F., et al.: Knowledge Management Systems and Intellectual Capital Measurement. In: Proceedings of XVI ISPIM Annual Conference, Porto, Portugal (2005)
16. Roos, J., Dragonetti, N., et al.: Intellectual Capital: Navigating in the New Business Landscape. Macmillan, New York (1997)
17. Bontis, N., Keow, W., et al.: Intellectual capital and business performance in Malaysian industries. Journal of Intellectual Capital 1(1), 85–100 (2000)
18. Mouritzen, J.: Overview intellectual capital and the capital market: the Circulability of Intellectual Capital. Accounting, Auditing & Accountability Journal 16(1), 18–30 (2003)
19. Andriessen, D.: Making sense of intellectual capital designing a method for the valuation of intangibles. Butterworth-Heinemann, Oxford (2004)
20. Williams, M.: Is a company's intellectual capital performance and intellectual capital disclosure practices related? In: McMaster's Intellectual Capital Conference, Toronto, Canada (2000)
21. Lindvall, M., Rus, I., et al.: Software systems support for knowledge management. Journal of Knowledge Management 7(5), 137–150 (2003)
22. Baroni, R.: Aplicações de Softwares de Gestão do Conhecimento: tipologia e usos. Ciência da Informação. Blo. Horizonte, Brazil, UFMG University (2000)
23. Abdullah, R., Selamat, M.: A framework for knowledge management system implementation in collaborative environment for higher learning institution. Journal of Knowledge Management Practice 6 (2005)
24. Nonaka, I., Kono, N.: The Concept of BA: Building a Foundation for Knowledge Creation. California Management Review 40(3), 40–54 (1998)
25. Loureiro, J.: Gestão do Conhecimento. V. N. Famalicão, Centro Atlântico (2003)
26. Nantel, R.: Knowledge Management Tools and Technology 2004: 35 Systems to Maximize Your Organization´s Intellectual and Human Capital. Brandon- hall.com (2003)
27. Zhou, A., Fink, D.: The Intellectual Capital Web: a systematic linking of intellectual capital and knowledge management. Journal of Intellectual Capital 4(1), 34–48 (2003)
28. Fortin, M.: Le processus de la recherce: de la conception à la realization. Décarie Éditeur (1996)
29. Birch, M., Uccardi, M., et al.: Conduction Background Investigations. In: APTA, Standards Development Program, American Public Transportation Association (2011)
30. Bebensse, T., Helms, R., Spruit, M.: Exploring Web 2.0 Applications as a Mean of Bolstering up Knowledge Management. The Electronic Journal of Knowledge Management 9(1), 1–9 (2011)

Semantic Patent Information Retrieval and Management with OWL

Maria Bermudez-Edo, Manuel Noguera, José Luis Garrido, and María V. Hurtado

University of Granada. Department of Software Engineering, E.T.S.I.I.,
c/ Saucedo Aranda s/n, 18071 Granada, Spain
{mbe,mnoguera,jgarrido,mhurtado}@ugr.es

Abstract. Patent information is mainly represented and stored in databases containing large amounts of information about the inventions and metadata of patents such as the technological field to which they belong, which can be retrieved in standard formats such as CSV or XML. These, however, provide few semantics to enable further relationships among patents to be inferred for analysis purposes. Ontologies, mostly represented in the Web Ontology Language (OWL), are increasingly being developed to represent, manage and reason about data in information systems. Unfortunately, the current patent ontologies do not seem to fully capture the implicit hierarchies present in patent technology codes to exploit the information that can be derived from the formal representation of patent code classification hierarchies through logic reasoning. This paper presents an approach to automatically index hierarchical codes with ontological categories and enrich the information retrieved and knowledge management from different patent repositories with new relationships, properties and inferred information.

Keywords: Information and knowledge management, Ontology, OWL, XML, eXtensible Stylesheet Language Transformations (XSLT), Patent.

1 Introduction

Patents have a huge impact on national and international economies and represent a great part of all the scientific and technological knowledge worldwide [1]. Several studies have in fact used patents to measure the innovative capacity of firms [2] [3] and even technological trends [4] [5].

Patents are usually stored in large databases which belong to the different patent offices around the world, e.g. the European Patent Office (EPO) and the United States Patent and Trademark and Office (USPTO). While most of these databases have become available online in recent years, they exhibit different datasets and data structures for patent representation and this makes it difficult to automate their processing.

Patent metadata comprise different types of information such as the name of the patent applicant, publication date and the technological patent classification. Each patent database defines a set (or sets) of technological codes according to hierarchical

Á. Rocha et al. (Eds.): *Advances in Information Systems and Technologies*, AISC 206, pp. 33–42.
DOI: 10.1007/978-3-642-36981-0_4

classifications which specify the technological fields a patent may pertain to or be associated with. These fields are widely used in database searches to discover, for instance, the field or fields in which a firm may infringe another company's industrial rights or existing gaps in a certain technology which a company could exploit [6].

When analysts need information about a firm's innovations (to find niche markets, to internationally extend its innovations, etc.), they can retrieve this information from patent databases in standard formats, such as Comma-Separated Values (CSV) or eXtensible Markup Language (XML) [7]. Since these formats lack formal semantics to enable the interrelation of patent information, it is also therefore difficult to share data from different databases. In this regard, an efficient retrieval and processing of patent information based on semantics could improve the information and knowledge management of patents [8].

Such semantics could be provided by ontology languages with a formal model-theoretic grounding. Ontologies have been increasingly developed to represent, manage and reason about information system data. Ontologies allow common vocabularies and relationships between domain entities to be defined [9]. The web ontology language (OWL) [10] with its formal semantics [11] based on description logics has become the de facto standard among ontology languages.

In the context of patent metadata, ontologies have already been developed [1] [12]. These ontologies have been populated by translating XML documents retrieved from patent databases into OWL ontologies. However, these approaches seem to disregard the hierarchical relationships between the technological fields into which a patent may be classified, thereby hindering the exploitation of information that can be derived from the formal representation of patent code classification hierarchies by means of logic reasoning. Another drawback is that they only represent patent data from a single patent office.

This paper proposes an approach which is explicitly intended or utilized for creating and processing knowledge about patents. This approach provides a practical mechanism to automatically build and populate patent metadata ontologies by indexing hierarchical codes, which can be retrieved from different patent repositories, and by defining ontological categories which enrich patent information management with new relationships, properties and enabling the inference of new knowledge.

An application study is presented in order to illustrate the applicability of the proposal in the information and knowledge management about firms' innovation, by means of a case study, that shows how to automatically infer information about the internationalization of environmental patents on the basis of the information provided in the metadata of patent documents.

The rest of the paper is organized in the following way: Section 2 studies related work; Section 3 shows the proposed method for translating hierarchical patent codes from XML files into hierarchies of concepts in OWL files, including the population of the ontology; Section 4 presents a case study that shows the benefits of arranging the patent technological field in a hierarchical manner in OWL; and Section 5 concludes the paper by discussing the contributions of the proposal.

2 Related Work

In the domain of patent ontologies, various efforts have been made to create ontologies with information retrieved from patent databases. The most relevant patent ontologies based on patent metadata and represented in OWL are Patexpert, which was created within the European Patexpert project [13], [1], and PatentOntology, which was developed at Stanford University [12]. Although these two ontologies are relevant and close to our field of interest, none of them reflect the structure of the technological codes and they do not allow fully exploiting the logical reasoning with technological codes.

Patexpert was created to homogeneously represent different patent information from several EPO databases and to provide this with semantic meaning. However, Patexpert does not merge information retrieved from different patent offices: the patent metadata ontology has been populated by XSLT (eXtensible Stylesheets Transformation Language)[14], stylesheets. Unfortunately, the public version of this ontology is not populated, but to the best of our knowledge this ontology does not automatically retrieve or represent the semantics of the hierarchy of technological codes.

PatentOntology was developed to avoid the limitations of Patexpert when integrating heterogeneous domains [8]. PatentOntology merges information from USPTO patent documents retrieved from the USPTO database with information from patent courts of USPTO retrieved from the LexisNexis database [15]. This ontology has been populated with a parser, but does not automatically retrieve the semantics of the hierarchy of technological codes neither merge information from different offices.

In other domains, there exist other proposals which have extracted OWL documents from XML documents by using XSLT, but none of them has dealt with the extraction and then indexing of hierarchical codes in ontological categories as our proposal does. While some of this work uses XML schema [16] [17], other work only creates the OWL model [18] and others the OWL model and instances [19] but none attempt to represent the code structure in an ontological categorization.

Other work has also been published on translating XML into OWL and the development of visual tools such as JXML2OWLMapper [20], [21] or the online XMLtoOWL tool [22]. These tools enable the visual assignment of XML labels of the XML instances or schema into OWL labels. However, neither of these tools combines information retrieved from different sources nor capture the semantics of the hierarchical organization of codes.

3 Semantic Information Retrieval from XML to OWL

The codes for patent technological fields exhibit a certain structure that needs to be identified before a transformation document is defined to translate patent metadata in XML into an OWL ontology. Furthermore, each patent office defines its own scheme

to represent patent metadata. In this section, we introduce an overview of the method proposed and describe the transformation process from XML to OWL by means of XSLT. We then propose that the stylesheets be customized in order to fully exploit the information gathered in the hierarchical codes. Finally, we show an example of the method for a particular hierarchical code.

3.1 Method for Transforming XML into OWL with Stylesheets

This paper presents a method for processing query results from different databases in XML format, and the XML files are converted into OWL by means of the corresponding XSL files (styleshhets) and an XSLT processor. Figure 1 shows an overview of the method proposed.

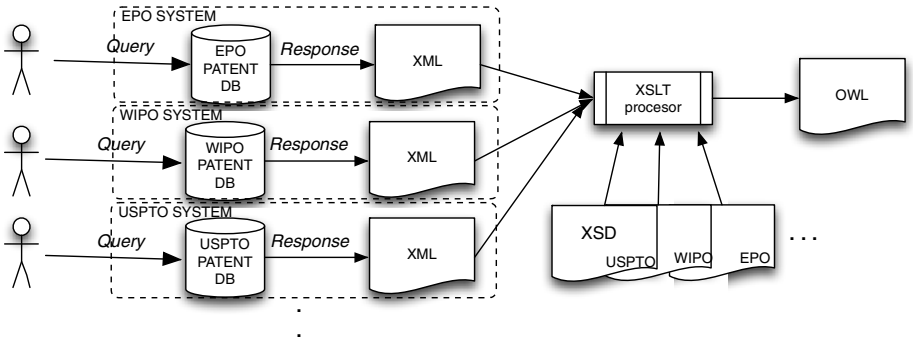

Fig. 1. Method Overview

The most important phase of this process is to create the stylesheet. Previous publications [12] [1] have used stylesheets to convert one node label of an XML document into a single class or property or instance of an OWL document.

In our proposal, we will automatically create an OWL model from the XML instance document for the technological fields of patents and OWL instances of this model. This method encompasses the translation of each instance of the technological code (a node label of the XML document) into a hierarchy of classes in the OWL document (several classes and one instance). For this purpose, it is necessary to divide the XML label into several parts and transform each into an OWL class or instance. The code is partitioned using XPath (XML Path language) [23].

The stylesheet developed in this work allows the ontology to be populated automatically with the hierarchical technological codes of the patents. Even when new codes appear, the stylesheet can create new classes and instances automatically without any further modification.

3.2 Structure of the Hierarchical Codes

Various patent technology codes exist such as the US classification or the European ECLA and ICO classifications or the International (IPC). The ICO and ECLA codes are based on the IPC codes and have the same underlying structure:

- Section, represented by one letter
- Class, represented by two digits
- Subclass, represented by one letter
- Main group, represented by one to three digits
- Group, represented by at least two digits

For instance, the ICO code Y02E10/20, which refers to energy generation, is made up of section (Y), class (02), subclass (E), main group (10) and group (20). As such, the complete code H04L25/02 means:

Y	new technological developments
Y02	technologies for mitigation of climate change
Y02E	reduction of greenhouse gases
Y02E10	energy generation for reduction of greenhouse gases
Y02E10/20	energy generation through renewable energy sources

Hierarchical codes have two main characteristics:

- Characteristic 1. Importance of the context of the code. The meaning of one part of the code depends on the meaning of the previous parts of the code. For example, the meaning of the part of the code "main group" (10) has different meanings if the previous part of the code is Y02E (Y02E10, energy generation for reduction of the greenhouse gases) or if the previous part of the code is Y02C (Y02C10, C02 capture or storage).
- Characteristic 2. Importance of parts of the code. The same representation of a part of the code (for example: 20) has different meanings depending on the part of the code to which it belongs. For example, the representation (20) has different meanings if it belongs to the "group" part of the code (Y02E10/20, energy generation through renewable energy sources) or to the "main group" part of the code (Y02E20, combustion technologies with mitigation potential).

Taking these two characteristics into account, the proposed method automatically creates an OWL ontology named HCOntology (Hierarchical Code Ontology) which represents the hierarchical codes with the full semantic meaning of the codes. The resulting HCOntology for the ICO codes with the tool Protégé [24] is shown in Figure 2. In OWL, since child classes inherit the meaning of the parent classes, HTCOntology therefore complies with the aforementioned Characteristic 1. Furthermore, HTCOntology represents different parts of the codes (for example the main group 10) with previous parts of the code (example ICO_Y02E10) and so HTCOntology fulfils the aforementioned Characteristic 2.

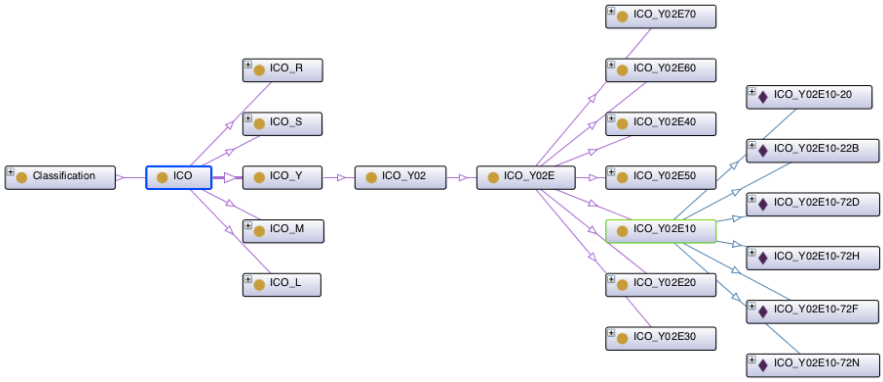

Fig. 2. HTCOntology focus on the ICO code Y02E10/20 in Protégé

3.3 Customization of the Stylesheets for Hierarchical Codes

In the method proposed, the customized stylesheets will allow the automatic translation of XML instance documents (where the instances are hierarchical codes) into an OWL ontology model and instances with the process shown in Figure 1. While other works translate one XML label (instance) into one single OWL label (class or instance), our work, on the other hand, proposes the translation of one XML label (instance) into several OWL labels (several classes and one instance) following an implicit hierarchy in the XML label.

The customization of the stylesheets proposed in this paper consists of the following steps:

1. Study the codes and define their parts, identifying the number of characters or digits in each part and whether they have separating characters
2. Study the structure of the labels in the XML and OWL files
   2.1. Location of the hierarchical codes in the XML file
   2.2. Detect the class from which implement HCOntology in the OWL file
3. Write the XSL file with its different parts:
   3.1. A header of the OWL file with the namespaces and if HTCOntology is implemented on top of an existing ontology, import the existing ontology
   3.2. Clean each label of the XML file, deleting unnecessary spaces, ensuring that each code is only written once in the OWL file, even if the code is repeated in several places in the XML file, etc.
   3.3. Create the hierarchical code structure
      3.3.1. Define the whole code except the last part as a subclass of the previous part of the code
      3.3.2. Repeat Step “3.3.1” for each part of the code until the first part of the code, and define the first part of the code as a subclass of the class of step “2.2”
      3.3.3. Insert the individuals (the whole code) in the corresponding created class
   3.4. Close the open labels

Table 1 shows an example for ICO codes. In which, we make use of the XPath functions related to substrings to split the hierarchical codes into its parts and create the corresponding ontology classes, subclasses and individuals for each code. This example shows the method steps to create an XSL file with the ICO codes (for clarity purposes we will show the instance of the ICO code in the XML file: Y02E10:20):

Table 1. Customizationof the stylesheets for ICO Codes

Step	Result
1	The structure of the ICO codes is shown in Section 3.2. The mark ":" separates the main group from the group in the XML file
2	Study the labels in XML and OWL file
2.1	<RESULT-LIST> <ROW> <ICO><p>
2.2	http://www.semanticweb.org/HCOntology#ICO
3	Write XSL file
3.1	The RDF namespace envelope and the ontology elements [25]
3.2	Clean the labels in the XML file that contains the hierarchical code
3.3	The code Y02E10:20 (figure 1) should have an OWL instance ICO_Y02E10-20, with the previous OWL hierarchy of classes (ICO_Y, ICO_Y02, ICO_Y02E, ICO_Y02E10). Figure 3 shows this step in the XSD file
3.3.1	In the XML instance Y02E10:20. Create the classes ICO_Y02E10 and (ICO_Y02E) and then define one as a subclass of the other (ICO_Y02E10 subclass of ICO_Y02E)
3.3.2	Create the class ICO_Y0E and define it as a subclass of ICO_Y02, create the class ICO_Y02 and define it as a subclass of ICO_Y, create the class ICO_Y and define it as a subclass of ICO
3.3.3	Add the instance ICO_Y02E10-20 to the class ICO_Y02E10
3.4	Close all the labels that remain opened

```
<xsl:variable name="var8" select="concat('ICO_',substring-before(normalize-space(.),':'))"/>
   <owl:Class rdf:about="{$var8}">
       <rdfs:subClassOf rdf:resource="{concat('ICO_',substring(normalize-space(.),1,4))}"/>
   </owl:Class>
   <owl:Class rdf:about="{concat('ICO_',substring(normalize-space(.),1,4))}">
        <rdfs:subClassOf rdf:resource="{concat('ICO_',substring(normalize-space(.),1,3))}"/>
   </owl:Class>

   <owl:Class rdf:about="{concat('ICO_',substring(normalize-space(.),1,3))}">
        <rdfs:subClassOf rdf:resource="{concat('ICO_',substring(normalize-space(.),1,1))}"/>
   </owl:Class>

   <owl:Class rdf:about="{concat('ICO_',substring(normalize-space(.),1,1))}">
        <rdfs:subClassOf rdf:resource=" http://www.semanticweb.org/HCOntology#ICO "/>
   </owl:Class>
   <owl:NamedIndividual rdf:about="{concat('ICO_',substring-before(normalize-space(.),':'),
                                           '-',substring-after(normalize-space(.),':'))}">
        <rdf:type rdf:resource="{concat('ICO_', substring-before(normalize-space(.),'/'))}"/>
   </owl:NamedIndividual>
```

Fig. 3. Excerpt of XSD file implementing step 3.3.1, 3.3.2 and 3.3

4 Reasoning through Ontology Hierarchy

A broader international scope of selected regions for the exploitation of patented environmental innovation provides existing patents a greater potential to influence

firms' financial performance [26]. Hence, our case study illustrates how the proposed method enables relationships to be created between concepts using the full semantics of the technological codes. In particular, it examines the patent portfolio of the German company and classifies its environmental patents (patents with an ICO code starting with Y02) as international patents if they have a family patent outside its region (Europe). In this example, the international patents of the family have been considered as the ones filed in the main patent offices outside Europe (the WIPO-WO- and the offices of China-CN-, USA-US-, Korea-KR- and Japan-JP-).

The XML file is a search result from the EPO database with the firms' patents. The OWL file is the result of applying the proposed methodology in Section 3 to the XML file. In order to exploit the reasoning over the resulting ontology OWL-DL expressivity is needed. Axiom 1 shows in description logic notation [11] the equivalent class created that classifies patents belonging to the class ICO_Y02 and that has one instance of a family patent with the representation WO, CN, KR, US or JP, in the class GlobalFamily. The representation of axiom 1 in Protégé is shown in the EquivalentClasses part of Figure 4.

The reasoner Pellet [27] has classified all the environmental patents with a family patent in one of the above patent offices as "globalPatent" (see the Menbers part of Figure 4). For example, Siemens patent PL356346A120040628, entitled "group of at least two windmills", with Y02E10/20 ICO code, and with family patents in WO and CN, among others, has been classified as GlobalPatent. Without HCOntology, these new relationships should not be possible and so HCOntology therefore enriches the information retrieval and management process and thus improves patent analysis.

$$
\begin{aligned}
&GlobalFamily \\
&\equiv PatentDocument \sqcap \exists hasICOCode.ICO\_Y02 \\
&\sqcap \big((\exists hasFamilyStates.\{WO\} \\
&\sqcup \exists hasFamilyStates.\{CN\} \\
&\sqcup \exists hasFamilyStates.\{US\} \\
&\sqcup \exists hasFamilyStates.\{KR\} \\
&\sqcup \exists hasFamilyStates.\{JP\})\big)
\end{aligned}
$$

Axiom 1: Definition of the equivalent class of GlobalFamily

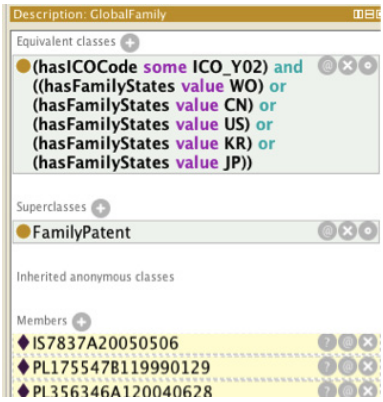

Fig. 4. Equivalent class GlobalFamily in Protégé

5 Conclusions

Information about patents is stored in large patent document databases with heterogeneous data representations. The technological field, although defined as a

single register in the databases, has ample implicit information about the classification of the technological field to which the patent belongs due to the hierarchical nature of the technological codes. However, this information is stored without any semantics and the information is not therefore explicitly represented to be automatically processed by computers. Patent documents can be retrieved from the databases in several standard formats, such as CSV or XML. Because of the limited semantics of such formats, additional relationships cannot be inferred between patents for analysis purposes, and makes hampers the sharing of heterogeneous data from different databases. Ontology languages such as OWL provide this semantics and have proved useful for representing and managing knowledge. Various efforts have been made to create patent ontologies in OWL, enabling new knowledge to be discovered and patent analysis improved, and these could be used by firms to optimize innovation management. Such ontologies have been populated by transforming XML instances into OWL instances. These translations are, however, limited to mirror hierarchies of patent codes and do not take further advantage of reasoning capabilities, and have not dealt with the heterogeneity of patent representation by different patent offices.

This article introduces an approach for automatically retrieving information from different patent repositories and for indexing hierarchical codes with ontological categories. This indexing enriches the information retrieval and management process with new relationships, properties and inferred information. The paper also discusses the importance and potential of this indexing.

More specifically, this paper provides a method for automatically translating XML instances from different data repositories into OWL classes and instances. This method is based on XSLT, and the XSD file is built with the help of XPath that splits the code into its structural parts. This methodology enables the future emerging codes to be translated automatically without the need for any reimplementation. The resulting hierarchy of concepts in OWL (which we called HCOntology) allows the exploitation of the information gathered in each part of the hierarchy.

We have also shown the potential of the proposal through a case study that creates relationships between concepts using the full semantics of the technological codes. In particular, we have classified as international environmental patents those patents that have family patents outside the region of the owner firm and are environmental patents. The detailed empirical comparison between our proposal and other methods is out of the scope of this paper because space restrictions, however we are implementing this analysis in a future paper that it is now under preparation.

Although this method has been applied in the domain of patents, it could equally be applied to any domain with hierarchical codes. The hierarchical codes would therefore be enriched with semantics, enabling the definition of new relationships, properties and inferred information.

References

1. Giereth, M., St, A., Rotard, M., Ertl, T.: Application of Semantic Technologies for Representing Patent Metadata. In: 1st International Workshop on Applications of Semantic Technologies AST 2006, vol. 94, pp. 297–305 (2006)
2. Almeida, P., Phene, A.: Subsidiaries and knowledge creation: the influence of the MNC and host country on innovation. Strategic Management Journal 25, 847–864 (2004)

3. Vries, D.: Leveraging Patents Financially. Gabler Verlag, Wiesbaden (2012)
4. Oltra, V., Kemp, R., Vries, F.: de: Patents as a measure of eco-innovation. International Journal of Environmental Technology and Management 13, 130–148 (2010)
5. Trappey, C.V., Wu, H.-Y., Taghaboni-Dutta, F., Trappey, A.J.C.: Using patent data for technology forecasting: China RFID patent analysis. Advanced Engineering Informatics 25, 53–64 (2011)
6. Foglia, P.: Patentability search strategies and the reformed IPC: A patent office perspective. World Patent Information 29, 33–53 (2007)
7. W3C: Extensible Markup Language (XML), 2nd edn. (2006)
8. Taduri, S., Lau, G.T., Law, K.H., Yu, H., Kesan, J.P.: Developing an Ontology for the U. S. Patent System. Text (2011)
9. Guarino, N., Giaretta, P.: Ontologies and Knowledge Bases: Towards a Terminological Clarification. In: Mars, N. (ed.) Towards Very Large Knowledge Bases Knowledge Building and Knowledge Sharing, pp. 25–32. IOS Press (1995)
10. W3C: OWL - Ontology Web Language (2009)
11. Baader, F.: The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press (2003)
12. Taduri, S., Lau, G.T., Law, K.H., Yu, H., Kesan, J.P.: An Ontology to Integrate Multiple Information Domains in the Patent System. In: 2011 IEEE International Symposium on Technology and Society ISTAS, pp. 23–25 (2011)
13. Wanner, L., Brügmann, S., Diallo, B., Giereth, M., Kompatsiaris, Y., Pianta, E., Rao, G., Schoester, P., Zervaki, V.: PATExpert/: Semantic Processing of Patent Documentation. Knowledge Creation Diffusion Utilization (2009)
14. W3C: XSL Transformations (XSLT), `http://www.w3.org/TR/xslt`
15. LexisNexis Website
16. Ferdinand, M., Zirpins, C., Trastour, D.: Lifting XML Schema to OWL. In: Koch, N., Fraternali, P., Wirsing, M. (eds.) ICWE 2004. LNCS, vol. 3140, pp. 354–358. Springer, Heidelberg (2004)
17. Bedini, I., Gardarin, G., Nguyen, B.: Deriving Ontologies from XML Schema. In: Proceedings EDA 2008 (2008)
18. Bosch, T., Mathiak, B.: XSLT transformation generating OWL ontologies automatically based on XML Schemas. IEEE (2011)
19. Lacoste, D., Sawant, K.P., Roy, S.: An efficient XML to OWL converter. In: Proceedings of the 4th India Software Engineering Conference 2011 ISEC 2011, pp. 145–154 (2011)
20. Rodrigues, T., Rosa, P., Cardoso, J.: Moving from syntactic to semantic organizations using JXML2OWL. Computers in Industry 59, 808–819 (2008)
21. Cardoso, J., Bussler, C.: Mapping between heterogeneous XML and OWL transaction representations in B2B integration. Data & Knowledge Engineering 70, 1046–1069 (2011)
22. Bohring, H., Auer, S.: Mapping XML to OWL Ontologies. Leipziger InformatikTage 72, 147–156 (2005)
23. Malhotra, A., Melton, J., Walsh, N.: XQuery 1.0 and XPath 2.0 Functions and Operators (2007)
24. Noy, N.F., Fergerson, R.W., Musen, M.A.: The knowledge model of Protégé-2000/: combining interoperability and flexibility. Knowledge Engineering and Knowledge Management Methods Models and Tools 1937, 1–20 (2000)
25. Smith, M.K., Welty, C., McGuinness, D.L.: OWL Web Ontology Language Guide (2004)
26. Kirca, A.H., Hult, G.T.M., Mena, J.A., Miller, J.C.: Firm-Specific Assets, Multinationality, and Financial Performance: A Meta-Analytic Review and Theoretical Integration. Academy of Management Journal 54, 47–72 (2011)
27. Sirin, E., Parsia, B., Grau, B., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. Web Semantics Science Services and Agents on the World Wide Web 5, 51–53 (2007)

Multilevel Clustering of Induction Rules for Web Meta-knowledge

Amine Chemchem, Habiba Drias, and Youcef Djenouri

USTHB, LRIA
BP 32 El Alia Bab Ezzouar, Algiers, Algeria
aminechemchem@gmail.com, hdrias@usthb.dz,
youcef062009@hotmail.fr

Abstract. The current World Wide Web is featured by a huge mass of knowledge, making it difficult to exploit. One possible way to cope with this issue is to proceed to knowledge mining in a way that we could control its volume and hence make it manageable. This paper explores meta-knowledge discovery and in particular focuses on clustering induction rules for large knowledge sets. Such knowledge representation is considered for its expressive power and hence its wide use. Adapted data mining is proposed to extract meta-knowledge taking into account the knowledge representation which is more complex than simple data. Besides, a new clustering approach based on multilevel paradigm and called multilevel clustering is developed for the purpose of treating large scale knowledge sets. The approach invokes the k-means algorithm to cluster induction rules using new designed similarity measures. The developed algorithms have been implemented on four public benchmarks to test the effectiveness of the multilevel clustering approach. The numerical results have been compared to those of the simple k-means algorithm. As foreseeable, the multilevel clustering outperforms clearly the basic k-means on both the execution time and success rate that remains constant to 100 % while increasing the number of induction rules.

Keywords: Knowledge mining, meta-knowledge, multilevel paradigm, k-means, k-nearest neighbors, induction rules, genetic algorithm.

1 Introduction

To accelerate knowledge discovery [1][2], it would be interesting to exploit the knowledge currently present on the web and proceed to its mining. The purpose of this paper is to propose a knowledge mining process and to show how to adapt some data mining tasks to knowledge. Let us point out that the concept of knowledge mining is different from the one we found in the literature [3][4][14][15]. To be clear, we are interested in this work in mining knowledge instead of elementary data and the result of the desired task is therefore meta-knowledge.

The only similar study we found in the literature deals with frequent sequential patterns [5]. In fact in this paper, the authors focus on clustering sets of items and not

Á. Rocha et al. (Eds.): *Advances in Information Systems and Technologies*, AISC 206, pp. 43–54.
DOI: 10.1007/978-3-642-36981-0_5

on simple items and more precisely sequential patterns. In our case we are interested in induction rules because of their closeness to the natural language. On the other hand, clustering is considered as a mining task for scalability concern. A new clustering approach based on multilevel paradigm is proposed.

The idea behind the multi-level approach is to be able to tackle very large scale knowledge sets. The reminder of the paper is organized as follows. The next section presents the concept of knowledge mining compared to data mining. Afterwards, induction rules representation and similarity measures are proposed according to the morphological aspect. Then a new clustering approach based on multilevel paradigm for induction rules is proposed in section3. The experimental evaluation presented in section 4 is performed on a benchmark including three different knowledge bases and another one containing hard SAT instances. Conclusions are finally summarized and some perspectives are suggested.

2 Knowledge Mining

In the literature [14] the knowledge mining concept is defined as an evaluation process of knowledge discovery, or as a selection process of interesting knowledge [15]. We can also cite many other works [3] [4] that define the concept of knowledge mining as a process of extracting knowledge from a tremendous database using prior knowledge. In other words, we illustrate the concept by the following scheme:

Prior Knowledge + data mining +goal → desired knowledge.

This definition deals with mining data not knowledge, it is different from the paradigm that we aim at introducing and studying.

We define knowledge mining as the process of extracting new knowledge from a knowledge set, such as a knowledge base or a knowledge warehouse. It can be then described as follows:

Knowledge + Prior Knowledge + Goal → Meta-Knowledge

2.1 Knowledge Representation

Knowledge representation consists in translating the natural knowledge into a symbolic formalism that can be processed by a machine. Knowledge is too vast and diverse to be represented and operated by a single formalism. Among the knowledge representations, the procedural and the declarative approaches have been investigated for a long time [6]. The procedural approach, invokes the simplicity and the ease of understanding reasoning, represented by algorithms simulating real behaviors. In addition, the procedural representation allows treating problems with algorithmic style which is fully analyzable and fully understandable. On the contrary, the declarative approach is more flexible because it provides heuristic expressions using statements. It allows to specify constraints and to learn independently methods of use. Fig. 1 shows the various knowledge representation formalisms from procedural which

is rigid and well structured to the declarative one which is on the contrary more open and free.

Clearly, induction rules are the closest to the natural language phrases. In addition, they represent an efficient framework to design reasoning systems and by producing meta-knowledge.

Another subsequent extension of induction rules mining is the design of super intelligent agents that are capable to reason on subsets of rules rather than on single rules. Lot of researches can be launched from this idea, such as knowledge represented by taxonomies or association rules.

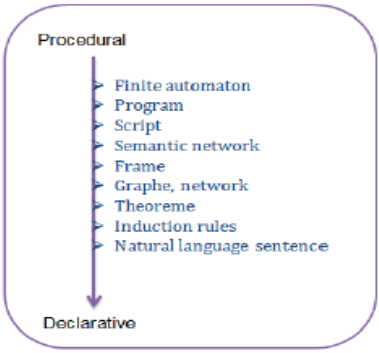

Fig. 1. Knowledge Representation

2.2 Induction Rules

An induction rule is a Boolean formula of the form: $R : X \rightarrow Y$, where X and Y are sets of clauses. X is called the premise of the rule and Y its consequence[3].

2.3 A Distance Measure for Knowledge Rules

In the previous work [7], a similarity measure between induction rules was proposed. It took into account the main items of the rule except the logical connectors in the design of the distance between two rules. This is a subtle way to evaluate the differences between rules. However, one important element was not considered at all, which is the position of the implication operator because this element distinguishes between the premise and consequent parts of the rule. In the present paper, we integrate this feature in designing similarity measures in order to be more precise and hopefully yield more accurate results. In addition to that fact, the clauses are taken as basic entities for measuring the disparity between rules. Moreover, if we consider the pair (variable, value) when ignoring the relational operator instead of the whole clause we obtain another measure with average bounded rationality. For a completely bounded rationality, we can take in account only the variables of the clauses to obtain a third measure. These three measures are summarized as follows:

Distance with Complete Rationality

$$Dist(C_i;C_j) = total(C_i;C_j) - shared\ (C_i;C_j) \quad (1)$$

$$Dist1(R_i;R_j) = Dist(antecedent(R_i),antecedent(R_j)) + Dist(consequent(R_i),consequent(R_j)) \quad (2)$$

where C_i and C_j are respectively two sets of clauses and the distance between them is equal to the number of clauses that are totally different from one clause to the other.

Antecedent(R) is the set of clauses appearing in the premise part of the rule R and consequent(R) is the set of clauses appearing in the consequence part of the rule R.

Distance with Average Bounded Rationality
In this case, the same formula is used except for $Dist(C_i;C_j)$ where we ignore the relational operators of the clauses.

Distance with Full Bounded Rationality
In this case, the same formula is used except for $Dist(C_i;C_j)$ where only the variables are considered. Of course we will obtain three kinds of clustering depending on which distance we select.

Let look at the following rules:

R1: If (temperature = hot) and (humidity = low) then (outlook=sunny),
R2: If (outlook=sunny) and (temperature = hot) and (wind = light) then (play_tennis=no),
dist(R1,R2)= total(R1,R2) - shared(R1,R2) = 7 - 4 = 3. computed by [7]
dist(R1,R2)= dist(antecedent (R1),antecedent(R2)) + dist(consequent(R1), consequent(R2)) = (5-1)+(2-0) = 6, computed by the complete rationality distance.

The defined distance is a reliable and valid metric measure across the whole of the induction rules, because we have demonstrated mathematically the four properties of a metric distance function, which are:

1- D is a function which is defined as follows:
$D : E\times E \rightarrow R$ // E is the whole of induction rules: $(X, Y) \rightarrow D(X, Y)$
2- $\forall x \in E : D(x, x) = 0$.
3- $\forall (x, y) \in E^2 : D(x, y) = D(y, x)$.
4- $\forall (x, y, z) \in E^3 : D(x, y) = D(x, z) + D(z, y)$.

2.4 Centroids Computation

Another concept, which is necessary to perform clustering using k-means is the centroid. The latter, which is the central element of the cluster represents somehow all the objects of the whole cluster. Finding formula that computes exact centroid of a set of rules is not evident. One possible idea is to calculate the distance that separates each pair of rules, then for each rule associate the sum of the distances that separates it from the others. The centroids will correspond to the rule that has the near-average sum of distances because intuitively, it is the closest to all the other rules and hence the most similar to them.

3 Multilevel Induction Rules Clustering

3.1 Related Works of Multilevel Paradigm

In the literature, the multilevel paradigm was first proposed by [8], as a method of speeding up spectral bisection, and improved by generalizing it to encompass local refinement algorithms [9]. It has been made popular by [10], and since then

it was commonly used for graph partitioning problems, as the work [11]. Also the authors in [12] hybridize a multilevel approach with an ant-colony meta-heuristic for mesh-partitioning. And recently the multi-level approach was used for solving problems such as combining memetic algorithms with multi-level approach for solving the problem SAT[13]. The multilevel paradigm is a simple technique which at its core involves recursive coarsening to produce smaller and smaller problems that are easier to solve than the original one. This paradigm consists of three phases: coarsening, initial solution, and uncoarsening. The coarsening phase aims at merging the components associated with the problem to form clusters. The clusters are used in a recursive manner to construct a hierarchy of problems. Each level of this hierarchy represents the original problem but with fewer degrees of freedom. The coarsest level can then be used to compute an initial solution.

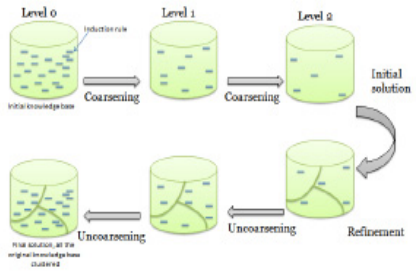

Fig. 2. Multilevel Induction Rules Clustering

This last, found at the coarsest level is extended to give another initial solution for the next level, and then improved using a chosen optimization algorithm.

3.2 The Proposed Algorithm

In this section, we develop a new general algorithm for induction rules clustering based on multilevel paradigm. First, it begins by eliminating the nearest induction rules from the original knowledge set in the coarsening step to keep only the most far away rules from each other. Thanks to a genetic algorithm that computes at the end of this stage a mini knowledge base that includes the different most distant induction rules as explained in fig.2. After this step, the k-means algorithm is launched on the small base and it is expectable that it will provide reliable centroids, because the selected rules are very distant from each other. Finally in the refinement step, the k-NN algorithm adapted to induction rules is invoked on the result of k-means of the previous step. This operation is repeated until obtaining the initial knowledge set. The purpose of using k-NN is to complete inserting the remaining rules in the corresponding classes.

3.3 Coarsening Step

The purpose of this step is to reduce the induction rules number. We use a genetic algorithm for this aim while ensuring a good sweeping of all the knowledge base. The rules are not removed randomly but in a way to keep only the rules that appear very far from each other.

A Chromosome (Individual)
It is an encoded solution in our case represented by an array of integers varying between 0 and the number of induction rules of the knowledge set, each gene representing a rule. Therefore the length of the chromosome represents the size of the small knowledge base. For instance, if we have the following chromosome (19,201,1987,2012,3033,4200), its size is equal to 6, the number of rules, which can be located as follows: The first rule is at position 19 in the set, the second one at position 201 and so on...

Fitness Function
The notion of fitness is fundamental to the application of genetic algorithms. It is a numerical value that expresses the performance of an individual (solution), so that different individuals can be compared. We propose as a fitness function the sum of the distances between each of the genes of the individual, that is:
$\text{fitness}(\text{chromosome}) = MAX\ (\sum_{i=1}^{n-1} \sum_{j=1}^{n} Dist(i,j))$, for all i different from j with i and j belonging to chromosome.

Coarsening Algorithm

```
Genetic Algorithm for Coarsening Step
Begin
   Generate randomly an initial population of solutions;
   Compute fitness for each solution of the population;
    For ( i:= 1 to maxIter)do
      Choose 2 individuals (or solutions);
        Generat Rc randomly RC, RM : a random numbers between [0 and
        length of chromosome];
          If (Rc < rate of crossover) then
               Apply the crossover;
          end_If
          If (Rm < rate of mutation) then
               Apply the mutation on one chromosome selected
                  randomly;
          end_If
     Evaluate fitness(S')
               /*where S' is the new individual*/
          If (fitness (S') > fitness (S)) then
               Remove S from the population and insert S';
          end_if
    end_for
   Consider the best found solution;
end
```

3.4 The Initial Solution

After the coarsening step, we obtain a solution that maximizes the distance between its induction rules. So it is sour that the last one sweeps all the areas of the knowledge set. In this step, we apply a simple k-means algorithm to cluster these rules as explained in Fig.3 and the reliable gravity centers will be achieved. The result of this clustering constitutes the initial solution.

3.5 The Refinement Step

After the clustering in the initial step, the different clusters are obtained with the rules belonging to the initial solution. In order to rebuild our initial knowledge set, we suggest applying the k-nearest-neighbors algorithm for the rest of the induction rules, as shown in the following algorithm.

Refinement Step Algorithm based on k-NN

```
Begin
rule_list := all rules (knowledge base) - rules (initial
 solution);
   while (rule_list is not empty) then
      Current_rule  :=  get_rule (rule_list);
         for  (each  already  classified  rule  Ri)  do
           Calculate  the  distance  Dist_Clauses  (curr-
           nt-rule,Ri);
         end_for
      Compute  k_nearest_neighbor(current-rule);
          for  (each  rule  belonging  to  k-NN)  do
           Calculate  the  number  of  frequency  of  each
           class;
         end_for
      Attribute  to  "current_rule"  the  most  frequent
      class;
   end_while
End.
```

4 Evaluation and Experimentation

One of the hardest problems in comparing different clustering algorithms is finding an algorithm to evaluate the quality of the clusters. Our main idea is to construct the induction rules set from three different benchmarks, and with fixing the parameter “k” of the clustering algorithms at 3, we can calculate the clustering success rate as explained in fig.3. And in a second step and only after having validated the clustering approach we develop it on a public SAT benchmark.

4.1 Benchmark Construction

Data Benchmark Adaptation
In this part, we build the knowledge set from three public different benchmarks. It includes 11000 induction rules, after transforming the data sets to induction rule sets, as follows:

If ($attribute_1 = value_1$) and ($attribute_2 = value_2$) and ... then ($attribute_n = value_n$)

Where n is the last attribute of the data set.

The first one is originally a big data set known as **Chess (King-Rook vs.King) Data Set**[1], which contains 28056 instances with 7 attributes.

The second benchmark is known as **Abalone Data Set**[2]. It contains 4177 instances with 9 attributes. The third data set is known as **Car Evaluation Data Set**[3]. It contains 1728 instances, with 7 attributes.

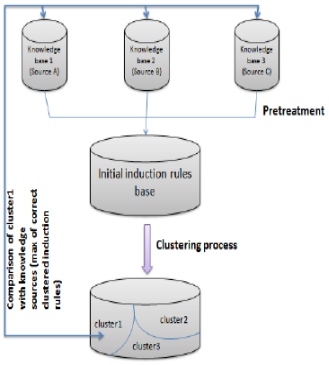

Fig. 3. Clustering Evaluation

SAT Benchmark Adaptation

We used the public SAT benchmark known as SATLIB[4], it contains 1000 SAT instances with 20 variables and 91 clauses. A SAT Benchmark is the set clause-conjunctions of literals. Each instance SAT can be modified as follows:

Let $C_i = \neg\alpha 1 \vee \alpha 2 \vee \neg\alpha 3$ be a clause, then

As $a \Rightarrow b \equiv \neg\alpha 1 \vee \beta$ then: $\neg\alpha 1 \vee \alpha 2 \vee \neg\alpha 3 \Rightarrow (\alpha 2 \vee \neg\alpha 3)$...(3)

Using Eq.(3), we can consider the following rule: if $\alpha 1 = 1$ then $\alpha 2 = 1$ or $\alpha 3 = 0$. Generally, each clause C is transformed as Eq.(3).

After that, for each variable,$\alpha_i \in C$, if α_i is true, then it set to 1, otherwise to 0.

4.2 Evaluation Pattern

Fig.3 shows the evaluation process of the clustering, when fixing the parameter k of the clustering algorithms to 3. Then after the clustering step, the three obtained clusters are compared to the initial rule bases. If the clustering process is efficient, the respective clusters should be identical. The success rate is calculated using the following formula:

$$\text{success rate} = \frac{ncr}{npr} = \frac{ncr1}{npr1} + \frac{ncr2}{npr2} + \frac{ncr3}{npr3} \quad \ldots \qquad (2)$$

Where, ncri=number of correct rules for knowledge base i.
npri=number of pertinent rules from knowledge base i (total number of its rules).

The number of correct rules of KB_i = *max*(number rules common (KB_i;C_j)) for all j in [1,3].

[1] http://archive.ics.uci.edu/ml/datasets/Chess+%28King-Rook+vs.+King%29
[2] http://archive.ics.uci.edu/ml/datasets/Abalone
[3] http://archive.ics.uci.edu/ml/datasets/Car+Evaluation
[4] http://www.satlib.org/

4.3 Experimentation

Performance Comparison

Fig.4 compares the performance of several variants of the designed clustering approaches. We remark that the results computed by the simple k-means and the multi-level k-means are very different. While the success rate of the simple k-means algorithms is reduced when increasing the number of rules to cluster, the multilevel algorithms success rate remain constant to 100%.

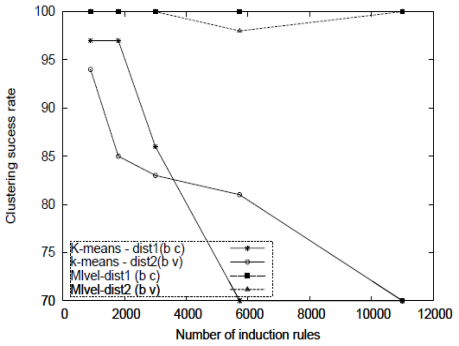

Fig. 4. Success rate of clustering approaches

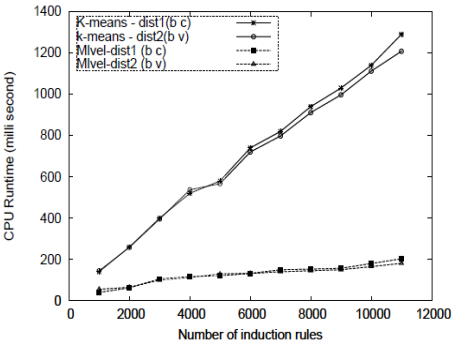

Fig. 5. CPU Time of clustering approaches

CPU Runtime Comparison

The execution times for the different clustering approaches are compared in Figure 5, when increasing the number of induction rules. We observe that the multilevel clustering algorithms are faster than the simple clustering. Moreover, when the simple clustering approaches cluster a number of induction rules varying between 100 and 11000, the execution time increases from 180 to 1270 milliseconds whereas the execution time of multilevel approaches increases just between 26 and 200 millisecond.

Application of Clustering Approaches on SAT Benchmark

Fig.6 shows how the execution time augments with the increase of number of sat instances for the four clustering approaches. According to this figure, we remark that the simple clustering approaches are slower than the two approaches based on multilevel paradigm, like in the previous subsection2. This is due to the fact that the simple clustering approaches repeat a computation of new centroids, until the stability of all induction rules in clusters, however, in the multilevel clustering approaches the gravity centers are computed only once. Execution time of the multilevel approaches do not exceed 35milliseconds, when the number of sat instances is 1000 rules, even though the simple clustering approaches execution time is 200 ms when the number rules is 1000.

In figure 7 the results of comparing the clustering approaches while increasing the parameter 'k' are shown. When the number of SAT instances is fixed at 1000 rules,

and with the increase of the number of clusters from 3 to 40 clusters, we remark that the CPU runtime of the simple clustering approaches increase between 125 to 290 milliseconds. However the CPU runtime of the clustering approaches based on multilevel paradigm increases just from 40ms to 90 milliseconds.

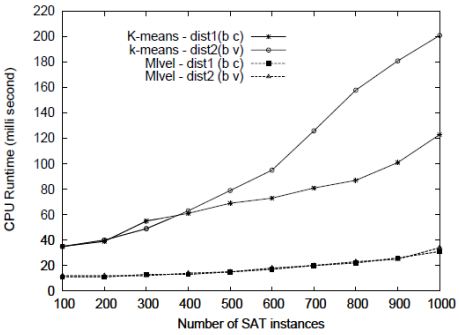

Fig. 6. CPU Runtime of clustering approaches for SAT instances

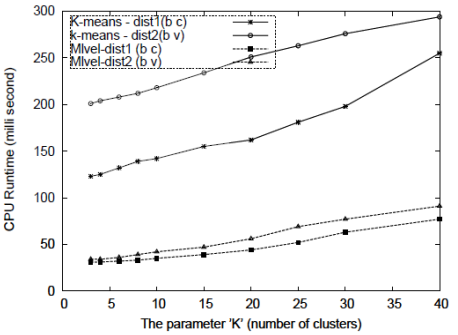

Fig. 7. CPU Time of clustering SAT instances Vs number of clusters

General Comparison

When comparing the four clustering approaches, we note that the approaches based on multilevel clustering are more efficient on both solution quality and runtime criteria than the simple clustering approaches, so they take until 83% less time than the simple clustering approaches to treat the same rule base.

Furthermore, even on the quality criterion expressed through the success rate, their performance reaches 100% whatever the number of rules whereas the simple clustering approaches success rate varies between 97% and 70%. Finally with applying the four approaches on SAT benchmark, and with increasing the number of clusters, the difference between the runtime of the approaches is very important. Therefore when increasing the number of clusters from 3 to 40 clusters, the runtime for the simple clustering approaches increase from 120 to 290 milliseconds, while the runtime of the multilevel clustering approaches remain constant between 40 and 90 milliseconds.

5 Conclusion

In this paper we presented a new multilevel k-means algorithm for clustering induction rules. The algorithm combines a traditional clustering technique in the initial solution step, a genetic algorithm in the coarsening step and a supervised classification especially the k-nearest-neighbors algorithm in the refinement step. As our experimental results proved, the best clustering solutions were produced by the multilevel clustering approaches when comparing to the simple k-means. Furthermore, the algorithm has the additional advantage of being extremely fast, as

it is based on genetic algorithm. For instance, the amount of time required by the proposed algorithm ranges from 26 milliseconds for a knowledge base with 900 induction rules to 115 milliseconds for a knowledge base with 11000 induction rules, on a I5 core PC. We believe that this paper presents the first attempt for developing a robust framework for a large scale clustering approaches. However, a number of key questions remain to be addressed, in particular the best way to design the different components of the multi-level paradigm. We can imagine in a future work, an application of the multilevel clustering of induction rules on a rule base of an agent intelligent, in order to speed up its reasoning and also to discover the new meta-knowledge.

References

1. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Elsevier (2011)
2. Mariscal, G., Marbn, Fernndez, C.: A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review 25, 137–166 (2010)
3. Kaufman, K.A., Michalski, R.S.: From Data Mining to Knowledge Mining. In: Rao, C.R., Solka, J.L., Wegman, E.J. (eds.) Handbook in Statistics. Data Mining and Data Visualization, vol. 24, pp. 47–75. Elsevier/North Holland (2005)
4. Michalski, R.S.: Knowledge mining: A proposed new direction, School of Computational Sciences George Mason University and Institute for Computer Science Polish Academy of Sciences (2003)
5. Saneifar, H., Bringay, S., Laurent, A.: S2MP: Similarity Measure for Sequential Patterns. In: Proceeding of the 7th Australian Data Mining Conference AusDM 2008, Adelaide, Australia, November 27-28, pp. 95–104 (2008)
6. Tuomi, I.: Data is More Than Knowledge Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory. Journal of Management Information Systems 16(3), 107–121 (fall 1999)
7. Drias, H., Aouichat, A., Boutorh, A.: Towards Incremental Knowledge Warehousing and Mining. In: DCAI 2012, pp. 501–510 (2012)
8. Barnard, S.T., Simon, H.D.: A fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems. Concurrency: Practice and Experience 6, 101–117 (1994)
9. Hendrickson, B., Leland, R.: A multilevel algorithm for partitioning graphs. In: Proceedings of the Supercomputing 1995 (1995)
10. Karypis, G., Aggarwal, R., Kumar, V.: A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Sci. Comput. 20(1), 359–392 (1999)
11. Dhillon, S., Guan, Y., Kulis, B.: Weighted Graph Cuts without Eigenvectors: A Multilevel Approach. IEEE Transactions on Patterns Analysis and Machine Intelligence 29(11) (2007)
12. Korosec, P., Silc, J., Robic, B.: A Multi-level Ant-Colony-Optimization: Algorithm for MESH Partitioning, Computer Systems Department, Jozef Stefan Institute, Ljubljana, Slovenia. IEEE 2003 Conference Publication (2003)

13. Bouhmala, N.: A Multilevel Approach Applied to Sat-Encoded Problems, Vestfold University College Norway. VLSI Design (2012) ISBN: 978-953-307-884-7
14. Poongothai, K., Sathiyabama, S.: Integration of Clustering and Rule Induction Mining Framework for Evaluation of Web Usage Knowledge Discovery System. Journal of Applied Sciences 12, 1495–1500 (2012)
15. Poongothai, K., Sathiyabama, S.: Efficient Web Usage Miner Using Decisive Induction Rules. Journal of Computer Science 8(6), 835–840 (2012)

Knowledge-Based Risk Management: Survey on Brazilian Software Development Enterprises

Sandra Miranda Neves[1,3], Carlos Eduardo Sanches da Silva[2], Valério Antonio Pamplona Salomon[3], and André Leonardo Almeida Santos[2]

[1] UNIFEI, Federal University of Itajuba, Irmã Ivone Drummond Street, 200 Industrial District, 35903-087, Itabira, Minas Gerais State, Brazil
sandraneves@unifei.edu.br
[2] UNIFEI, Federal University of Itajuba, BPS Ave. 1303, Pinheirinho, 37500-903, Itajuba, Minas Gerais State, Brazil
sanches@unifei.edu.br, andre.leonardo.cco@gmail.com
[3] UNESP, Sao Paulo State University, Ariberto Pereira da Cunha Ave. 333, 12516-410, Guaratingueta, Sao Paulo State, Brazil
salomon@feg.unesp.br

Abstract. A risks management, carried on in an effective way, leads the software development to success and may influence on the organization. The knowledge takes part of such a process as a way to help taking decisions. This research aimed to analyze the use of Knowledge Management techniques to the Risk Management in software projects development and the possible influence on the enterprise revenue. It had, as its main studying subject, Brazilian incubated and graduated software developing enterprises. The chosen research method was the Survey type. Multivariate statistical methods were used for the treatment and analysis of the obtained results, this way identifying the most significant factors, that is, enterprise´s achievement constraining factors and those outcome achievement ones. Among the latter we highlight the knowledge methodology, the time of existence of the enterprise, the amount of employees and the knowledge externalization. The results encourage contributing actions to the increasing of financial revenue.

Keywords: Software development, Knowledge Management, Risks Management, Incubated and Graduated Technological Based Enterprises.

1 Introduction

All projects have risks. Once a project management aims to increase the success rate of projects, it is, in its essence, the risk management [1]. In this context, the software projects are part of them and, especially susceptible to faults [2]. The high failure rates, associated with the information system projects, suggest that organizations need

Á. Rocha et al. (Eds.): *Advances in Information Systems and Technologies*, AISC 206, pp. 55–65.
DOI: 10.1007/978-3-642-36981-0_6

to improve not only their ability to identify, but also manage the risks linked to those projects [3].

An organization cannot effectively manage the risks provided it doesn´t manage its knowledge [4]. However, it is not enough to have the knowledge in some part of the organization; it needs to be accessible to be used and to serve as a basis to new knowledge to be created [5]. Farias et al. [6] consider that projects managers may make mistakes from the past simply for not knowing the mitigation actions that have been successfully applied, once an applied ineffective risk Knowledge Management contributes to maximize such a problem. A research, carried on by Wong e Aspinwall [7], from a survey with 72 small and medium information technological enterprises, identified that only 26 enterprises used Knowledge Management formally. The main reason for not using it was due to the uncertainty as to the potential benefits (45.7%). For Alhawari et al. [8], to use Knowledge Management processes to improve the applications of processes, linked to Risk Management, is a recent and important research area. And yet, this research area is not so addressed. According to Massingham [9] the Knowledge Management and Risk Management relationship is an academic research emergent field. So, this research has as its aim to analyze the use of Knowledge Management (KM) techniques to the Risk Management (RM) in software projects development and the possible influence on the enterprise revenue.

The article is structured as follows. Section 1 presents justifications and the research aim. Section 2 discusses the theoretical foundation on Technological Incubated and Graduated Based Enterprises and the relationship between RM and KM. Section 3 contemplates the research method and Section 4 the resulting presentation and analysis. Finally, Section 5 presents conclusions and suggestions for future research.

2 Literature Review

2.1 Incubators and Technological Incubated and Graduated Based Enterprises

Enterprises incubators have become a phenomenon in many parts of the world and are seen as a tool to promote the growing of technological based enterprises development [10]. Dahlstrand [11] defines a Technological Based Enterprise (TBE) as that one depending on technology for its development, not meaning necessarily, in most of the cases, it has to be new or in innovation. The enterprises incubators supply technical support, networking capacity, infrastructure, shared services and facilitates access to capital, making it vital to business development in their early stages [12].

At Graduated Enterprises the information management and its procedures have been identified as being more aware and structured. In addition, incubated enterprises presented an organization environment with better conditions for knowledge creation [13].

2.2 Risk Management and the Relationship with Knowledge Management

Risks in software projects are a series of factors or conditions that may represent a serious threat for the success of the project achievement [14] and they imply to quantify the importance of a risk, assessing and its possible impact on the project, as well as in the strategies development to control it [15]. Despite the improvements already achieved, many software development projects still use more sources than planned, take longer to be finished and supply less quality and functionality than expected [16].

For Davenport e Prusak [5], KM is composed by set of processes that seek to support in the organizational environment the knowledge generation, their register and their transfer. Anantatmula e Kanungo [17] state that "knowledge is recognized as a critical resource to acquire and keep up competitive advantage in business". So, several enterprises have expectations that KM, if accomplished the right way, may transform knowledge into competitive advantage [18].

Normally micro and small enterprises use knowledge more than other traditional resources to compete. However, a micro and small enterprises (MSE) significant majority is not employing KM techniques [19]. These techniques, or practices, help enterprises to empower their knowledge generation capacity. According to Nonaka and Takeuchi [20], the knowledge generation goes on information interactions and its effective transformation occurs in 4 conversion modes, the so called SECI model: Socialization (S) - knowledge conversion from tacit into tacit; Externalization (O) - articulation process of the tacit knowledge into explicit concepts.; Combination (C) - the explicit knowledge conversion into explicit; Internalization (I) - incorporation of the explicit knowledge into tacit one.

The relationship between KM and RM is also approached by Farias et al. [6], who describe a software project risk planning approach, based on the reuse of organizational risk knowledge. Verhaegen [21] shows a KM as a tool to decision makers and management capacity improvement, especially regarding to risk related issues. Karadsheh et al. [22] present the KM as a strategic resource for organizations, and can have a major influence on risk reduction. Massingham [9] proposes and tests a Knowledge Risk Management (KRM). This author also discusses the application of KM tools and techniques for the organizational KM. Jafari et al. [23] elaborate and apply a model for Risk Management of knowledge loss at the projects management. Recently, Alhawari et al. [8] presented a structure of a Knowledge-Based Risk Management (KBRM) for information technology projects.

3 Research Method

It has been planned a Survey type research for obtaining empirical evidences. The steps will be carried out according to the sequence established in the works of Forza [24] e Bryman e Bell [25].

The objects of this study are Brazilian incubated and graduated software development enterprises. The choice of these enterprises is justified by researches as the one by Radas and Bozic [26], where Small and Medium Enterprises are

considered as economic growing propellers, as well as of employment generation, once, due to such an importance, the developed and developing countries are interest to meet ways so these enterprises accomplish innovations. Besides that, incubated enterprises present an organizational environment with better conditions for knowledge generation [13].

The population has been defined as of 89% (eighty nine) of Incubated and Graduate software Development Enterprises belonging to an Incubators, Technologic Parks and Tecnopolis association. The size of the sample, according to Malhotra [27] must be defined according to the type of the study to be carried on. The used sample was the population itself, chosen for the research, that is, all the 89 software development enterprises. This was not a random and intentional sample. For the collection of data it has been used a questionnaire validated in the research carried on by Neves [28] and made up for 31 (thirsty one) questions on different kinds of scale. The data were collected in the period from January till June 2012. There were performed eight rounds for sending the enterprises the questionnaire. At the end of the eight rounds, the total was 23 (twenty three) answers. The return rate was 26%, which, according to Forza [24] is an acceptable one.

4 Results Obtained and Analyzes

4.1 Obtained Results

The internal validation was carried out through Cronbach´s Alpha. It has been estimate the value of Cronbach´s Alpha through the Minitab 15® software, being the smallest result 0.6682 which, according to Malhotra [27] is regarded as acceptable. The external validity was obtained by reliability of the respondents (96% in charge of management and had more than two years at the enterprise), which assures validity to the obtained data. The respondents average age is 30% (thirty), whereas, considering the latter completed course, 57% concluded graduation, 30% specialization and 13% Bachelor. The average time of enterprises existence is of 4 (four) years, and the oldest is 13.5 (thirteen and a half) years old and it is a graduated enterprise, and the youngest is only 8 (eight) months old, and it an incubated enterprise. The enterprises have an average of 10 (ten) employees and the average projects time is of 8 (eight) months. They have an average annual billing of R$ 70 thousand.

Based mainly on the approach by Nonaka and Takeuchi [20], there have been identified the main techniques that contribute to the enterprises to maximize their capacity to generate knowledge. These techniques were associated with the SECI model. Fig. 1 presents the results with relation to the KM Techniques in the enterprise, in accordance with criteria (0) Never used, (1) Rarely used, (2) Little used, (3) Used and (4) Frequently used. It is worth mentioning, as a KM technique "Never used" by the enterprises, the Database Skills (43%, n=10) and as "Frequently used" the observation, imitation and practice (52%, n=12). Among the most cited we find the Telephone/computer network (65%, n=15).

Table 1 presents the evaluation of KM frequency use techniques (Fig. 1), associated to the conversion modes proposed through SECI model [20].

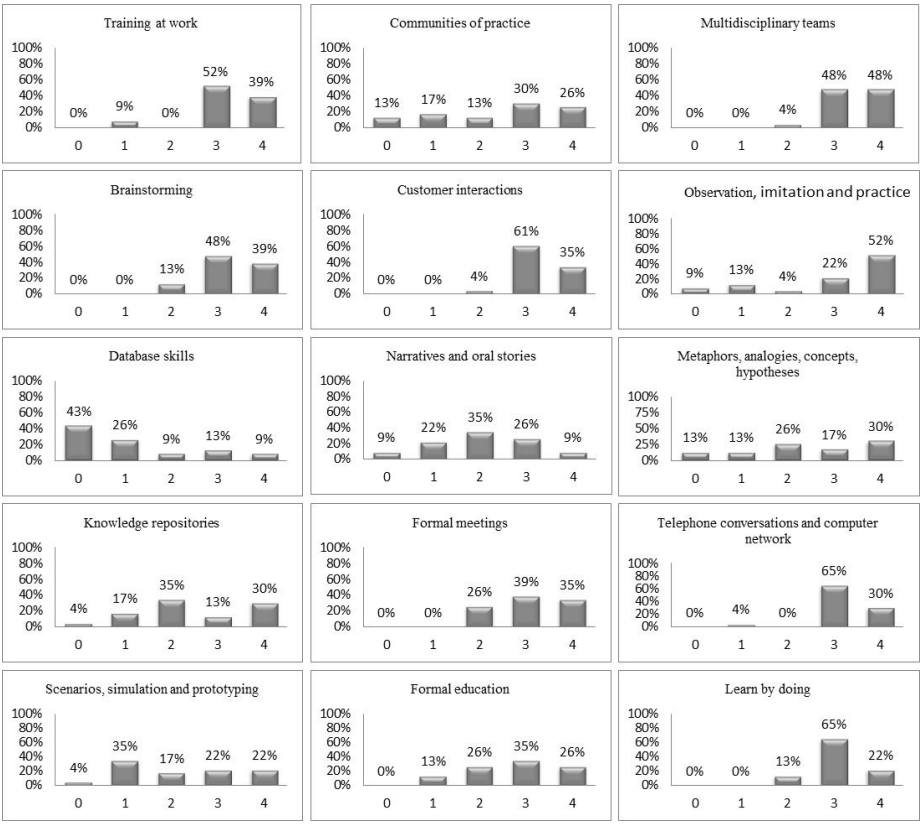

Fig. 1. KM Techniques use frequency

Table 1. Use of KM techniques frequency according to the conversion mode

Frequency	Socialization	Externalization	Combination	Internalization
0 – Never	9%	9%	1%	0%
1 - Rarely used	9%	17%	13%	0%
2 - Little used	7%	32%	17%	13%
3 - Used	39%	19%	40%	65%
4 - Frequently used	35%	23%	28%	22%

There are two aspects in this table: the "Externalization" conversion mode as "Little used" (32%) and the "Internalization" conversion mode as "Used" (65%). Most enterprises have an average level of awareness for RM (43%). However, a methodological formal use is bigger to RM (48%, n=11) than to KM (39%, n=9). As to RM, enterprises employ, in its major part, the Probability versus Impact analysis. Table 2 presents the same evaluation considering the separation of the enterprises in incubated (65%, n=15) and graduated (35%, n=8).

Table 2. Formal RM and KM methodologies use by Incubated and Graduated enterprises

Frequency	Incubated		Graduated	
	Yes	No	Yes	No
RM Formal Methodology	33% (n = 5)	67% (n = 10)	50% (n = 4)	50% (n = 4)
KM Formal Methodology	40% (n = 6)	60% (n = 9)	63% (n = 5)	38% (n = 3)

One notices the Graduated enterprises had the highest indexes regarding the use of formal methodology for RM (50%, n=4) and for KM (63%, n=5). The results highlight the research by Vick et al. [13]. According to the authors, the information management and its procedures were identified as being more aware and structured in Graduated companies. The obtained data, which were discrete variables, were transformed into continuous values, employing the Minitab l5® software and, later on, calculated their correlation [27]. The correlated factors are identified in Table 3.

Table 3. Correlation found in data analyses

Hypothesis	P-Value
There is a relationship between the enterprise annual billing and its employees number	**0.000**
There is a relationship between the enterprise annual billing and the existence time of the enterprise	**0.049**
There is a relationship between the enterprise annual billing and the project average time	0.455
There is a relationship between the enterprise annual billing and the existence of an RM methodology	0.234
There is a relationship between the enterprise annual billing and the existence of a KM methodology	**0.008**
There is a relationship between the existence of a KM methodology and the existence of an RM methodology	**0.002**
There is a relationship between the number of employees and the existence of an RM methodology	0.255
There is a relationship between the number of employees and the existence of a KM methodology	0.128
There is a relationship between the project average time and the existence of a KM methodology	0.542
There is a relationship between the project average time and the existence of an RM methodology	0.104
There is a relationship between the existing time of an enterprise and the existence of an RM methodology	0.483
There is a relationship between the existing time of an enterprise and the existence of a KM methodology	0.492
There is a relationship between the Board of Directors´ awareness level and the importance of an RM projects average time	**0.039**
There is a relationship between the enterprise annual billing and the Board of Directors´ awareness level as to the RM importance.	0.411

Observing the results (Table 3) it turns out that the hypotheses, statistically considered validated, with a significance of 5%, are the existence of a relationship among: (I) The enterprise annual billing and its number of employees, this relationship is directly pro rata, once this number is a key factor for the production amount, influencing this way the billing; (II) The annual billing and the enterprise existing time are two proportionally values, which makes sense, once a enterprise with a longer existence time tends to have a certain know-how, so having a larger project number and in a more effective way; III) a methodology existence for KM and its association with the billing, it corroborates with the statements by Anantatmula e Kanungo [17]. According to the authors, knowledge is recognized as a critical weapon to acquire and keep up competitive knowledge in business; IV) Relationship between a KM methodology and the existence of an RM methodology, the establishment of such a relationship strengthens and quantifies statements by Karadsheh et al. [22], where they present the KM processes as a strategic resource for organizations, being able to have a great influence on the risk reduction; (V) The Board of Directors awareness level as to the importance of the RM and the projects average time, the risks incorporation at projects events turns out to be feasibly timed for the projects, these times scheduled on a more realist way, considering the threats and opportunities or even bigger or smaller projects can be abandoned or encouraged, as well like decisions related to resources allocation.

4.2 Data Analysis through Partial Least Square (PLS)

The Partial Least Square regression method applies when there are: one or multiple dependent variables; highly correlated predictors; more predictors than observations [29], [30].

The calculations for analysis were performed in the Minitab15® software. It has been considered as the result the annual billing (Y), once, based on this number, one can evaluate the enterprise success in its line, once the goal is to maximize the billing of all the enterprises. The predictors variable (X) are: 1) the number of employees; (2) the enterprise existing time; (3) project average time; (4) the Board of Directors as to the importance of RM; (5) RM methodology; (6) KM methodology; (7) Socialization; (8) Externalization; 9) Combination; (10) Internalization. 5 (five) main components were selected, and the PLS mode presented explanation rates of 83.44% and P-Value of 0.000 (trust intervals of 95%). First, a PLS model was generated, when one checked the normalcy of the resulting waste, considered to be normal.

Fig. 2 presents a chart of the result of a PLS model that used the first 5 (five) main components. One used the PLS Std Coefficient Plot, which patterns allow the coefficients to be compared.

Negative coefficients indicate that the factor restrain the getting the results and positive coefficients indicate that factor contributes to achieve the result. This way, Table 4 presents the results of contributing factors, neutral and the billing restraining ones.

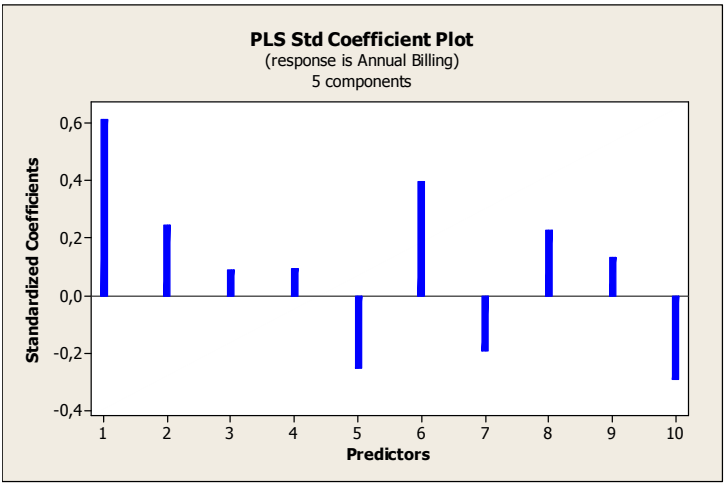

Fig. 2. PLS for the "Annual billing" result - deployed to the KM application

Table 4. Contributing factors to achieve results

Result: annual billing	10 factors (deployed the KM application)
Enhancing factors	Number of employees (1) KM methodology (6) Existing time of enterprise (2) Externalization (8)
Neutral factors	Combination (9) The Board of Directors awareness level as to the importance of RM (4) Projects average time (3)
Restraining factors	Internalization (10) RM methodology (5) Socialization (7)

As to the obtained results from the analysis of contributing factors for getting results, one highlights the existing number of employees at the enterprise (1) and use of a methodology for KM (6). Another factor that potentiates is associated to the "Externalization" (8) conversion mode, which, according to Nonaka and Takeuchi [20] has been neglected at organizations and should be taken into consideration. On the other hand, the financial results may be restricting for the use of a RM methodology (5). Considering this specified result and its relationship with the billing item, this restriction may be due to time issues and possible investment, like human resources and materials for the RM deployment.

5 Conclusions

This research aimed to analyze the use of KM techniques to the RM in software projects development and the possible influence on the enterprise revenue. It was obtained, as factors that contribute to achieving the financial results, the number of employees, the KM methodology and the enterprise existing time. It has been highlighted the "Externalization" conversion methodology as one of the factors to be observed by the enterprise, once it influences on the revenue, as a way of articulation of tacit knowledge for explicit ones. Such an articulation would allow the use of this knowledge for organization benefit, contributing especially with the help for taking decisions related to the risks. This is why it is important the existence of an environment that stimulates the knowledge sharing. The KM and methodology use techniques are still on an initial stage at the evaluated enterprises. It had been seen the many of them use KM techniques in their processes, but they haven´t formalized them so far as being KM. This attitude can make enterprises not to have the benefits the KM practice can provide, especially related to the help of taking decisions.

Taking into account that 53% of the respondents still doesn´t prioritize KM, it is still necessary to increase the awareness level as to the importance of this process for high tech enterprises, allowing the creation a proper environment for improving the organizational performance. Here is a reflection on "if and how far" these enterprises are performing a transfer strategy of their risks for the funding bodies. As the research main contribution it has been highlighted the fact of presenting, in a quantitative way, the benefits of the joint use of KM techniques and methodologies for RM, an item not very much exploited item in literature. The obtained results may show the software developing incubated and graduated enterprises managers the development of politics and strategies aiming to maximize the enterprise profits.

It is suggested, as for future research, to evaluate the efficiency of Incubated and Graduated Technological Based enterprises, through the use of Multiobjective Programming techniques, like the Goal Programming and Data Envelopment Analysis (GPDEA), having as the main variables the methodologies use for KM and RM and its relationship with the enterprises efficiency.

Acknowledgments. The authors need to express their acknowledgments to two Brazilian research agencies: the CAPES Foundation (Grant No. PE024/2008) and FAPEMIG (Grant No. PPM-00586), and especially all interviewees.

References

1. Kendrick, T.: Identifying and managing project risk: essential tools for failure-proofing your project. Amacom, NY (2003)
2. Bannerman, P.L.: Risk and risk management in software projects: A reassessment. The Journal of Systems and Software 81, 2118–2133 (2008)
3. Jiang, J.J., Klein, G., Discenza, R.: Information system success as impacted by risks and development strategies. IEEE Transactions on Engineering Management 48, 46–55 (2001)

4. Neef, D.: Managing corporate risk through better knowledge management. The Learning Organization 12, 112–124 (2005)
5. Davenport, T.H., Prusak, L.: Working Knowledge: How organizations Manage what they know. Harvard Business School Press (1998)
6. Farias, L.L., Travassos, G.H., Rocha, A.R.: Managing Organizational Risk Knowledge. Journal of Universal Computer Science 9, 670–681 (2003)
7. Wong, K.Y., Aspinwall, E.: An empirical study of the important factors for knowledge-management adoption in the SME sector. Journal of Knowledge Management 9, 64–82 (2005)
8. Alhawari, S., Karadsheh, L., Talet, A.N., Mansour, E.: Knowledge-Based Risk Management framework for Information Technology Project. International Journal of Information Management 32, 50–65 (2012)
9. Massingham, P.: Knowledge risk management: a framework. Journal of Knowledge Management 14, 464–485 (2010)
10. Bergek, A., Norrman, C.: Incubator best practice: A framework. Technovation 28, 20–28 (2008)
11. Dahlstrand, A.L.: Technology-based entrepreneurship and regional development: the case of Sweden. European Business Review 19, 373–386 (2007)
12. Bollingtoft, A.: The bottom-up business incubator: Leverage to networking and cooperation practices in a self-generated, entrepreneurial-enabled environment. Technovation 32, 304–315 (2012)
13. Vick, T.E., Nagano, M.S., Santos, F.C.A.: Identifying the information management process and knowledge creation in technology-based companies: a Brazilian comparative case study. Knowledge Management Research & Practice, http://www.palgrave-journals.com/kmrp/
14. Wallace, L., Keil, M., Rai, A.: Understanding software project risk: a cluster analysis. Information & Management 42, 115–125 (2004)
15. Huang, S.-J., Han, W.-M.: Exploring the relationship between software project duration and risk exposure: A cluster analysis. Information & Management 45, 175–182 (2008)
16. Barros, M.O., Werner, C.M.L., Travassos, G.H.: Supporting risks in software Project management. The Journal of Systems and Software 70, 21–35 (2004)
17. Anantatmula, V., Kanungo, S.: Modeling Enablers for Successful KM Implementation. Journal of Knowledge Management 14, 100–113 (2010)
18. Fan, Z.-P., Feng, B., Sun, Y.-H., Ou, W.: Evaluating knowledge management capability of organizations: A fuzzy linguistic method. Expert Systems with Applications 36, 3346–3354 (2009)
19. Maguire, S., Koh, S.C.L., Magrys, A.: The adoption of e-business and knowledge management in SMEs. Benchmarking: An International Journal 14, 37–58 (2007)
20. Nonaka, I., Takeuchi, H.: The knowledge-creating company: how Japanese companies create the dynamics of innovation. Oxford University Press, New York (1995)
21. Verhaegen, T.: Knowledge makes risks manageable. Business Insurance: Industry Focus 3, 16–17 (2005)
22. Karadsheh, L., Alhawari, S., El-Bathy, N., Hadi, W.: Incorporating knowledge management and risk management as a single process. In: Proceedings of GBDI, USA, pp. 207–214 (2008)
23. Jafari, M., Rezaeenour, J., Mazdeh, M.M., Hooshmandi, A.: Development and evaluation of a knowledge risk management model for project-based organizations A multi-stage study. Management Decision 49, 309–329 (2011)

24. Forza, C.: Survey research in operations management: a process-based perspective. International Journal of Operations & Production Management 22, 152–194 (2002)
25. Bryman, A., Bell, E.: Business research methods. Oxford University Press, NewYork (2007)
26. Radas, S., Bozic, L.: The antecedents of SME innovativeness in an emerging transition economy. Technovation 29, 438–450 (2009)
27. Malhotra, N.K.: Pesquisa de Marketing: uma orientação aplicada. Bookman, Porto Alegre (2006)
28. Neves, S.M.: Análise de riscos em projetos de desenvolvimento de software por meio de técnicas de gestão do conhecimento. Federal University of Itajuba (UNIFEI), Itajuba (2010)
29. Yacoub, F., Macgregor, J.F.: Product Optimization and Control in the Latent Variable Space of Nonlinear PLS Models. Chemometrics and Intelligent Laboratory Systems 70, 63–74 (2004)
30. Helland, I.S.: On the Structure of Partial Least Saquares Regression. Communications in Statistics, Part B Simulation and Computations 17, 581–607 (1988)

Leveraging Knowledge from Different Communities Using Ontologies

Herlina Jayadianti[1,2], Carlos Sousa Pinto[1], Lukito Edi Nugroho[2], Paulus Insap Santosa[2], and Wahyu Widayat[2]

[1] Departemento de Sistemas da Informacao, Universidade do Minho, Campus de Azurem, Guimaraes, Portugal
[2] Electrical Engineering and Information Technology, Gadjah Mada University, Yogyakarta, Indonesia
[3] Economic Development, Gadjah Mada University, Yogyakarta, Indonesia
herlinajayadianti@gmail.com, csp@dsi.uminho.pt, {Lukito,Insap}@mti.ugm.ac.id, wahyu@mep.ugm.ac.id

Abstract. The purpose of this paper is to provide research based understanding of leveraging knowledge and managing knowledge within and across several communities using the poverty domain as a case study. We hypothesize that leveraging knowledge with a good taxonomy and a good integration process are good approaches to organize and share knowledge. Problems appear when a group of people in different communities share data and collaborate using different perceptions, different concepts, different terms (terminologies), and different semantics to represent the same reality. In this paper we present an approach to solve this problem. We will generate a common set of terms based on the terms of several different storage devices, used by different communities, in order to make data retrieval independent of the different perceptions and terminologies used by those communities. We use ontologies to represent the particular knowledge of each community and discuss the use of mapping and integration techniques to find correspondences between the concepts used in those ontologies.

Keywords: Leveraging Knowledge, Knowledge Management, Common Ontology, Perception, Terminology, Ontology.

1 Introduction – Leveraging Knowledge

Information technology has lead many institutions or communities to imagine a new world of leverage knowledge. Internet made it possible for professionals allowing them to draw on the latest thinking of their peers no matter where they are located. As a result many communities are rethinking how works gets done, linking people to electronic media so they can leverage each other's knowledge. Knowledge is different from information and sharing it requires a different set of concepts and tools. Four characteristics of knowledge distinguish it from information: [1]–[3]

Á. Rocha et al. (Eds.): *Advances in Information Systems and Technologies*, AISC 206, pp. 67–76.
DOI: 10.1007/978-3-642-36981-0_7

- Knowing is a human act; whereas information is an object that can be filed, stored and moved around.
- Knowledge is created in present moment; whereas information is fully made and can sit in storage. To share knowledge we need to think about the current situation.
- Knowledge belongs to communities.
- Knowledge circulates through communities in many ways.

Leveraging knowledge involves a unique combination of human and information systems [3], [4]. Leveraging knowledge also allows us to tap into many innovations to access diverse knowledge bases and integrate them to create new competencies with multi-dimensional concepts [5], [6]. Ironically to leverage knowledge we need to focus on the community that owns it and the people who use it, not the knowledge itself. Six implications for leveraging knowledge are: [3]

- Different communities. Focus on knowledge important to both the business and the people. People naturally seek help, share insights and build knowledge in areas they care about.
- Sharing information. The ways to share knowledge should be as multidimensional as knowledge itself. Most corporate knowledge sharing efforts revolve around tools.
- Let the community decide what to share and how to share it. Knowledge needs to have an "owner" who cares. It is tempting to create organization-wide systems for sharing knowledge.
- Community support structure. Communities are held together by people who care about the community.
- Use the community's terms for organizing knowledge.
- Integrate sharing knowledge into the natural work flow.

Knowledge and learning are the only capabilities that can provide sustained competitive advantage. 'Knowledge' is the content of learning. 'Learning' is the process of gaining new knowledge, so that the firm is constantly accumulating and assimilating knowledge and this becomes the basis for creating and improving organizational routines [7]. Knowledge is a critical resource that warrants much more attention. If we are serious about managing knowledge, then we need to embrace the concepts associated with knowledge management [3], [6], [8]. Since knowledge is the sense we make of information, then the way information is organized is also a sense making device. A good taxonomy should be intuitive for those who use it.

To be "intuitive" it needs to tell the story of the key distinctions of the field, reflecting the natural way discipline members think about the field [3]. There are great temptations to make all systems for organizing knowledge the same, such as formatting information – to make it easily transferred, and having the same metadata – to make it easy searched, indexed and used in different context. However beyond that, the system for organizing information should be the community's. If a community of people sharing knowledge spans several disciplines, then such thing of terms and structures should be the *common* among those communities [3].

Having some *common* ground, among those communities, either within an application area or for some high-level general concepts, this could alleviate the problem of integrating knowledge [9]–[12]. Based on the presented reasons, we believe that ontologies with common terms and common concepts are very important in a knowledge sharing process. In this paper we describe an approach of leveraging knowledge using a common set of terms derived from several different ontologies. This paper is organized as follows: (1) Introduction; (2) Knowledge and Common ontology; (3) Implementation of the solution; (4) Conclusions.

2 Knowledge and Common Ontology

Figure 1 shows the relation between knowledge and ontologies.

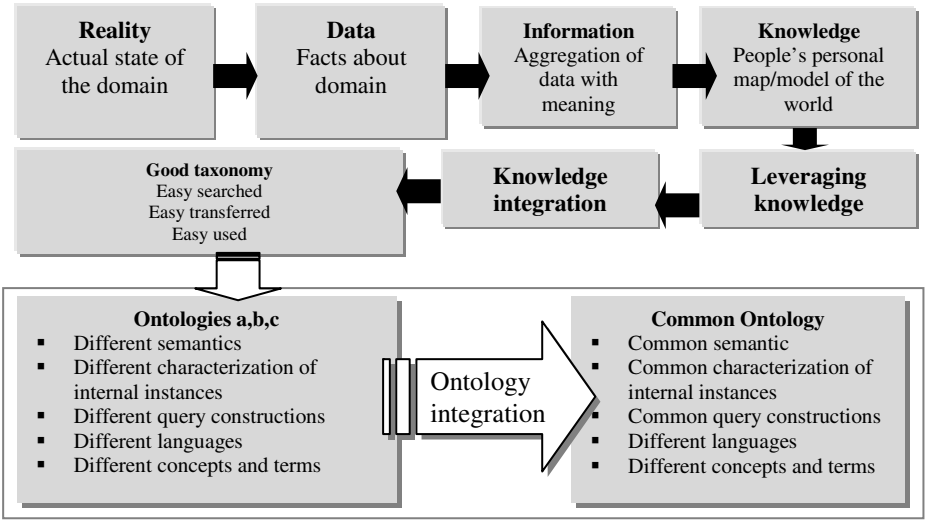

Fig. 1. Managing Knowledge, and Leveraging Knowledge with Ontologies

At the level "Reality" we represent the actual state of a particular domain. At this level we can find lots of data. Data are facts in the context of a domain of discourse. At the next level, establishing relationships between data, it is possible to derive information and expand it beyond the limits of understanding of each person. Knowledge is obtained by adding experience, reflection and reasoning to information. If different information is discussed by people, it is easy to understand what is inside their minds, either by arguments or communication, but what happens if those differences exist at the machine level? We need to combine information so that machines can "think" and understand the concepts we can find inside human brains. To do that, we can use ontologies to represent data and information of the several communities. Ontology is some formal description of a domain of discourse. However, ontology is not enough to make computers understand what is necessary. Scattered ontologies should then be incorporated and integrated into a new ontology,

a Common Ontology (See Fig.2). Ontology integration is one way to solve the problem of data, information, and semantic heterogeneity. Semantic heterogeneity on naming includes problems with synonyms (same concept with different terms) and homonyms (same term with different meanings). Semantic heterogeneity occurs when the same reality is modeled by two or more different people or systems [13].

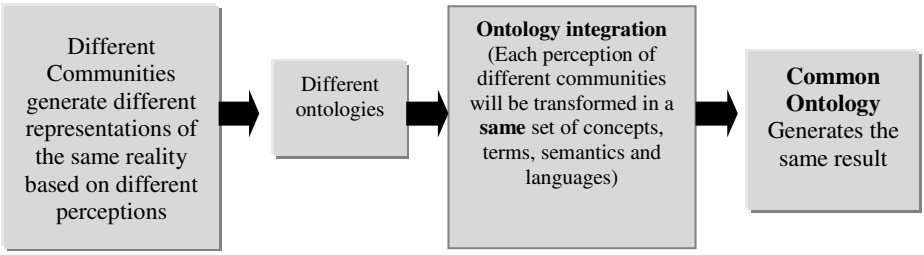

Fig. 2. Towards a Solution of Different Perceptions

The goal of ontology integration is to derive a more general domain ontology (Common Ontology) from other several ontologies in that domain. Every person has his/her own knowledge. They can justify everything based on their thoughts, perceptions and conceptualizations. Conceptualization is an abstraction of the external world inside an individual mind. It can be used to construct one or several concepts and also to interpret some reality in a conceptual way [14].

3 Implementing the Solution

Ontology integration is one way to solve the problem of semantic heterogeneity and it can be done using several approaches. For example, merging, matching or mapping. The integration of ontologies creates a new ontology by reusing other available ontologies through assembling [15]–[17], extending [18] , or specializing operations [19]. In integration processes the source ontologies and the resultant ontology can have different amounts of information [14]. Ontology process integration implies several steps. According to Noy [20] there are some specific challenges in the ontology integration process:

- Finding similarities and differences between ontologies in an automatic and semi-automatic way;
- Defining mappings between ontologies;
- Developing an ontology integration architecture;
- Composing mappings across different ontologies;
- Representing uncertainty and imprecision in mappings.

Particularly, in ontology integration, some tasks should be performed to eliminate differences and conflicts between those ontologies [20]. Ontology integration is used to find similarities and differences between ontologies. The goal of ontology

integration is to derive a more general domain ontology (Common Ontology) from several other ontologies in the same domain, into a consistent unit. The domain of both the integrated and the resulting ontologies is the same.

3.1 How to Get Common Terms – Common Ontology (CO)

Groups of people in different communities will probably have a different way of view to the same reality. Different view of each set of users is then called user view (UV) and can be implemented using an ontology. Common Ontology is expected to overcome the differences that exist in the several user views (UVs). CO will contain common terms which will then be equated with each term in the UVs. Common term is a common word recognized and used with the same meaning by different communities. To get the CO terms we use WordNet[1] , Thesaurus[2] and Swoogle[3] (See Table 1). Wordnet is a large lexical database or electronic dictionary for English [21], [22]. WordNet implements measure of similarity and relatedness among terms. Measures of similarity use information found in an *is–a* hierarchy of concepts, and quantify how much concept A is similar to concept B [23]. Thesaurus is a reference work that lists words grouped together according to similarity of meaning. Swoogle is the first Web search engine dedicated to online semantic data. Its development was partially supported by DARPA and NFS (National Science Foundation).

Table 1. Equivalences for some terms related to poverty from different applications

Search string	Synonym		
	Wordnet 2.1 (Noun)	Thesaurus (Noun)	Swoogle (Terms)
Hospital	Infirmary, medical institution	Clinic, emergency room, health service, hospice, infirmary, rest home.	Hospital, hospital,
Clinic	Medical institution, Session, Medical building, health facility, healthcare facility	Emergency room, hospice, infirmary, nursing home, rest home.	Clinic, Clinical, ClinicalTreatment

There are two senses for the term hospital in Wordnet (version 2.1).

Sense 1. *hospital, infirmary -- (a health facility where patients receive treatment)*
=> Medical building, health facility, healthcare facility -- (building where medicine is practiced)

Sense 2. *hospital -- (a medical institution where sick or injured people are given medical or surgical care)*
=> Medical institution -- (an institution created for the practice of medicine)

[1] http://wordnet.princeton.edu/
[2] http://thesaurus.com/
[3] http://swoogle.umbc.edu/
Swoogle was the first search engine dedicated to online semantic data. Its development was partially supported by DARPA and NFS (National Science Foundation).

There are three senses for the term clinic in Wordnet (version 2.1).

Sense 1. *clinic -- (a medical establishment run by a group of medical specialists)*
=> Medical institution -- (an institution created for the practice of medicine)

Sense 2. *clinic -- (meeting for diagnosis of problems and instruction or remedial work in a particular activity)*
=> Session -- (a meeting for execution of a group's functions)

Sense 3. *clinic -- (a healthcare facility for outpatient care)*
=> Medical building, health facility, healthcare facility -- (building where medicine is practiced)

3.2 Implementation

Let's consider the referred case study of poverty and two ontologies or user views both representing that domain. By using WordNet, Thesaurus and Swoogle we can find synonyms and similarities that were chosen from ontology UV1 and Ontology UV2. The next step is to find the number of each synonym with Google and Swoogle. The results provided by Google and Swoogle are different due to the number of documents that are available in each system. Google provides more documents than Swoogle. Currently, Swoogle only indexes some metadata about Semantic Web documents[4].

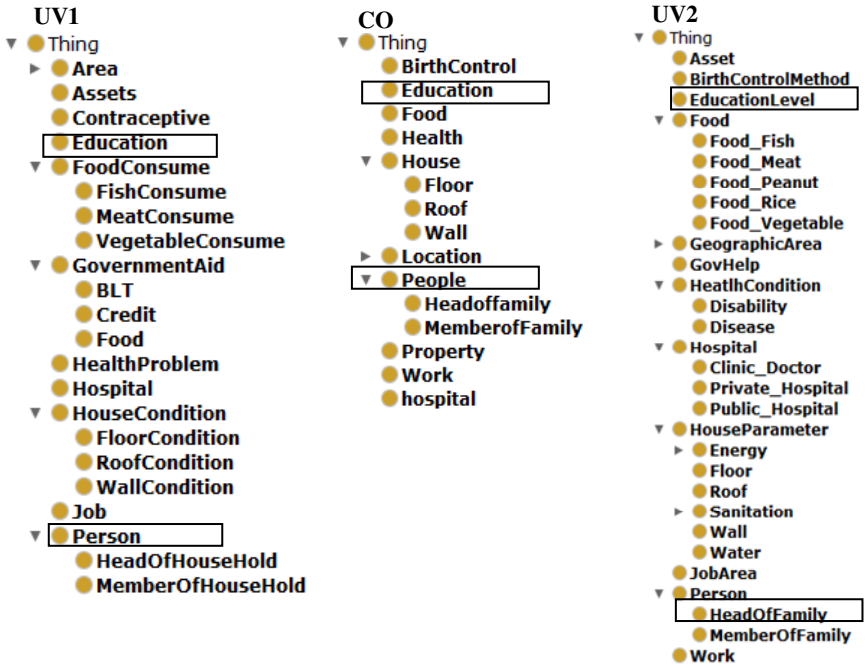

Fig. 3. Terms in ontology UV1, UV2 and CO

[4] `http://swoogle.umbc.edu/index.php?option=com_swoogle_manual&manual=faq`

Based on data from Swoogle and Google, then we selected the term with highest references. We assume that the number of references reflects a widely and commonly usage of the term by users. We use a common term as a term in CO. For example: term `Person` is more specific than `People` (See Fig.3). We use Wordnet, Swoogle and Google not only for comparing the number of result to get common terms but also to find common ObjectProperties. For example: ObjectProperties `hasFloorMadeFrom` (CO) is more general than `hasHouseFloorMadeFrom` (UV2) and `hasLargestFloorAreaMAdeFrom` (UV1) (See Fig. 4). Poverty is not the focus of our research. We just use that case as a real scenario that allows us to demonstrate our approach. We combine different existing terminologies about the same reality (poverty in this case) used by different communities in order to get a common set of terms that can be transparently used by those communities, while maintaining the original terms in the data sources. We use Indonesia as the country for the example because in that country there are several communities in charge of dealing with poverty data, generating problems due to differences in the criteria used to make their surveys, even considering that the semantics of these different criteria are the same. For example, let's consider the two communities, BKKBN[5] and BPS[6],that are responsible for collecting data on poverty. Each community has a different system and use different sets of terms to describe the same domain and

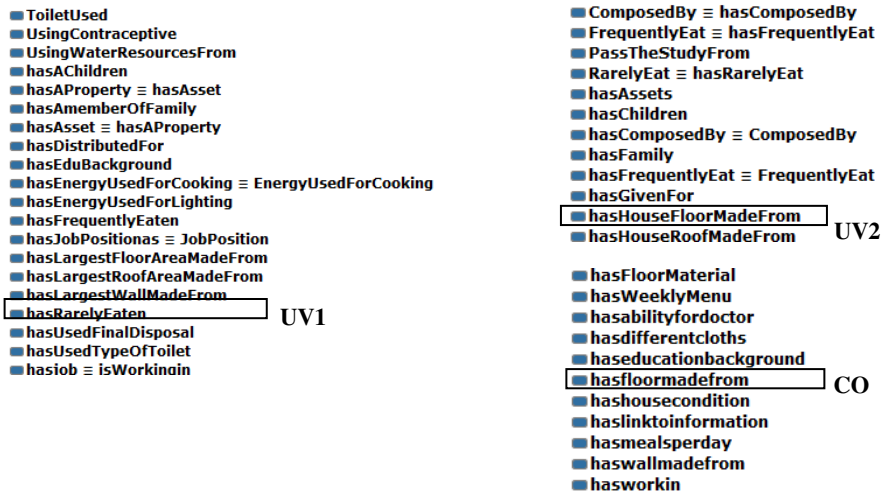

Fig. 4. ObjectProperties in ontology UV1, UV2 and CO

[5] Badan Keluarga Berencana Nasional (BKKBN) or National Population and Family Planning Board is a governmental agencies that appointed to conduct a survey of poverty in Indonesia. `www.bkkbn.go.id`

[6] Badan Pusat Statistik (BPS) or Central Berau of Statistic is a non departmental government institution directly responsible to the President of Indonesia. `www.bps.go.id`

different criteria to classify people as poor or not. To be similar *(≅)* or not equal *(≠)* depend on several factors, such as the programmer's interpretation, the needs of the system itself, and last but not least the domain/area that we are talking about. Currently, both communities are working separately to collect and manage data on poverty. Each community sends data to the government based on its perception.

3.3 Testing Queries

SPARQL[7] is a query for Resource Description Framework (RDF)[8]. SPARQL can be used to express queries across diverse data sources whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying and also supports extensible values for testing and constraining queries [24]. SPARQL commands from our work shows below.

```
Prefix :<http://www.semanticweb.org/CO.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?People ?Food ?Floor ?Location
WHERE { ?People :hasWeeklyMenu?Food. ?People :hasFloorMaterial?Floor.
?People :islivinginvillage?Location.
?Food :FoodName ?value1. ?Floor :Material ?value2. ?Location
:VillageName ?value3.
FILTER (?value1 ='Vegetable' && ?value2 ='Soil'  && ?value3
='Widodomartani')
}
Prefix : <http://www.semanticweb.org/UV1.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?Person ?Food ?Job ?Floor ?Area
WHERE {?Person :RarelyEat ?Food. ?Person :hasJobPositionAs ?Job. ?Person
:hasHouseFloorMadeFrom ?Floor. ?Person :isLivinginVillage ?Area.
?Food :FoodName ?value1. ?Job :JobName ?value2. ?Floor :TypeOfFloor
?value3. ?Area :hasName ?value4.
FILTER (?value1 ='Chicken' && ?value2 ='Farmer' && ?value3 ='Soil' &&
?value4 ='Widodomartani')
}
```

ObjectProperties `hasHouseFloorMadeFrom` in UV1 is equivalent to Object-Properties `hasFloorMaterial` in CO. ObjectProperties `hasFloorMaterial` is more common than ObjectProperties has `HouseFloorMadeFrom`. In this work, we found the same result these queries (See Fig.5).Our future work will include

[7] `http://www.w3.org/TR/rdf-sparql-query/`
[8] RDF is a standard model for data interchange on the web. `http://www.w3.org/RDF/`

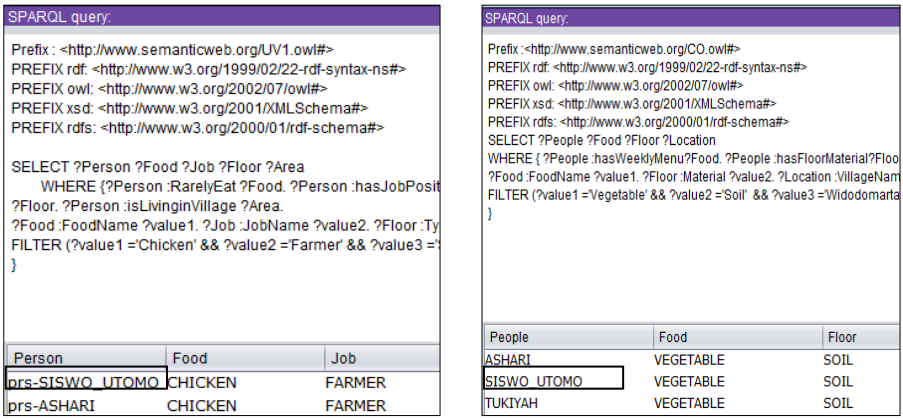

Fig. 5. The same result between ontology UV1 and CO

functionalities that will allow users ask queries using JSP[9] (Java Server Pages) and JENA[10] ontology API against OWL/RDF files. Through the ontology API, JENA provides a consistent programming interface for ontology applications.

4 Conclusions

In this research we try to leveraging knowledge by using an ontology integration as a process to create a new ontology (Common Ontology). Using this approach it is possible to share different conceptualizations, different terminologies, and different meanings between different systems. We believe that ontology integration is one of the best approaches to solve the problem of data and semantic heterogeneity.

Acknowledgment. We would like to acknowledge the support of the Erasmus Mundus EuroAsia program for the research foundation of this research, and also to acknowledge Universidade do Minho and Universitas Gadjah Mada for the collaboration.

References

[1] McDermott, R.: Knowing in community. IHRIM, 19 (2000)
[2] McDermott, R.: Knowing is a human act, Informatik-Informatique. Zeitschrift der Schweizerischen Informatikorganisationen 1, 7–9 (2002)
[3] McDermott, R.: Why information technology inspired but cannot deliver knowledge management. Knowledge and Communities 41(4), 21–35 (2000)

[9] Javaserver Pages is a technology provides a simplified, fast way to create dynamic web content.
http://www.oracle.com/technetwork/java/javaee/jsp/index.html
[10] http://jena.apache.org/

[4] Lesser, E.L., Fontaine, M.A., Slusher, J.A.: Knowledge and Communities. Routledge (2000)
[5] Hatteland, C.J.: Commentary on leveraging knowledge based resources: The role of contracts. Journal of Business Research 65(2), 162–163 (2012)
[6] Mudambi, R., Swift, T.: Leveraging knowledge and competencies across space: the next frontier in international business. Journal of International Management 17(3), 186–189 (2011)
[7] Carayannis, E.G.: Knowledge-driven creative destruction, or leveraging knowledge for competitive advantage strategic knowledge arbitrage and serendipity as real options drivers triggered by co-opetition, co-evolution and co-specialization. Industry and Higher Education 22(6), 343–353 (2008)
[8] Quinn, J.B.: Strategic outsourcing: leveraging knowledge capabilities. Sloan Management Review 40(4), 9–21 (1999)
[9] Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag New York Inc. (2007)
[10] Gangemi, A., Pisanelli, D., Steve, G.: Ontology integration: Experiences with medical terminologies. In: Formal Ontology in Information Systems, vol. 46, pp. 98–94 (1998)
[11] Noy, N.F.: Semantic integration: a survey of ontology-based approaches. ACM Sigmod Record 33(4), 65–70 (2004)
[12] Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S.: Ontology-based integration of information-a survey of existing approaches. In: IJCAI 2001 Workshop: Ontologies and Information Sharing, vol. 2001, pp. 108–117 (2001)
[13] Bouzeghoub, M., Goble, C.A., Kashyap, V., Spaccapietra, S. (eds.): ICSNW 2004. LNCS, vol. 3226. Springer, Heidelberg (2004)
[14] Xue, Y.: Ontological View-driven Semantic Integration in Open Environments. The University of Western Ontario (2010)
[15] Pinto, H.S., Martins, J.P.: A methodology for ontology integration. In: Proceedings of the 1st International Conference on Knowledge Capture, pp. 131–138 (2001)
[16] Kim, K.Y., Manley, D.G., Yang, H.: Ontology-based assembly design and information sharing for collaborative product development. Computer-Aided Design 38(12), 1233–1250 (2006)
[17] Farquhar, A., Fikes, R., Rice, J.: Tools for assembling modular ontologies in Ontolingua. In: Proc. of Fourteenth American Association for Artificial Intelligence Conference (AAAI 1997), pp. 436–441 (1997)
[18] Heflin, J., Hendler, J.: Dynamic ontologies on the web. In: Proceedings of the National Conference on Artificial Intelligence, pp. 443–449 (2000)
[19] Zimmermann, A., Krotzsch, M., Euzenat, J., Hitzler, P.: Formalizing ontology alignment and its operations with category theory. Frontiers in Artificial Intelligence and Applications 150, 277 (2006)
[20] Noy, N.F., Crubézy, M., Fergerson, R.W., Knublauch, H., Tu, S.W., Vendetti, J., Musen, M.A.: Protégé-2000: An Open-Source Ontology-Development and Knowledge-Acquisition Environment: AMIA 2003 Open Source Expo. In: AMIA Annual Symposium Proceedings, vol. 2003, p. 953 (2003)
[21] Fellbaum, C.: WordNet. Theory and Applications of Ontology: Computer Applications, 231–243 (2010)
[22] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: An on-line lexical database*. International Journal of Lexicography 3(4), 235–244 (1990)
[23] Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet: Similarity: measuring the relatedness of concepts. In: Demonstration Papers at HLT-NAACL 2004, pp. 38–41 (2004)
[24] Pérez, J., Arenas, M., Gutierrez, C.: Semantics and Complexity of SPARQL. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 30–43. Springer, Heidelberg (2006)

An Approach for Deriving Semantically Related Category Hierarchies from Wikipedia Category Graphs

Khaled A. Hejazy and Samhaa R. El-Beltagy

Center for Informatics Science, Nile University, Cairo, Egypt
khaledhejazy86@yahoo.com, samhaa@computer.org

Abstract. Wikipedia is the largest online encyclopedia known to date. Its rich content and semi-structured nature has made it into a very valuable research tool used for classification, information extraction, and semantic annotation, among others. Many applications can benefit from the presence of a topic hierarchy in Wikipedia. However, what Wikipedia currently offers is a category graph built through hierarchical category links the semantics of which are undefined. Because of this lack of semantics, a sub-category in Wikipedia does not necessarily comply with the concept of a sub-category in a hierarchy. Instead, all it signifies is that there is some sort of relationship between the parent category and its sub-category. As a result, traversing the category links of any given category can often result in surprising results. For example, following the category of "Computing" down its sub-category links, the totally unrelated category of "Theology" appears. In this paper, we introduce a novel algorithm that through measuring the semantic relatedness between any given Wikipedia category and nodes in its sub-graph is capable of extracting a category hierarchy containing only nodes that are relevant to the parent category. The algorithm has been evaluated by comparing its output with a gold standard data set. The experimental setup and results are presented.

Keywords: Wikipedia, Semantic relatedness, Semantic similarity, Graph analysis, Category hierarchy, Hierarchy extraction.

1 Introduction

Wikipedia is an online encyclopedia that has more than 23,000,000 articles in which, more than 4 Millions articles are in English covering a wide variety of topics. Articles are maintained by more than 100,000 active volunteer contributors. As Wikipedia is written collaboratively by anonymous volunteers, anyone can write and change Wikipedia articles. It is assumed that contributors will follow a set of policies and guidelines developed by the Wikipedia community. However, there is nothing in place to enforce editing policies before or during contributing[1] which means that breaches to Wikipedia's policies and guidelines are being conducted by its community, greatly affecting its quality.

[1] http://en.wikipedia.org/wiki/Wikipedia

Á. Rocha et al. (Eds.): *Advances in Information Systems and Technologies*, AISC 206, pp. 77–86.
DOI: 10.1007/978-3-642-36981-0_8

Just like articles, Wikipedia's categories are socially annotated. When creating new categories and relating them to previously created ones, there is no strict enforcement of which higher-level categories a child sub-category can belong to; thus, Wikipedia's category structure is not a tree, but a graph in which links between nodes, have loosely defined semantics.

Consequently a sub-category in Wikipedia does not necessarily comply with the concept of a sub-category in a hierarchy. A category label in Wikipedia is simply intended as a way for users to navigate among articles, and only signifies that there is some sort of a relationship between the parent category and its sub-category that is not necessarily of the type "is-a" which is expected in a hierarchical Knowledge Organization System. This problem causes irregularity in semantics between categories that is amplified in deeper levels. For example, following the category of "Computing" down its sub-category links, the totally unrelated category of "Theology" appears. Also, the graph nature of the Wikipedia category structure means that following the sub-category links of any given category, can eventually lead back to the same category. Detecting and eliminating cycles is a minor issue. Detecting sub-categories that should be considered as belonging to any given category is the main challenge addressed by this work. To address this challenge, an approach for measuring lexical semantic relatedness between Wikipedia's categories and nodes in their sub-graphs and using this as an indicator for relatedness, was developed.

In this paper, we introduce this new approach for deriving semantically related category hierarchies from Wikipedia category graphs and extracting a category hierarchy containing only sub-categories that are relevant to the parent category.

The rest of the paper is organized as follows; firstly, related work is presented in section (2), the proposed approach is described in section (3), the procedure followed for evaluating our approach and the experimental results are presented in section (4). Analysis of the results is discussed in section (5). And finally section (6) concludes this paper.

2 Related Work

Since its inception, Wikipedia has undergone tremendous growth, and today it is the largest online encyclopedia known to date. Wikipedia has been widely used as a huge resource of concepts and relationships for text mining tasks; like classification, information extraction, and computing semantic relatedness of natural language texts, among others. Most research works that make use of Wikipedia have used Wikipedia's concepts and relationships as is, except for some preprocessing and slight modifications. No previous research (as far as the authors are aware) addressed semantic irregularity between categories in Wikipedia's categorization system.

Wikipedia's categories' growth has previously been analyzed in [1], where an algorithm that semantically maps articles by calculating an aggregate topic distribution through the articles' category links to the 11 top Wikipedia categories (manually selected). Semantic relatedness for category nodes is then calculated through link distance metrics, such as the length of the shortest path between two nodes.

The evolution of Wikipedia's category structure over time has been studied in [2]. Results of this research have shown that the Wikipedia category structure is relatively stable for a bottom-up evolved system. However, the work did not address the accuracy of the category structure.

Wikipedia has been used for measuring lexical semantic relatedness between words or text passages. Explicit Semantic Relatedness (ESA) [3] has been shown as a successful measure for semantic relatedness. It treats each Wikipedia article as a dimension in a vector space. Texts are then compared by projecting them into the Wikipedia articles' space, then measuring the similarity between vectors using conventional metrics like cosine similarity. Because this work relies mostly on individual articles, the category structure of Wikipedia was not an issue.

Wikipedia has been used to compute semantic relatedness by taking the categorization system of Wikipedia as a semantic network [4].

Wikipedia Link-based Measure [5] also measures the semantic similarity of two Wikipedia pages by comparing their incoming and outgoing links. The score is determined using several weighting strategies applied to the overlap score of the articles' links.

In this paper, we propose an approach for deriving semantically related category hierarchies from Wikipedia category graphs. Our approach is somehow similar to ESA, except the fact that we are measuring semantic relatedness between categories instead of articles or words. Also, we use a key-phrase extraction for dimensionality reduction.

3 Methodology

Detecting semantically related categories based on measuring lexical semantic relatedness between them requires an efficient representation for each category. A TF-IDF scheme [6] has been used to assign weights to the feature vectors representing Wikipedia categories. In the following subsections, we start with the pre-processing step; in which we discuss the data sources with their components and the pre-processing steps conducted before these data are used, and then we discuss the steps of generating the feature vectors of Wikipedia categories.

3.1 Pre-processing

Wikipedia's backups are created regularly by the Wikimedia Foundation[2]. These dumps are publicly available. We have used Wikipedia's XML dump release 02-05-2012, which contains all Wikipedia article pages. The size of the uncompressed dump is around 38 GB.

Pages in this xml dump are represented by multiple tags. From those our system uses the page's unique ID, page's title, page's time stamp, and page's text.

[2] www.wikimedia.org

In order to handle this large XML dump file, apache Solr [7] has been used to index it through its Data Import Handler, which also facilitated the searching processes required by our followed approach.

We also acquired some relevant SQL files from the same source in order to allow us to re-construct the categorization graph of Wikipedia. One of these files is the enwiki-20120502-category.sql.gz which is an SQL file containing metadata for each category in Wikipedia; its category ID (differs from the page ID), its title, and the number of its pages and subcategories. The other is 20120502-categorylinks.sql.gz which is an SQL file has been acquired, and used for building Wikipedia's categorization graph. The SQL file contains the page IDs of any page defined as a category member, the page title of the category's description page, the time stamp of the approximate addition time of the link, the category link type that determines whether the page ID is a page, a sub-category or a file, along with some other attributes for sorting and for defining the collation of the category links.

Pages in Wikipedia are not only articles; categories are special pages that are used to group articles together, and their titles start with the namespace "Category:". Also, there are administrative pages with different namespaces that are not used to share encyclopedic information, but rather to preserve policies created and implemented

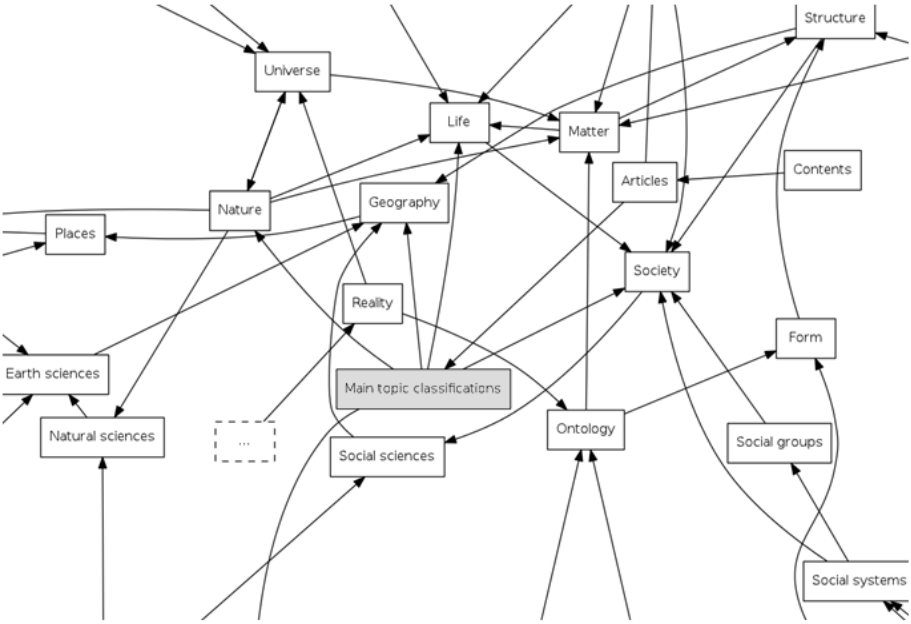

Fig. 1. A sub-graph in Wikipedia showing the category "Main topic classifications" and some of its super-categories and sub-categories