On the Equivalence of Information Retrieval Methods for Automated Traceability Link Recovery: A Ten-Year Retrospective

Rocco Oliveto
rocco.oliveto@unimol.it
University of Molise
Pesche (IS), Italy

Denys Poshyvanyk
denys@cs.wm.edu
The College of William and Mary
Williamsburg, VA, USA

Malcom Gethers*
mbgethers@wm.edu
The College of William and Mary
Williamsburg, VA, USA

Andrea De Lucia
adelucia@unisa.it
University of Salerno
Fisciano (SA), Italy

ABSTRACT

At ICPC 2010 we presented an empirical study to statistically analyze the equivalence of several traceability recovery methods based on Information Retrieval (IR) techniques [1]. We experimented the Vector Space Model (VSM) [2], Latent Semantic Indexing (LSI) [3], the Jensen-Shannon (JS) method [4], and Latent Dirichlet Allocation (LDA) [5]. Unlike previous empirical studies we did not compare the different IR based traceability recovery methods only using the usual precision and recall metrics. We introduced some metrics to analyze the overlap of the set of candidate links recovered by each method. We also based our analysis on Principal Component Analysis (PCA) to analyze the orthogonality of the experimented methods. The results showed that while the accuracy of LDA was lower than previously used methods, LDA was able to capture some information missed by the other exploited IR methods. Instead, JS, VSM, and LSI were almost equivalent. This paved the way to possible integration of IR based traceability recovery methods [6].

Our paper was one of the first papers experimenting LDA for traceability recovery. Also, the overlap metrics and PCA have been used later to compare and possibly integrate different recommendation approaches not only for traceability recovery, but also for other reverse engineering and software maintenance tasks, such as code smell detection, design pattern detection, and bug prediction.

KEYWORDS

Traceability, Information Retrieval Methods, Evaluation metrics

ACM Reference Format:
Rocco Oliveto, Malcom Gethers, Denys Poshyvanyk, and Andrea De Lucia. 2020. On the Equivalence of Information Retrieval Methods for Automated Traceability Link Recovery: A Ten-Year Retrospective. In *28th International Conference on Program Comprehension (ICPC '20), October 5–6, 2020, Seoul, Republic of Korea.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3387904.3394491

*Former student of The College of William and Mary.

ICPC '20, October 5–6, 2020, Seoul, Republic of Korea

ACM ISBN 978-1-4503-7958-8/20/05.
https://doi.org/10.1145/3387904.3394491

ABOUT THE AUTHORS

Rocco Oliveto is associate professor at University of Molise (Italy), where he is also the founder of the Software and Knowledge Engineering Lab (STAKE Lab). He is co-founder and CEO of datasound—a spin-off of the University of Molise—aiming at producing innovative recommender systems. More information available at: https://dibt.unimol.it/staff/oliveto/.

Malcom Gethers received the PhD degree in computer science from the College of William and Mary in 2012. He was an assistant professor in the Department of Information Systems at the University of Maryland, Baltimore County (UMBC).

Denys Poshyvanyk is the class of 1953 term distinguished associate professor of computer science with the College of William and Mary, in Virginia. He currently serves on the editorial board of the IEEE Transactions on Software Engineering, the Empirical Software Engineering Journal and the Journal of Software: Evolution and Process. He is a recipient of the NSF CAREER award (2013). More information available at: http://www.cs.wm.edu/ denys/.

Andrea De Lucia is full professor and head of the software engineering lab in the Department of Computer Science, University of Salerno, Italy. He is the director of the International Summer School on Software Engineering and the co-editor in chief of Science of Computer Programming. More information available at: https://docenti.unisa.it/003241/home.

REFERENCES

[1] R. Oliveto, M. Gethers, D. Poshyvanyk, A. De Lucia, "On the Equivalence of Information Retrieval Methods for Automated Traceability Link Recovery", *in Proc. of the International Conference on Program Comprehension*, pp. 68-71, 2010.
[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.
[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, vol. 41, n. 16, pp. 391?407, 1990.
[4] A. Abadi, M. Nisenson, and Y. Simionovici, "A Traceability Technique for Specifications", *in Proc. of the International Conference on Program Comprehension*, pp. 103-112, 2008.
[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation", *The Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
[6] M. Gethers, R. Oliveto, D. Poshyvanyk, A. De Lucia, "On Integrating Orthogonal Information Retrieval Methods to Improve Traceability Recovery", *in Proc. of the International Conference on Software Maintenance*, pp. 133-142, 2011.