Mining Software Repositories

What is MSR?

- Mining Software Repositories (MSR) uses data available in repositories to support development activities
  - For example, defect assignment, software validation, evolution and planning
- Increased amount of data available in software archives through large open source projects
- Process applies data extraction and analysis to make decisions and predictions

Repositories

- Three repository types are the focus of research activity:
  - Bug Repositories (e.g. Bugzilla)
  - Source Repositories (e.g. CVS, SVN)
  - Communication Repositories (e.g. mailing lists)

Research focus areas

- Bug repositories
  - Defect triage
  - Estimate hours for a fix to plan schedules
  - Report quality
- Source Repositories
  - Code clones
  - Code evolution and project growth
  - Bug detection
- Communication Repositories
  - Mailing list interaction

Process

- Extract data from the repository (e.g. Eclipse Mylyn demo)
- Generally apply a machine learning algorithm or data mining algorithm to the data (e.g. WEKA demo)
  - Typical strategies:
    - Supervised learning
    - Unsupervised learning (e.g. kNN)

Data Analysis

- In analyzing the effectiveness of the application, typically use precision and recall
- Precision
  - How well provide concise recommendations [Yin04]
- Recall
  - How well recommend items in the correct solution [Yin04]

Future direction

- Increased support through tooling
- Ease of use for the end user

Resources

[Anv06] Anvik, John, Lyndon Hiew, and Gail C. Murphy. "Who should Fix this Bug?". *28th International Conference on Software Engineering 2006, ICSE '06, May 20, 2006 - May 28.* 2006, Shanghai, China: Inst. of Elec. and Elec. Eng. Computer Society , 2006. 361-370. Print.

[Anv07] Anvik, John, and Gail C. Murphy. "Determining Implementation Expertise from Bug Reports". ICSE 2007 Workshops: Fourth International Workshop on Mining Software Repositories, MSR 2007, May 20, 2007 - May 26. 2007, Minneapolis, MN, United states: Inst. of Elec. and Elec. Eng. Computer Society , 2007. IEEE Computer Society; SIGSOFT. Print.

[Bet08] Bettenburg, N., et al. "What Makes a Good Bug Report?".November 09 - 14, 2008, Atlanta, Georgia. New York, NY: ACM , 2008. 308-318. Print.

[Can06] Canfora, Gerardo, and Luigi Cerulo. "Supporting Change Request Assignment in Open Source Development". *2006 ACM Symposium on Applied Computing, April 23, 2006 - April 27.* 2006, Dijon, France: Association for Computing Machinery , 2006. 1767-1772. Print.

[Cub04] Cubranic, Davor, and Gail C. Murphy. "Automatic Bug Triage using Text Categorization." Proceedings of the Sixteenth International Conference on Software Engineering and Knowledge Engineering (2004)Print.

[Fis03] Fischer, Michael, Martin Pinzger, and Harald Gall. "Populating a Release History Database from Version Control and Bug Tracking Systems". *International Conference on Software Maintenance, September 22, 2003 - September 26.* 2003, Amsterdam, Netherlands: IEEE Computer Society , 2003. 23-32. Print.

[Go4] Godfrey, M., et al. "Four Interesting Ways in which History can Teach Us about Software". *"International Workshop on Mining Software Repositories (MSR 2004)" W17S Workshop - 26th International Conference on Software Engineering.* 25 May 2004, Stevenage, UK: IEE , 2004. 58-62. Print.

[Go09] Godfrey, Michael W., et al. "Future of Mining Software Archives: A Roundtable." *IEEE Software* 26.1 (2009): 67-70. Print.

[Has06] Hassan, Ahmed E. "Mining Software Repositories to Assist Developers and Support Managers". *ICSM 2006: 22nd IEEE International Conference on Software Maintenance, September 24, 2006 - September 27.* 2006, Philadelphia, PA, United states: IEEE Computer Society , 2006. 339-342. Print.

[Has08] Hassan, Ahmed E. "The Road Ahead for Mining Software Repositories". *16th Frontiers of Software Maintenance, FoSM 2008, September 30, 2008 - October 2.* 2008, Beijing, China: Inst. of Elec. and Elec. Eng. Computer Society , 2008. 48-57. Print.

[Has08-1] Hassan, Ahmed E., and Tao Xie. "Mining Software Engineering Data." 2008.Web. <http://ase.csc.ncsu.edu/dmse/dmse-icse08-tutorial.pdf>.

[Nag09] Nagappan, Zeller, and Thomas Zimmermann. "Guest Editors' Introduction: Mining Software Archives." *IEEE Software* 26.1 (2009): 24-5. Print.

Resources continued

[Pan07] Panjer, Lucas D. "Predicting Eclipse Bug Lifetimes". *ICSE 2007 Workshops: Fourth International Workshop on Mining Software Repositories, MSR 2007, May 20, 2007 - May 26.* 2007, Minneapolis, MN, United states: Inst. of Elec. and Elec. Eng. Computer Society , 2007. IEEE Computer Society; SIGSOFT. Print.

[Yin04] Ying, Annie T. T., et al. "Predicting Source Code Changes by Mining Change History." *IEEE Transactions on Software Engineering* 30.9 (2004): 574-86. Print.

[Wei07] Weiß, C., et al. "How Long Will it Take to Fix this Bug?".2007. Print.

"IEEE Working Conference on Mining Software Repositories."Web. <http://www.msrconf.org>.

"Eclipse Mylyn."Web. <http://www.eclipse.org/mylyn>.

"Weka."Web. <http://www.cs.waikato.ac.nz/ml/weka/>.