

Scalable Software Code Representation, Search and Classification

AU\$50K / Queensland University of Technology(Geva) / Labs(Cifuentes) / #1396

Proposal Details

1. Basic Details:

- **CR ID:** 1396
- **Project Title:** Scalable Software Code Representation, Search and Classification
- **Current Status:** Completed; last updated on 16-SEP-16
- **Is this a project renewal?** No
- **Expected Start Date:** 01-OCT-15
- **Project Duration:** 9 month/s
- **Funding Request Amount:** 50K
- **Funding Source:** Labs
- **Funding Type:** Gift
- **Institution:** [Queensland University of Technology](#)
- **Executive summary and value to Oracle:**

The goal of this research is to investigate the effectiveness of document signature approaches for source code classification tasks. We want to determine whether document signature techniques can be used to create a representation of source code that preserves semantic similarities in a way that can be used to quickly retrieve potentially matching source code segments. If these techniques work, the signatures can be used in conjunction with highly efficient approximate retrieval methods in order to classify source code segments based on an existing database of classified source code segments.

We believe that this research will be of value to Oracle due to the high processing requirements of existing source code analysis tools. If signature classification is at least effective enough to reduce the amount of code that needs to be more rigorously analyzed by traditional accurate but computationally more expensive methods, then adopting this approach will produce substantial performance dividends for source code classification.

- **Research Summary:**

The goal of this research is to investigate the effectiveness of document signature approaches for source code classification tasks. We want to determine whether document signature techniques can be used to create a representation of source code that preserves semantic similarities in a way that can be used to quickly retrieve potentially matching source code segments. If these techniques work, the signatures can be used in conjunction with highly efficient approximate retrieval methods in order to classify source code segments based on an existing database of classified source code segments.

We believe that this research will be of value to Oracle due to the high processing requirements of existing source code analysis tools. If signature classification is at least effective enough to reduce the amount of code that needs to be more rigorously analyzed by traditional accurate but computationally more expensive methods, then adopting this approach will produce substantial performance dividends for source code classification.

- **Prior Work:**

The research done as part of this collaboration builds on existing research into generating compact representations of searchable inputs that maintains semantic relationships as documented by Sahlgren [2] and Geva [3]. These representations are in the form of binary 'signatures' obtained through locality sensitive hashing (many variations exist). The signatures are generally much more compact and not human readable, but can efficiently be compared to each other for similarity using the Hamming distance. While these locality-sensitive hashes do not provide an exact representation of the original input, they are often a sufficiently accurate approximation for many information retrieval, clustering and classification tasks.

While signature representations of source code are a currently unexplored facet of this research, Kuhn [1] found success with the related approach of Latent Semantic Indexing for source code classification in conjunction with a bag-of-words approach. We seek to significantly extend this approach and apply it to the much richer, yet complex, space of code representation at multiple levels, and with the incorporation of metadata and other knowledge about code structure.

One of the main reasons for pursuing a signature representation of source code segments is that it enables the use of high performance approaches that take advantage of the semantic clustering of these signatures to retrieve approximate nearest neighbours. This allows even data sets of many millions of signatures to be searched within milliseconds on a single machine. We were able to index and cluster a collection of about a billion HTML web documents (Clueweb-2012) in a matter of hours, and search it in a matter of milliseconds. We expect this to be particularly useful when programmatic search and classification operations are built into code analysis and may require the

execution of possibly thousands of searches.

References:

1. Kuhn, Adrian, Stéphane Ducasse, and Tudor Gírba. "Semantic clustering: Identifying topics in source code." *Information and Software Technology* 49.3 (2007): 230-243.
2. Sahlgren, Magnus. "An introduction to random indexing." *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*. Vol. 5. 2005.
3. Geva, Shlomo, and Christopher M. De Vries. "Topsig: Topology preserving document signatures." *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011.
4. Chappell, Timothy, Shlomo Geva, and Guido Zuccon. "Approximate Nearest-Neighbour Search with Inverted Signature Slice Lists." *Advances in Information Retrieval*. Springer International Publishing, 2015. 147-158.
5. De Vries, C.M., De Vine, L., Geva, S., Nayak, R.: Parallel Streaming Signature EM-tree: A Clustering Algorithm for Web Scale Applications. In: *Proceedings of the 24th International Conference on World Wide Web Conference (2015)*

2. Oracle Principal Investigators:

- [Cristina Cifuentes \(cristina.cifuentes@oracle.com\)](mailto:cristina.cifuentes@oracle.com)

3. External Principal Investigators:

- Shlomo Geva (s.geva@qut.edu.au)

Other External Team Members:

- Funds management:
 - Kym Mansfield (kym.mansfield@qut.edu.au)
- Banking contacts:
 - Russell Mallet (r.mallet@qut.edu.au)

4. Expected Value to Oracle:

- **Business Benefit:**

The benefit of this research is to explore the suitability of machine learning and information retrieval techniques as they apply to the analysis of source code. This is an area that has not been researched in the literature.

Should the results of the research be positive, the research would require further development to be incorporated into a tool that internal Oracle organisations could use, to find bugs and/or vulnerabilities in their source code, as a way to complement static analysis tool already in use. The approach has the potential to be more efficient than existing tools.

- **Source Code Useability:**

Code would be useful only as a research prototype.

- **Group Contacts:**

- **Internships:** Yes

- **Publications:** Yes

5. IP and License Considerations:

- **Proposed Outbound IP/License (what Oracle provides to do the research):** None
- **Proposed Inbound IP/License (what Oracle receives to do the research):** None
- **Proposed Resulting IP/License (ownership and rights of the research outcome):**

Source code to be released under the UPL or BSD license.

- **Unfiled or Pending Patents:** None

- **Filed Patents:** None

- **Rationale:**

The work will be done at the university based on earlier research and software artifacts that the researchers have developed. Availability of results will be made available during meetings. The final software artifact will be made available under UPL or BSD license, so that we can further explore with it. Other software that they reuse is available under open source, they need to ensure the licenses are compatible, and so cannot determine whether to use UPL or BSD until such time as the project starts.

6. Students:

Not Available

7. Benefit to Oracle:

- **Artifacts:** Not Yet Available
- **Patents/Disclosures:** Not Yet Available
- **Publications:** Not Yet Available
- **Hires:** Not Yet Available
- **External Synopsis:** Not Yet Available
- **Internal Synopsis:** Not Yet Available


8. Governance:

Not Yet Available

9. Summary of Benefit and Governance from previous projects:

Not Available

Documents

Document	Type	Uploaded 	Download	File Name
Approval trail	Administration	17-DEC-2015 13:39	Download	CR 1396 approval trail.pdf
Payment request	Administration	10-DEC-2015 06:08	Download	CR 1396 payment request QUT -updated.xlsx

Legal Compliance and Ethics

Approval Request for University Research Activity (JAPAC)

Executive Summary:

The goal of this research is to investigate the effectiveness of document signature approaches for source code classification tasks. We want to determine whether document signature techniques can be used to create a representation of source code that preserves semantic similarities in a way that can be used to quickly retrieve potentially matching source code segments. If these techniques work, the signatures can be used in conjunction with highly efficient approximate retrieval methods in order to classify source code segments based on an existing database of classified source code segments.

We believe that this research will be of value to Oracle due to the high processing requirements of existing source code analysis tools. If signature classification is at least effective enough to reduce the amount of code that needs to be more rigorously analyzed by traditional accurate but computationally more expensive methods, then adopting this approach will produce substantial performance dividends for source code classification.

1. Project Title: Scalable Software Code Representation, Search and Classification / CR #1396
2. Expected Start Date and Duration: 01-OCT-15, 9 months
3. External Research Organization:
 - Name of the External Research Organization: Queensland University of Technology
 - Public Organization
 - Complete Mailing Address: GPO Box 2434 Brisbane QLD 4001 Australia
 - Location (if different from mailing address): 2 George St Brisbane QLD 4000
 - Name of the External Principal Investigator (EPI): Shlomo Geva (s.geva@qut.edu.au)
4. Name of the Oracle Principal Investigator: Cristina Cifuentes (cristina.cifuentes@oracle.com, Labs, Research Director, Oracle Labs Australia)
5. Additional Oracle Investigators:
6. Project Background and Description:

The background to this project is Oracle's existing work on using source code analysis techniques on large bodies of program code in order to find bugs, vulnerabilities and other properties of interest. Due to the problems associated with standard analysis tools when it comes to handling very large projects (many millions of lines of code), as well as the fact that many common bugs/malware/etc tend to share various characteristic features with other instances of the same kind of bugs/malware/etc, it makes sense to draw on existing knowledge about various bugs and vulnerabilities that have been found when searching for new ones. If the incorporation of this existing knowledge can be applied automatically to reduce the search space, the scalability problems inherent with automated source code analysis could be ameliorated.

The goal of this research is to investigate whether certain information retrieval approaches that have been found to be effective for searching and classification tasks can be used in this space to speed up the process of finding suspect/interesting source code. If a certain segment of source code happens to contain features of interest, other segments of source code that are highly similar (e.g. in structure) may also be found to contain the same or similar features of interest. Approaches from machine learning that will be explored in this project include feature selection and engineering, dimensionality reduction through random projections, clustering and classification. The references list of this proposal contains several articles that are quite detailed with respect to the underlying approaches. This project will require the fusion of multiple feature types associated with code snippets to form a suitable representation. The open source software tools that we developed at QUT (TopSig, [k-tree](#) and [LMW-tree](#)) are scalable and

effective in storage, retrieval and classification of multi modal objects (e.g. DNA sequences, text, XML, images, and more.) An important outcome from this project will be a suitable efficient and effective representation of code snippets for storage, retrieval, clustering and classification.

Technical Objectives:

- Investigate different representations of source code in the signature space to find out which representations produce a better picture of the code for classification purposes. Potential representations include text-based representations, as used when representing documents for search purposes, as well as representations that make use of the code at different levels (for instance, after being compiled into assembler or IL), as well as features produced in pre-processing of source/compiled code. Any metadata about code which is available will also be investigated (e.g. any temporal or geospatial or cyber-coordinates of code, or code features generated by code analysis using existing tools – such as by the research team at Oracle Labs in Brisbane).
- Determine the appropriate features and approach required to classify these representations, based on inputs from a database of existing tagged code samples.
- Determine the granularity of code segmentation necessary to achieve the desired results – in our experience it is possible to get improved results by working with very fine granularity snippets rather than with whole (large) documents. This means splitting programs/code into smaller snippets and working with two to three orders of magnitude more objects in storage and retrieval - hence turning millions of programs into billions of code parts. Scaling to handle such big data collections is where our (QUT) expertise is.
- Develop prototype software for use in Oracle's existing workflow, based on the research performed in the course of meeting previous objectives, to incorporate use of the document signature approach and improve the computational efficiency of Oracle's existing processes.
- Testing of hybrid approaches developed during this project, in search, clustering and classification, against baseline state-of-the art approaches that are currently being developed and/or in use by Oracle Labs in Brisbane.

7. Who initiated this (is the University/Professor soliciting the support or is it Oracle initiated)? Professor

8. Is this a project renewal? No

If Yes, please comment on past research and researcher collaboration and contributions, including deliverables and publications:

9. Describe how collaboration and interactions will be managed by Oracle:

Fortnightly or monthly meetings with the Professor and his student.

Meetings to be held at the Oracle Labs Brisbane office.

10. Does this project involve any of the following (check all that may apply and provide exact name of entity to be paid):

- Membership in an Industrial Affiliates or similar program
- A gift, grant, donation or other payment: directly to a university division or department, Name: Queensland University of Technology, School of Electrical Engineering and Computer Science
- A gift, grant, donation or other payment: to a university foundation or other non-profit
- A gift, grant, donation or other payment: to a consortium, collaborative or other third party
- Hiring of a university professor or graduate student as a part-time employee
- Retaining a university professor or graduate student as a consultant
- Providing or loaning hardware or software to a university or any university professor or graduate student
- Other

11. Will the university, any university professor or student create any software or other intellectual property that Oracle may wish to use in the future? Yes

If Yes, please describe the nature of the software or intellectual property:

Software artifact.

12. Does Oracle have, or has Oracle had in the past, any business relationships with this University or any of the professors or graduate students involved in this project (include any past or pending research payments from Oracle to this University and any professors or graduate students from this University that worked for Oracle in the past three (3) years or are presently working for Oracle)? No

If Yes, please provide details about the relationship(s) and the Oracle LOB(s) involved:

13. Does/Do the professor(s) or graduate student(s) have any other ties to Oracle, such as a relative employed by Oracle or any ownership/interest in a company that may be a vendor to Oracle? No

If Yes, please provide details:

14. Are you aware of any pending Oracle business deals or decisions with this University? No

If Yes, please provide details:

15. Are you aware of any recently closed Oracle business deals or decisions with this University? No

If Yes, please provide details: