



Association for
Computing Machinery

Advancing Computing as a Science & Profession

Kaiserslautern, Germany
October 9-10, 2008



ESEM'08

Proceedings of the 2008 ACM-IEEE International Symposium on
Empirical Software Engineering and Measurement

Sponsored by:

ACM SIGSOFT & IEEE CS

Supported by:

Siemens AG & Robert Bosch GmbH

Kaiserslautern, Germany
October 9-10, 2008



Association for
Computing Machinery

Advancing Computing as a Science & Profession

8
ACM-IEEE International Symposium on
Emerging Engineering and Measurement

ICS

Bosch GmbH





**Association for
Computing Machinery**

Advancing Computing as a Science & Profession

The Association for Computing Machinery
2 Penn Plaza, Suite 701
New York, New York 10121-0701

Copyright © 2008 by the Association for Computing Machinery, Inc. (ACM). Permission to make digital or hard copies of portions of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permission to republish from: Publications Dept., ACM, Inc. Fax +1 (212) 869-0481 or <permissions@acm.org>.

For other copying of articles that carry a code at the bottom of the first or last page, copying is permitted provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Notice to Past Authors of ACM-Published Articles

ACM intends to create a complete electronic archive of all articles and/or other material previously published by ACM. If you have written a work that has been previously published by ACM in any journal or conference proceedings prior to 1978, or any SIG Newsletter at any time, and you do NOT want this work to appear in the ACM Digital Library, please inform permissions@acm.org, stating the title of the work, the author(s), and where and when published.

ISBN: 978-1-59593-971-5

Additional copies may be ordered prepaid from:

ACM Order Department
PO Box 11405
New York, NY 10286-1405

Phone: 1-800-342-6626
(US and Canada)
+1-212-626-0500
(all other countries)
Fax: +1-212-944-1318
E-mail: acmhelp@acm.org

ACM Order Number 594087

Printed in the USA

Message from the General Chair

It is my great pleasure to welcome you to the Second International Symposium on Empirical Software Engineering and Measurement (ESEM 2008) in Kaiserslautern, Germany. Empirical software engineering and measurement are important sub-disciplines aimed at advancing software engineering towards an engineering discipline. They focus on measurement in order to better understand, control, and improve software development processes, and on empirical studies to better model products and processes and understand process products.

The 1st ESEM was held in Madrid in September 2007. Although this is only the second ESEM symposium, this series continues the long tradition of two high-quality conference series. ESEM resulted from the merger of the IEEE/ACM International Symposium on Empirical Software Engineering (ISESE) and the IEEE International Software Metrics Symposium (METRICS). ISESE was held five times, first in Japan in 2002 and last in Brazil in 2006. As its name suggests, ISESE's primary goal was to disseminate knowledge on and from experimental research on software engineering. METRICS is a long-standing conference series, which was held first in the United States in 1993 and last in Italy in 2005. METRICS' foremost objective was to disseminate knowledge on and from research on software metrics.

ESEM 2008 is part of the week-long Experimental Software Engineering International Week (ESEIWI) in Kaiserslautern, Germany. It is an honor to welcome this conference to Kaiserslautern. Kaiserslautern has evolved into one of the premier centers for software & systems research in Europe. About 800 scientists and engineers – spread across 3 departments of the University of Kaiserslautern and several institutes – are working on advancing the state of the art in our field. The institutes include a Max-Planck Institute on Software Systems, two Fraunhofer Institutes – for Industrial Mathematics (ITWM) and Experimental Software Engineering (IESE) – and the German Research Institute for Artificial Intelligence (DFKI). I invite you to learn more about any of these institutes during your stay.

We have a full two-day program of three parallel tracks on each day including two keynotes, technical papers, short papers, and a panel (please see the message from the Program Chairs for details).

Putting together *ESEM 2008* has been a team effort. First of all, I would like to thank the authors of all submitted papers. Furthermore, I would like to express my gratitude to the program committee and to all external reviewers, who worked very hard on reviewing papers and providing suggestions for their improvements. I would also like to thank Sebastian Elbaum and Jürgen Münch, the program co-chairs Maria Teresa Baldassarre - the short papers chair, and Haruka Nakao - the poster program chair, as well as Andreas Jedlitschka, the treasurer, and Marcus Ciolkowski, the local arrangements chair. Finally, I would like to thank our society sponsor, ACM SIGSOFT, for their continued support of these successful meetings, and our industrial supporters Robert Bosch GmbH and Siemens AG for their financial support for the social events surrounding the conference.

I hope that you will find this program interesting and thought-provoking and that the symposium will provide you with a valuable opportunity to share ideas with other researchers and practitioners from institutions around the world.

Dieter Rombach

ESEM08 General Chair

*University of Kaiserslautern & Fraunhofer IESE
Kaiserslautern, Germany*

Message from the Program Co-Chairs

We would like to welcome you to Kaiserlautern, Germany, for the Second International Symposium on Empirical Software Engineering and Measurement – ESEM 2008. This year's technical program continues to advance the field of empirical software engineering and metrics with thoroughly reviewed technical papers complemented by two remarkable keynotes and a vibrant panel.

The call for research papers attracted over 100 submissions. To assess these submissions, the program committee members performed over 320 reviews and carried out dozens of online discussions. This process resulted in the selection of 28 papers that form our technical program. The papers cover a variety of topics, ranging from observational studies of programmers to measurement mechanisms for assessing technologies and organizations. In addition to the technical papers, short paper sessions and a poster exhibition will help to provide further coverage of ongoing work.

The keynote speakers and the panel will broaden and enrich the program by providing insights and reflections on our community strengths, weaknesses, and promising research directions. The keynotes will be given by Professor Mary Shaw from Carnegie Mellon University, USA, and Mr. Harald Hönninger, vice-president of Bosch corporate research division. The panel consists of highly renowned experts and promises interesting insights with respect to success stories, lessons learned from failures, and the next frontiers of empirical software engineering.

We would like to thank the authors who provided the content for the program, and to express our sincere gratitude to the program committee members and the external reviewers who volunteered time and resources to make this program possible. We would also like to thank the General Chair, Dieter Rombach, for his support and guidance, and Andreas Jedlitschka and Marcus Ciolkowski for the local organization of this conference and the maintenance of the website.

We hope that you will enjoy this program and that the symposium provides you with a valuable opportunity to share ideas with other researchers and practitioners.

Sebastian Elbaum

*Program Co-Chair
University of Nebraska-Lincoln (USA)*

Jürgen Münch

*Program Co-Chair
Fraunhofer Institute for Experimental
Software Engineering (Germany)*

Table of Contents

ESEM'08 Organization.....x

Keynote Address

Session Chair: J. Münch (*Fraunhofer Institute for Experimental Software Engineering*)

- **Using Empirical Methods to Improve Industrial Technology Transfer**.....1
Harald Hoeminger, Mark Müller (*Robert Bosch GmbH*)

Full Papers

Session 1A: Coordination and Communication

Session Chair: F. Lanubile (*University of Bari*)

- **Socio-Technical Congruence: A Framework for Assessing the Impact of Technical and Work Dependencies on Software Development Productivity**.....2
Marcelo Cataldo (*Bosch Corporate Research*), James D. Herbsleb, Kathleen M. Carley (*Carnegie Mellon University*)
- **A Multiple Case Study Investigating the Interaction between Manufacturing and Development Organizations in Automotive Software Engineering**.....12
Joakim Pernstål, Ana Magazinic, Peter Öhman (*Chalmers, Computer Science & Engineering*)

Session 1B: Testing and Analysis

Session Chair: S. Vegas (*Universidad Politécnica de Madrid*)

- **Empirical Evaluations of Regression Test Selection Techniques: A Systematic Review**..22
Emelie Engström, Mats Skoglund, Per Runeson (*Lund University*)
- **Capture-recapture in Software Unit Testing — A Case Study**.....32
Hanna Scott, Claes Wohlin (*Blekinge Institute of Technology*)
- **On Establishing a Benchmark for Evaluating Static Analysis Alert Prioritization and Classification Techniques**.....41
Sarah Heckman, Laurie Williams (*North Carolina State University*)

Session2A: Estimation Models I

Session Chair: C. Seaman (*University of Maryland, Baltimore*)

- **Comparative Studies of the Model Evaluation Criteria MMRE and PRED in Software Cost Estimation Research**.....51
Dan Port (*University of Hawaii at Manoa*), Marcel Korte (*University of Applied Sciences and Arts Dortmund*)
- **Phase Distribution of Software Development Effort**.....61
Ye Yang, Mei He, Mingshu Li, Qing Wang (*Chinese Academy of Sciences*), Barry Boehm (*University of Southern California*)
- **Combining Regression and Estimation by Analogy in a Semi-parametric Model for Software Cost Estimation**.....70
Nikolaos Mitrás, Lefteris Angelis (*Aristotle University of Thessaloniki*)

Session2B: Modeling and Architecture

Session Chair: J. Carver (*University of Alabama*)

- **An Industrial Case Study of Architecture Conformance**.....80
Jacek Rosik (*Lero, University of Limerick*), Andrew Le Gear, Jim Buckley (*University of Limerick*), Muhammad Ali Babar (*Lero, University of Limerick*)
- **A Survey into the Rigor of UML Use and its Perceived Impact on Quality and Productivity**.....90
Ariadi Nugroho, Michel R.V. Chaudron (*Leiden University*)
- **Model-based Functional Size Measurement**.....100
Luigi A. Lavazza, Vicri del Bianco (*University of Insubria*), Carla Garavaglia (*Syrea - Intecy*)

Keynote Address

Session Chair: S. Elbaum (*University of Nebraska*)

- **Empirical Challenges in Ultra Large Scale Systems** 110
Mary Shaw (*Carnegie Mellon University*)

Session 3A: From the Programmers' Trenches

Session Chair: H. Erdogmus (*NRC Institute for Information Technology*)

- **Problems in Agile Trenches** 111
Mira Kajko-Mattsson (*Stockholm University and Royal Institute of Technology*)
- **Pair Programming: What's in it for Me?** 120
Andrew Begel, Nachiappan Nagappan (*Microsoft Research*)
- **Why Do Programmers Avoid Metrics?** 129
Medha Umarji, Carolyn Seaman (*University of Maryland, Baltimore County*)

Session 3B: Inspections

Session Chair: D. Pfahl (*Simula Research Laboratory*)

- **The Impact of Time Controlled Reading on Software Inspection Effectiveness and Efficiency: A Controlled Experiment** 139
Kai Petersen, Kari Rönkkö, Claes Wohlin (*Blekinge Institute of Technology*)
- **Defect Categorization: Making Use of a Decade of Widely Varying Historical Data** 149
Carolyn B. Seaman, Forrest Shull, Myrna Regardie, Denis Elbert, Raimund L. Feldmann (*University of Maryland*),
Yuepu Guo (*University of Maryland Baltimore County*),
Sally Godfrey (*NASA Goddard Space Flight Center*)
- **Evaluation of Capture-Recapture Models for Estimating the Abundance of Naturally-Occurring Defects** 158
Gursimran Singh Walia (*Mississippi State University*), Jeffrey C Carver (*University of Alabama*)

Session 4A: Metrics and Methodology

Session Chair: M. Oivo (*University of Oulu*)

- **Some Lessons Learned in Conducting Software Engineering Surveys in China** 168
Junzhong Ji (*Beijing University of Technology*),
Jingyue Li, Reidar Conradi (*Norwegian University of Science and Technology*),
Chunhuan Liu (*Beijing University of Technology*),
Jianqiang Ma (*Norwegian University of Science and Technology*),
Weibing Chen (*Beijing University of Technology*)
- **Strength of Evidence in Systematic Reviews in Software Engineering** 178
Tore Dybå, Torgeir Dingsøyir (*SINTEF ICT*)
- **Refining the Axiomatic Definition of Internal Software Attributes** 188
Sandro Morasca (*Università degli Studi dell'Insubria*)

Session 4B: Faults and Failures

Session Chair: L. Williams (*North Carolina State University*)

- **Quantitative Analysis of Faults and Failures with Multiple Releases of SoftPM** 198
Shujian Wu, Qing Wang, Ye Yang (*The Chinese Academy of Sciences*)
- **Iterative Identification of Fault-Prone Binaries Using In-Process Metrics** 206
Lucas Layman (*North Carolina State University*),
Gunnar Kudrjavets, Nachiappan Nagappan (*Microsoft Corporation*)

Session 5A: Estimation Models II

Session Chair: M. Ciolkowski (*Fraunhofer Institute for Experimental Software Engineering*)

- **A Constrained Regression Technique for COCOMO Calibration** 213
Vu Nguyen, Bert Steece, Barry Boehm (*University of Southern California*)
- **Reducing Biases in Individual Software Effort Estimations: A Combining Approach** 223
Qi Li, Qing Wang, Ye Yang, Mingshu Li (*Chinese Academy of Sciences*)
- **Any Other Cost Estimation Inhibitors?** 233
Ana Magazinic, Joakim Pernstål (*Chalmers*)
- **Session 5B: From the Manager's Trenches**
Session Chair: P. Runeson (*Lund University*)
- **Empirical Results from Using Custom-Made Software Project Control Centers in Industrial Environments** 243
Marcus Ciolkowski, Jens Heidrich (*Fraunhofer IESE*), Frank Simon (*SQS AG*), Mathias Radtke (*BTU Cottbus*)
- **A Survey on Software Cost Estimation in the Chinese Software Industry** 253
Da Yang, Qing Wang, Mingshu Li, Ye Yang, Kai Ye, Jing Du (*Chinese Academy of Sciences*)
- **A Survey of Software Project Managers on Software Process Change** 263
Yuepu Guo, Carolyn B. Seaman (*University of Maryland, Baltimore County*)

Short Papers

Session 1C: Evaluation and Comparison of Techniques and Models

Session Chair: M. Genero (*University of Castilla La Mancha*)

- **A Pilot Study of Comparative Customer Comprehension between Extreme X-Machine and UML Models** 270
Christopher Thomson, Mike Holcome, Tony Cowling, Tony Simons, George Michaelides (*University of Sheffield*)
- **Enhancing Predictive Models Using Principal Component Analysis and Search Based Metric Selection: A Comparative Study** 273
Rodrigo Vivanco, Dean Jin (*University of Manitoba*)
- **Evaluating the Usefulness of Software Visualization in Supporting Software Comprehension Activities** 276
Glauco de F. Carneiro, Rodrigo Magnavita, Eduardo Spinola, Fábio Spinola, Manoel Mendonça (*Salvador University*)
- **A Hybrid Faulty Module Prediction Using Association Rule Mining and Logistic Regression Analysis** 279
Yasutaka Kamei, Akito Monden, Shuji Morisaki, Ken-ichi Matsumoto (*Nara Institute of Science and Technology*)
- **Mining Software Code Repositories and Bug Databases using Survival Analysis Models** 282
Michael Wedel (*Universität Stuttgart*), Uwe Jensen (*University of Hohenheim*), Peter Göhner (*Universität Stuttgart*)
- **Adding Planned Design to XP Might Help Novices' Productivity (or Might Not): Two Controlled Experiments** 285
René Noël, Gonzalo Valdes, Marcello Visconti, Hernán Astudillo (*Universidad Técnica Federico Santa María*)

Session 2C: Empirical Studies of Processes and Products

Session Chair: D. Winkler (*Vienna University of Technology*)

- **Using Students as Subjects — An Empirical Evaluation** 288
Mikael Svahnberg (*Blekinge Institute of Technology*), Aybülke Aarum (*University of New South Wales*), Claes Wohlin (*Blekinge Institute of Technology*)
- **Empirical Study of How Personality, Team Processes and Task Characteristics Relate to Satisfaction and Software Quality** 291
Silvia T. Acuña (*Universidad Autónoma de Madrid*), Marta N. Gómez (*Universidad San Pablo-CEU*), Juan de Lara (*Universidad Autónoma de Madrid*)
- **Empirical Evaluation of Analogy-X for Software Cost Estimation** 294
Jacky Keung (*NICTA Ltd. and University of New South Wales*)
- **Improving Application and Understanding of Experience Packages through Learning Spaces** 297
Eric Ras (*Fraunhofer IESE*)
- **Does the Use of Stereotypes Improve the Comprehension of UML Sequence Diagrams?** 300
Marcela Genero, José A. Cruz-Lemus (*University of Castilla-La Mancha*), Danilo Caivano (*University of Bari*), Sílvia Abrahão, Emilio Infrán, José A. Carsi (*Universidad Politécnica de Valencia*)
- **Web Application Fault Classification — An Exploratory Study** 303
Yuepu Guo, Sreedevi Sampath (*University of Maryland, Baltimore County*)

Session 3C: Development of Predictive Models

Session Chair: O. Dieste (*Universidad Politécnica de Madrid*)

- **Exposure Model for Prediction of Number of Customer Reported Defects** 306
Keld Raaschou (*SimCorp A/S*), Austen W. Rainer (*University of Hertfordshire*)
- **Analysis of the Reliability of a Subset of Change Metrics for Defect Prediction** 309
Raimund Moser (*Free University of Bolzano-Bozen*), Witold Pedrycz (*University of Alberta*), Giancarlo Succi (*Free University of Bolzano-Bozen*)
- **An Over-sampling Method for Analogy-based Software Effort Estimation** 312
Yasutaka Kamei (*Nara Institute of Science and Technology*), Jacky Keung (*National ICT Australia Ltd.*), Akito Monden, Ken-ichi Matsumoto (*Nara Institute of Science and Technology*)
- **An Empirical Model to Predict Security Vulnerabilities using Code Complexity Metrics** 315
Yonghee Shin, Laurie Williams (*North Carolina State University*)
- **Ensemble of Software Defect Predictors: A Case Study** 318
Ayse Tosun, Burak Turhan, Ayse Bener (*Bogazici University*)
- **Managing Software Quality through a Hybrid Defect Content and Effectiveness Model** 321
Michael Klás, Frank Elberzhager (*Fraunhofer Institute for Experimental Software Engineering*), Haruka Nakao (*Japan Manned Space Systems Corporation*)

Session 4C: Experience in Process Improvement

Session Chair: T. Gorschek (*BTH*)

- **Surveying Model Based Testing Approaches Characterization Attributes** 324
Arlito Claudio Dias Neto, Guilherme Horta Travassos (*Federal University of Rio de Janeiro*)
- **Statistical Process Control for Software: A Systematic Approach** 327
Nicola Boffoli, Giovanni Bruno, Danilo Caivano, Gemma Mastelloni (*University of Bari*)
- **Issues and Effort in Integrating Data from Heterogeneous Software Repositories and Corporate Databases** 330
Rudolf Ramler, Klaus Wolfmeier (*Software Competence Center Hagenberg*)

- **A Defect-Driven Process for Software Quality Improvement**.....333
Brian Robinson, Patrick Francis, Fredrik Ekdahl (*ABB Robotics*)
- **IMPS: An Experimentation Based Investigation of a Nationwide Software Development Reference Model**.....336
Marcos Kalinowski (*COPPE/UFRJ*), Kival Chaves Weber (*SOFTEX*), Guilherme Horta Travassos (*COPPE/UFRJ*)
- **Using the ProdFLOW™ Approach to Address the Myth of Productivity in R&D Organizations**339
Melanie Rube (*Siemens AG, CT SE3*), Stefan Wagner (*Technical University of Munich*)

Session 5C: Empirical Evidence and Systematic Review

Session Chair: D. Caivano (*University of Bari*)

- **A Mapping Study on Empirical Evidence related to the Models and Forms used in the UML**342
Rialette Pretorius, David Budgen (*Durham University*)
- **Software Process Simulation over the Past Decade: Trends Discovery from A Systematic Review**345
He Zhang (*National ICT Australia*), Barbara Kitchenham (*Keele University*), Dietmar Pfahl (*University of Oslo*)
- **An Empirical Investigation of Scenarios Gained and Lost in Architecture Evaluation Meetings**.....348
Dietmar Winkler, Stefan Biffl (*Vienna University of Technology*), Muhammad Ali Babar (*Lero, University of Limerick*)
- **Are Good Code Reviewers Also Good at Design Review?**.....351
Hideake Uwano, Akito Monden, Ken-ichi Matsumoto (*Nara Institute of Science and Technology*)
- **Understandability Measurement in an Early Usability Evaluation for Model-Driven Development: An Empirical Study**354
Jose Ignacio Panach, Nelly Condori-Fernández, Francisco Valverde, Nathalie Aquino, Óscar Pastor (*Universidad Politécnica de Valencia*)

Posters

- **Automatic Extraction of the Main Terminology used in Empirical Software Engineering through Text Mining Techniques**.....357
Francisco P. Romero, José A. Olivás, Marcela Genero, Mario Piattini (*University of Castilla La Mancha*)
- **Exploring Effort Distribution in RUP Projects**359
Werner Heijstek, Michel R.V. Chaudron (*Leiden University*)
- **Fit Data Selection for Software Effort Estimation Models**.....360
Koji Toda, Akito Monden, Ken-ichi Matsumoto (*NARA Institute of Science and Technology*)

- Organization**.....362

ESEM 2008 Organization

General Chair:

Dieter Rombach - University of Kaiserslautern and Fraunhofer Institute for Experimental Software Engineering (Germany)

Program Chairs:

Sebastian Elbaum - University of Nebraska-Lincoln (USA)

Jürgen Münch - Fraunhofer Institute for Experimental Software Engineering (Germany)

Steering Committee:

Natalia Juristo - Universidad Politécnica de Madrid (Spain)

Sebastian Elbaum - University of Nebraska - Lincoln (USA)

James Miller - University of Alberta (Canada)

Jürgen Münch - Fraunhofer IESE (Germany)

Dieter Rombach - University of Kaiserslautern and Fraunhofer IESE (Germany)

Carolyn Seaman - University of Maryland at Baltimore County (USA)

Rick Selby - Northrop Grumman (USA)

Sira Vegas - Universidad Politécnica de Madrid (Spain)

Laurie Williams - North Carolina State University (USA)

Program Committee:

Silvia Abraham - Universidad Politecnica de Valencia (Spain)

Muhammad Ali Barbar - Lero, University of Limerick (Ireland)

Anneliese Amschler Andrews - University of Denver (USA)

James Andrews - University of Western Ontario (Canada)

Stefan Biffl - TU Vienna (Austria)

Lionel Briand - Simula Research Laboratory (Norway)

Luigi Buglione - Engineering.IT (Italy)

Gerardo Canfora - University of Sannio (Italy)

Jeff Carver - University of Alabama (USA)

Marcus Ciolkowski - University of Kaiserslautern (Germany)

Reidar Conradi - Norwegian University of Science and Technology (Norway)

Tore Dyba - Sintef (Norway)

Christof Ebert - Alcatel (France)

Khaled El Emam - University of Ottawa (Canada)

Hakan Erdogmus - NRC Institute for Information Technology (Canada)

Marcela Genero - University of Castilla La Mancha (Spain)

Tracy Hall - University of Hertfordshire (UK)

Lorin Hochstein - University of Nebraska-Lincoln (USA)

Martin Höst - Lund University (Sweden)

Frank Houdek - Daimlerchrysler (Germany)

Andreas Jedlitschka - Fraunhofer IESE (Germany)

Ross Jeffery - University of New South Wales (Australia)

Philip Johnson - University of Hawaii (USA)

Magne Jorgensen - Simula Research Laboratory (Norway)

Natalia Juristo - Universidad Politecnica de Madrid (Spain)

Shinji Kusumoto - Osaka University (Japan)

Program Committee
(Continued):

Filippo Lanubile - University of Bari (Italy)
Jose Carlos Maldonado - Universidade de Sao Paulo (Brazil)
Kenichi Matsumoto - Nara Institute of Science and Technology (Japan)
Emilia Mendes - University of Auckland (New Zealand)
James Miller - University of Alberta (Canada)
Audris Mockus - Avaya Labs Research (USA)
Sandro Morasca - University degli Studi dell'Insubria (Italy)
Maurizio Morisio - Politecnico di Torino (Italy)
Nachi Nagappan - Microsoft (USA)
Markku Oivo - University of Oulu (Finland)
Dietmar Pfahl - Simula Research Laboratory & University of Oslo (Norway)
Mario Piattini - University of Castilla La Mancha (Spain)
Marc Roper - University of Strathclyde (UK)
Guenther Ruhe - University of Calgary (Canada)
Per Runeson - Lund University (Sweden)
Sreedevi Sampath - University of Maryland, Baltimore (USA)
Carolyn Seaman - University of Maryland, Baltimore (USA)
Rick Selby - Northrop Grumman (USA)
Martin Shepperd - Brunel University (UK)
Forrest Shull - Fraunhofer Center, Maryland (USA)
Janice Singer - National Research Council (Canada)
Harvey Siy - University of Nebraska (USA)
Dag Sjoberg - Simula Research Laboratory (Norway)
Arie Van Deursen - Delft University of Technology (Netherlands)
Sira Vegas - Universidad Polit cnica de Madrid (Spain)
June Verner - NICTA (Australia)
Laurie Williams - North Carolina State University (USA)
Claes Wohlin - Blekinge Institute of Technology (Sweden)

Finance Chair: Andreas Jedlitschka - Fraunhofer Institute for Experimental Software Engineering (Germany)

Local Arrangements Chair: Marcus Ciolkowski - Fraunhofer Institute for Experimental Software Engineering (Germany)

Publicity Chair: James Miller - University of Alberta (Canada)

Proceedings Chair: Marcus Ciolkowski - Fraunhofer Institute for Experimental Software Engineering (Germany)

Short Papers Chair: Teresa Baldassarre - Universita' degli Studi di Bari (Italy)

Poster Chair: Haruka Nakao - Japan Manned Space Systems Cooperation (Japan)

Automatic Extraction of the Main Terminology used in Empirical Software Engineering through Text Mining Techniques

Francisco P. Romero, José A. Olivas
University of Castilla-La Mancha
Ciudad Real, Spain
{FranciscoP.Romero, JoseAngel.Olivas}@uclm.es

Marcela Genero, Mario Piattini
University of Castilla-La Mancha
Ciudad Real, Spain
{Marcela.Genero, Mario.Piattini}@uclm.es

ABSTRACT

The need for an explicit common terminology within Empirical Software Engineering (an ESE-Glossary of terms) was highlighted in the ISERN 2007 meeting [2]. The goal was to define a glossary of terms related to ESE based on an initial glossary published in <http://lens-ese.cos.ufrj.br/wikiese>. This initial glossary was built manually, based on expert knowledge. However, owing to the dynamic nature of the research works in ESE, this glossary must be dynamically updated with information extracted from the relevant documents in the research domain. Automation is, therefore, mandatory. We propose a text mining technique for the automatic extraction of the most relevant terms used in ESE documents. Our technique also provides the relationships between terms, with the degree of affinity between them. Our approach could, therefore, be useful in the improvement of the initial glossary of terms and in discovering relationships between terms.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *Linguistic processing*.

D.2.0 [Software Engineering]: General. Standards.

General Terms

Documentation, Standardization.

1. INTRODUCTION

In this work the thirteenth volume of Empirical Software Engineering (EMSE) Journal [1] is used as a source of documents. In this preliminary work, it was decided to limit the number of reviewed documents to eleven, as this would be a sufficient quantity with which to carry out a preliminary study. These papers have been randomly selected. This is an ongoing work whose eventual purpose is to build an ESE Ontology.

2. FINDING RELEVANT TERMS IN ESE

We shall now describe the three main steps of which our text mining technique is composed. The result of this process is a hierarchical classification of terms.

1) Documents Preprocessing. It is first necessary to carry out a linguistic preprocess of the documents. The pre-processing scheme employs similar techniques, such as stemming and stop word removal, which are popularly used in the text mining community to extract terms for vector-space representation. The

extraction process takes into account the presence of multiple-word terms. Abstract, references and acknowledgments are excluded since they contain redundant information. Finally, an index with these multi-word terms is built. This index contains the frequency of each term in each document processed.

2) Term Relevance. The automatic key terms extraction is performed through term ranking and by using term weighting measures. The frequency of the terms represents a good measure of the relevance in the domain. We propose the following extension of the classic tf-idf formula to calculate the relevance weight of a term i :

$$w_i = \frac{\text{docs}(i) + \sum_{j \in D} \frac{tf(i,j) \cdot k}{|j|}}{|D|}$$

Thus, the relevance weight of a term i (w_i) is calculated by using the number of documents in which the term i appears ($\text{docs}(i)$), and its frequency in each document j ($tf(i,j)$). The normalization of the value is realized by using the number of processed documents ($|D|$). The formula also contains a multiplicity factor k ($k = 1.000$).

Table 1 shows the top ten individual and pairs of terms extracted from the afore mentioned EMSE Journal documents.

Table 1. Top ten of terms and pair of terms

Terms	Weight	Couple of terms	Weight
software	23.93	<i>software engineering</i>	9.51
<i>study</i>	16.93	empirical study	2.82
data	13.55	controlled experiment	1.82
engineering	13.18	<i>research interest</i>	1.74
<i>project</i>	12.40	<i>experiment design</i>	1.72
result	11.14	independent variable	1.63
experiment	10.27	<i>software development</i>	1.58
model	9.56	dependent variable	1.44
software eng.	9.51	research question	1.34
<i>test</i>	9.04	external validity	1.24

The terms in italics, which we consider to be relevant terms in the ESE field, do not appear in the published glossary. For that reason, our approach could, therefore, enrich the existent glossary.

3) Term Relationships. This can be used to build a hierarchical classification of terms using the constructed index. An ontology, in this context, may be considered as a set of related trees in which each node represents a term.

In these trees, node *A* is a descendant of node *B* if the term *A* is “more concrete than” the term *B*, or if the term *B* is “more general than” the term *A*.

The fuzzy measure defined in [3] is used to obtain the generality degree (*GD*) between each pair of words contained in the document collection.

$$GD(A, B) = \frac{\text{Number of co- occurrences of A and B in the same subsection}}{\text{Number of occurrences of B in the subsections}}$$

The T-DiCoR tool (Three Dimensional Conceptual Representation) could be used to visualize this hierarchical classifications of terms.

3. FUTURE WORK

In the near future, we shall expand the application of the text mining technique proposed in this paper to a greater volume of EMSE Journal documents. Publications in the ISESE and ESEM conferences should, moreover, be explored. Later, we shall obtain the relationships between terms and we go on build an ESE ontology.

4. REFERENCES

- [1] Basili, V., Briand L. (eds). 2008 Empirical Software Engineering Journal. Vol 13. Springer.
- [2] Travassos, G., Barker, M. 2007, Experimental Software Engineering Glossary of Terms, ISERN 2007. Widyanoro D., Yen J. 2001 Incorporating fuzzy ontology of term relations in a search engine, Proceedings of the BISC Int. Workshop on Fuzzy Logic and the Internet, (Berkeley, USA), FLINT’01. 155-160.

Additional reviewers:

Martin Auer
Maria Teresa Baldassarre
Fabio Calefato
Marcio Delamaro
Oscar Dieste
Evgenia Egorova
Mohamed El-Attar
Fabiano Ferrari
Michael Gegick
Domenico Gendarmi
Stein Grimstad
Anna Grimán
Lauren Hayward
Sarah Smith Heckman
Jens Heidrich
Toan Huynh
Vigdis By Kampenes
Jacky Keung

Lucas Layman
Otávio Lemos
Jingyue Li
Teresa Mallardo
Rajwinder Kaur Panesar-Walawege
Francisco Pino
Daniel N. Port
Yonghee Shin
Ben Smith
Martin Soto
Tor Stålhane
Marco Torchiano
Adam Trendowicz
Irfan Ullah
Giuseppe Visaggio
Gursimran Walia
Byron Williams

Sponsors:



Supporters:

SIEMENS

Invented for life



BOSCH

Author Index

Abrahão, Silvia	300	Göhner, Peter	282	Nakao, Haruka	321
Acuña, Silvia T.	291	Gómez, Marta N.	291	Nguyen, Vu	213
Ali Babar, Muhammad	80	Guo, Yuepu	149, 263, 303	Noël, René	285
Angelis, Lefteris	70	He, Mei	61	Nugroho, Ariadi	90
Aquino, Nathalie	354	Heckman, Sarah	41	Öhman, Peter	12
Astudillo, Hernán	285	Heidrich, Jens	243	Olivas, José A.	357
Aurum, Aybüke	288	Hejstek, Werner	359	Panach, Jose Ignacio	354
Babar, Muhammad Ali	348	Herbsleb, James D.	2	Pastor, Óscar	354
Begel, Andrew	120	Hoeningger, Harald	1	Pedrycz, Witold	309
Bener, Ayse	318	Holcome, Mike	270	Pernstål, Joakim	12, 233
Biffi, Stefan	348	Insfrán, Emilio	300	Petersen, Kai	139
Boehm, Barry	61, 213	Jensen, Uwe	282	Pfahl, Dietmar	345
Boffoli, Nicola	327	Ji, Junzhong	168	Piattini, Mario	357
Bruno, Giovanni	327	Jin, Dean	273	Port, Dan	51
Buckley, Jim	80	Kajko-Mattsson, Mira	111	Pretorius, Rialette	342
Budgen, David	342	Kalinowski, Marcos	336	Raaschou, Keld	306
Caivano, Danilo	300, 327	Kamei, Yasutaka	279, 312	Radicke, Mathias	243
Carley, Kathleen M.	2	Keung, Jacky	294, 312	Rainer, Austen W.	306
Carneiro, Glaucio de F.	276	Kitchenham, Barbara	345	Ramler, Rudolf	330
Carsi, José A.	300	Klås, Michael	321	Ras, Eric	297
Carver, Jeffrey C.	158	Korte, Marcel	51	Regardie, Myrna	149
Cataldo, Marcelo	2	Kudrjavets, Gunnar	206	Robinson, Brian	333
Chaudron, Michel R. V.	90, 359	Lavazza, Luigi A.	100	Romero, Francisco P.	357
Chen, Weibing	168	Layman, Lucas	206	Rönkkö, Kari	139
Ciolkowski, Marcus	243	Le Gear, Andrew	80	Rosik, Jacek	80
Condori-Fernández, Nelly	354	Li, Jingyue	168	Ruhe, Melanie	339
Conradi, Reidar	168	Li, Mingshu	61, 223, 253	Runeson, Per	22
Cowling, Tony	270	Li, Qi	223	Sampath, Sreedevi	303
Cruz-Lemus, José A.	300	Liu, Chunnian	168	Scott, Hanna	32
de Lara, Juan	291	Ma, Jianqiang	168	Seaman, Carolyn	129, 149, 263
del Bianco, Vieri	100	Magazinovic, Ana	12, 233	Shaw, Mary	110
Dias Neto, Arilo Claudio	324	Magnavita, Rodrigo	276	Shin, Yonghee	315
Dingsøy, Torgeir	178	Mastelloni, Gemma	327	Shull, Forrest	149
Du, Jing	253	Matsumoto, Ken-ichi	279, 312, 351, 360	Simon, Frank	243
Dybå, Tore	178	Mendonça, Manoel	276	Simons, Tony	270
Ekdahl, Fredrik	333	Michaelides, George	270	Skoglund, Mats	22
Elbert, Denis	149	Mittas, Nikolaos	70	Spinola, Eduardo	276
Elberzhager, Frank	321	Monden, Akito	279, 312, 351, 360	Spinola, Fábio	276
Engström, Emelie	22	Morasca, Sandro	188	Steece, Bert	213
Feldmann, Raimund L.	149	Morisaki, Shuji	279	Succi, Giancarlo	309
Francis, Patrick	333	Moser, Raimund	309	Svarnberg, Mikael	288
Garavaglia, Carla	100	Müller, Mark	1	Thomson, Christopher	270
Genero, Marcela	300, 357	Nagappan, Nachiappan	120, 206	Toda, Koji	360
Godfrey, Sally	149			Tosun, Ayse	318