ELEC 875
Design Recovery and Automated Evolution

Week 5 Class 2
Clones
Scalable Clone Detection

- What Are Clones?
  - in simplest form, where a programmer has used copy paste during development
  - various levels of clones
  - identical clones (type 1)
  - similar clones (i.e. variable names/constants have been changed)(type 2)
  - similar clones (lines have been added, deleted)(type 3)
  - semantic clones (different code but same computation)
Scalable Clone Detection

- Some clones are harmful, some are not so harmful...
  - sometimes clones are done on purpose
  - origin of errors (minor vs conceptual)
  - reduce coupling
  - reduce testing
  - some clones never change (central data structure definitions)
  - design recovery tools should know about them
    - knowledge about system
    - Roy & Cordy
    - Godfrey et al.
Scalable Clone Detection

- Clone detection is hard problem
- Various approaches
  - character based (string based)
    - unix diff command
  - token based
    - allows for differences formatting
  - syntax
  - tree based, comparison of parsed trees
  - byte/machine code (execution)
    - generated code comparison
    - output execution
  - program dependency graphs...
Scalable Clone Detection

- Uses
  - copyright infringement...
  - system dependence....
- Units of clones
  - are individual statements clones?
  - sequences of statements clones?
  - blocks clones?
  - functions?
  - how much is similar before it is called a clone
Map Reduce

- Introduced by Google
- Large Data on Multiple Computers
  - Map Step (distribute data in key value pairs)
  - Reduce Step (combine data based on keys)
- Example: count unique words in documents
  - map step = document split into individual words, each gets a value of 1
  - reduce step, words with the same key (i.e. the word) are added together
    - result is number of times each word occurs in a document
- What are the keys (i.e. vectors)
Semantic Web

- Inclusion of semantic markup in web pages
- HTML markup is primarily layout. Some tokens have structure meaning (heading/paragraph), but most (bold, emphasize, etc.) are only formatting
- Ontologies
  - set of concepts in a domain, and relationships between concepts
  Description Logic -> concepts, roles, individuals
    - between propositional logic and 1st order predicate logic
Normalization

- Blocks, data types
  - blocks are functions, loops, etc.
  - data types used in blocks
  - not statement level
- Mapping of equivalent structures
  - loops, conditions, operator assignment
  - small function folding...
Vectors

- A expression is built for each code block
  - Which types are used (built in, user defined)
  - which functions are called
  - control blocks

- Expressed in the ontology
  - Class (method, data type, return data type)
  - object property (use of data type, function call, etc)

relations between expressions inferred by reasoner
  - subsumption, exclusion, identical
Clones

- So a clone is a block that uses the same data types and calls the same functions in the same way (i.e. condition calls vs loop calls).
- using a function from a class implies you are using the class (subsumption)

- Scalability
-- see also I. Keivanloo, J. Rilling, P. Charland, “Internet-scale Real-time Code Clone Search via Multi-level Indexing” 2011 18th Working Conference on Reverse Engineering, Limerick Ireland
Map-Reduce

- The authors don't really return to map-reduce, instead refer to an alternate paper about implementing DL using map reduce.
- Converting code to DL can easily be done in parallel.
Nicad

- uses syntax to format code consistently
- uses least common sequence to compare formatted code (token approach)