Smart Assessment of and Tutoring for Computational Thinking MOOC Assignments using MindReader

Hasan M. Jamil
Department of Computer Science
University of Idaho, USA
jamil@uidaho.edu

Abstract—One of the major hurdles toward automatic semantic understanding of computer programs is the lack of knowledge about what constitutes functional equivalence of code segments. We postulate that a sound knowledge base can be used to deductively understand code segments in a hierarchical fashion by first de-constructing a code and then reconstructing it from elementary knowledge and equivalence rules of elementary code segments. The approach can also be engineered to produce computable programs from conceptual and abstract algorithms as an inverse function. In this paper, we introduce the core idea behind the MindReader online assessment system that is able to understand a wide variety of elementary algorithms students learn in their entry level programming classes such as Java, C++ and Python. The MindReader system is able to assess student assignments and guide them how to develop correct and better code in real time without human assistance.

Index Terms—Authentic assessment; computational thinking; automated assessment; computer programming; program equivalence; semantic similarity

I. INTRODUCTION

A significant demand is known to exist for computer science (CS) graduates, and the US government has responded with the passing of the America Competes Act of 2007 [1] and subsequent refunding in 2011 to help train the much needed workforce. Additionally, the National Science Foundation has introduced the “CS for All” program with the goal that “all students should have the opportunity to learn CS in school.” This imperative requires that CS education move into K-12 poorly funded schools with woefully under prepared staff to provide education in this field. There are few teachers with the skills necessary to teach CS courses. In rural schools the situation is exacerbated by the fact that there may only be one teacher with math or science skills for the entire school.

Technological advances and economic realities are also prompting a shift in the way we learn, teach and deliver instructions to train our labor force. Tech savvy younger generation today find personalized online systems engaging and useful and are welcoming online and digital learning in all three settings – formal or institutional, blended, and self-paced and non-formal learning spaces. The expectation is that online systems will overcome much of the hurdles we face in formal education systems and will complement it in a larger way. Although some skepticism exists [2], [3], the excitement around Massive Open Online Courses (MOOC) and more institutional approach to digital learning using BBLearn[4] that are at the two ends of a spectrum, are fueled by these promises. All online universities, academies and institutes such as Coursera[2], KhanAcademy[3] and MITOpenCourseware[4] are then immediately faced with problems in three axes – content delivery, teaching and tutoring and assessment, much like the traditional systems do. They also grapple with enrollment and coverage, retention, cost, teaching effectiveness, and so on much like their formal counterparts. To combat these problems, new learning environments such as immersive, game-based, blended, personalized, self-regulated and self-paced, social, peer, and pair learning have been proposed, the effectiveness studies of which are ongoing [4], [5], [6].

But what we anecdotaly know already are of significant concern. For example, the retention rate in first year programming classes is extremely low nationally. A recent online study [7] found that about 60% STEM subject students drop out or transfer and about 55% never graduate in state and community colleges. MOOCs and other online institutes’ retention rates are even worse – about 90% enrollees never complete their courses [7], [8]. We believe the environment that currently exists within the online education community does not support many of the recommendations of experienced educationists summarized in reports such as [7], [9], and appear to retain the drawbacks of the traditional systems, and offer a mixed mode hodgepodge or a “succeed on your own” online setting.

However, the encouraging fact is that there have been significant progress in several areas of computer science that we believe can be leveraged and assembled together to build effective and smart cyber systems for online teaching, tutoring and assessment of entry level computing classes, and other STEM subjects. In our vision, such a system will complement a human instructor or mentor, and take on the role of a human observer to monitor students in real time and detect where she is making a mistake in her coding exercise, and immediately offer assistance by providing diagnostic comments and helpful pointers that most likely will cure the error [10].

In this paper, we introduce a novel prototype online system for tutoring and assessment, called the MindReader, for high school and freshman college students to aid learning programming languages. We develop necessary computational technologies to advance the science of computer program

---

1^http://www.blackboard.com/
2^https://www.coursera.org/
3^https://www.khanacademy.org/computing/computer-programming
4^https://ocw.mit.edu/courses/intro-programming/
understanding needed for digital tutoring, and online real time assessment of programming assignments. This system is aimed at complementing human instructors at a more massive scale fully automatically. For the want of space and brevity, however, we highlight only the salient features of MindReader and refer readers to [11] for a more detailed discussion.

II. RELATED RESEARCH

While some progress has been made in online instruction delivery, as well as in creating exciting learning environments, real time assessment and tutoring of STEM subjects online still remain at its infancy. Most often than not, these two areas rely largely on human interaction or MCQ tests, effectiveness of which are still being debated in general [12], and in STEM settings [13] and for computer programming classes [14] in particular. The general consensus appear to be that MCQ tests are great tools for formative and diagnostic assessments but for summative assessment, tests such as authentic assessment are more appropriate, particularly in computing courses.

The challenges in designing smart cyber systems for tutoring and summative assessment are manifold. Ideally, a tutoring or assessment system should not rely on a specific procedure for establishing correctness of a proof in mathematics, for example. Rather the logical argument in any order must be the basis. Such an assumption rules out most of the current approaches to establishing correspondence of a student response to a known solution. A few automated systems have attempted to capture this spirit in subjects such as mathematics [15] and physics [16] education with extremely narrow success. In computer science, the success has been mixed [17], [18].

To assess programming assignments, usually understanding the code semantically is required, and for a machine it would essentially mean determining the functional equivalence of a reference solution and the student solution, which is theoretically hard – deciding functional equivalence of two programs in general is NP-complete [19], and only in limited instances and for special classes of programs we are able to do so [20], [21]. Undeterred by this weakness, researchers took a different route and tried to assess correctness of programs by various means so that the method can be used in learning exercises and online settings [22] but faced complexity barriers of a different nature [23]. Other approaches used test data to assess correctness [24], [25] to match with known outcomes and “assume” correctness. We, however, are not aware of a system capable of tutoring or assessing computer programs fully automatically and comprehensively.

III. CDGs: HIERARCHICAL CONCEPT STRUCTURE

The program dependence graph (PDG) [20] based matching approach to determine code equivalence for the purpose of grading programming assignments is too simplistic although such approaches have been narrowly effective in detecting code clones [27], [28] and plagiarized codes [29], [30]. In particular, such techniques call for a complete enumeration of all possible solutions for every assignment, a largely daunting task, if not impossible. For example, consider an assignment that involves writing a code segment to swap the values of two variables. As shown in figure 1 a student cannot be penalized if she offered the code segment as a possible solution even though a PDG based matching approach will most likely fail to accept it as a possible solution. If a student offers a more sophisticated but unanticipated solution instead as shown below, she should be assigned higher credits, not less, though a PDG based grading will certainly be ineffective.

```cpp
#include <iostream>
using std::cout;

int main()
    int a=27, b=43, t;
    cout << "Before " << a
         << " * " << b << endl;
    t = a;
    a = b;
    b = t;
    cout << "After " << a
         << " * " << b << endl;
    return 0;
}
```

| Reference Solution | Student Solution |
|--------------------|------------------|
| `int main();` | `void swap(int i, int j)` |
| `int a=27, b=43, t;` | `int t = i;` |
| `cout << "Before " << a << " * " << b << endl;` | `i = j;` |
| `t = a;` | `j = t;` |
| `a = b;` | `int main(){}` |
| `b = t;` | `int a=27, b=43, t;` |
| `cout << "After " << a << " * " << b << endl;` | `cout << "Before " << a << " * " << b << endl;` |
| `swap(a, b);` | `swap(a, b);` |
| `cout << "After " << a << " * " << b << endl;` | `cout << "After " << a << " * " << b << endl;` |
| `return 0;` | `return 0;` |

![Fig. 1. Equivalent swapping code segments.](image)

In this paper, we propose a novel and a more effective approach to matching solutions based on the idea of concept dependence graphs (CDGs) in which nodes are matched semantically as opposed to syntactic matching using PDGs. In a CDG, each node represents a hierarchically defined concept, and the graph represents the precedence relationship among the concepts. Thus, the matching of two CDGs have a much higher likelihood of determining functional and semantic equivalence of two code segments necessary for grading assignments, and offering tutoring help. We illustrate the idea using a simple problem of averaging a list of values in C++.

For a list of $n$ values $x_i$, their average is the simple mathematical formula $a = \frac{\sum_{i=1}^{n} x_i}{n}$, represented as the CDG in figure 2 in which rectangles are declarations, ellipses are computable concepts, solid arrows represent precedence, and dashed arrows represent possible replacements. In the concept symbols, there are four quadrants which represents name (upper left), contextual concept parameters (lower left), node ID (upper right), and node membership (lower right). In CDGs, concepts are defined hierarchically, and terminal nodes are either declaration, or base computable concepts such as print, decide, for loop, while loop, and so on such that all variables needed for the base concepts are also in the CDG. CDGs can be simple or complex. A simple CDG is a connected and directed acyclic graph of base concepts and declarations, while a complex CDG is a forest of simple CDGs and CDGs involving concept nodes connected using dashed arrows.
Technically, a CDG is a graph \( \langle N, \prec \rangle \) of a set of nodes \( N = \bigcup_i n_i \subseteq V \) and a precedence relation \( \prec \subseteq \bigcup_j e_j \subseteq E \), where \( V \) is the set of all possible concepts, and \( E \) is all pairs \( 2^V \times V \). In figure 2 the CDG \( \langle \{u_1, u_4, v_6\}, \{u_1 \prec v_6, u_4 \prec v_6\} \rangle \) is a simple, while \( \langle \{u_1, u_4, v_3, v_7, v_8, v_9\}, \{u_1 \prec v_3, u_4 \prec v_3, v_3 \prec v_7, v_7 \prec v_8, v_8 \prec v_9\} \rangle \) is a complex. In figure 2 the concept counter loop is replaceable with the sequence \( v_7, v_8, v_9_\), or with node \( v_6 \) alone. In other words, the CDG in figure 2 assumes a counter loop can be implemented in two possible ways, and thus defines an equivalence relation.

Note that the concepts are also hierarchically defined. The concept average is defined as an aggregation of a list of values, followed by a division by the size of the list. An aggregate on the other hand is defined as the summation of the elements of the list inside a counter loop (note node \( v_4 \) is part of node \( v_3 \), the counter loop). Finally a counter loop is defined as a for loop or a while loop. For the student program \( P_s \) below, we can transform it to construct a corresponding CDG, and match it with the conceptual solution in the knowledgebase even if the student solution is implemented using a for loop.

1: #include <iostream>
2: void main() {
3:   int k=0, total, size=9, mean, elements[10];
4:   while (k<size) {
5:     total=total + elements[k];
6:     k++;  
7:   std::cout << total/(size+1);  
}

IV. FORMAL MODEL

Let \( L \) be a programming language, \( \mu_L \) be a function that can parse a program \( P \) in \( L \) and convert each sentence into either a declaration concept or a computational concept in the language \( L \) of MindReader. \( L \) consists of two types of expressions – abstract statements and precedence relations. Abstract statements are of two types: declaration and computational type. Declaration type expressions are tuples of the form \( \langle N, V, T, C \rangle \), where \( N \) is the statement number in \( P \), \( V \) is the variable name, \( T \) is the class of variable such as individual variable, boolean or a list, and \( C \) is the statement or program in which the statement is included. For example, statement number 5 and 6 in the program \( P_s \) above are contained in statement number 4, while the statement number 4 is contained in statement number 2. Likewise, statements 3 and 7 are contained in statement 2.

Similarly, computable expressions are tuples of the form \( \langle N, E, P, C \rangle \), where \( N \) is the statement number in \( P \), \( E \) is the type of executable statement such as assignment, loop or decision statement, \( P \) is a list of context sensitive parameters, and \( C \) is the statement number of which the statement is a part of. For example, statement 4 in program \( P_s \) is a while loop, represented as the expression \( \langle 4, \text{whileLoop}, \text{param}(\text{cond}(i<=n)), 2 \rangle \), and the expression \( \langle 6, \text{tran}, \text{param}(k,k+1), 4 \rangle \) represents statement 6. Finally, precedence relation is a set of expressions of the form \( n_1 \prec n_2 \), where \( n_1 \) and \( n_2 \) are statement numbers such that \( n_1 \) precedes \( n_2 \).

The language \( \mathcal{L} \) of MindReader is a tuple \( \langle \mu_L, \mathcal{C}, \Sigma, \Gamma, \Psi \rangle \) of a concept extractor \( \mu_L \), concept hierarchy \( \mathcal{C} \) (e.g., figure 3), concept mapper \( \Sigma \), concept dependence graph \( \Gamma \), and a subgraph isomorphism function \( \Psi \). The concept hierarchy organizes higher level concepts from computable expressions. For example, a counter loop can be a composite of an assignment, a while and an increment statement as discussed earlier in the context of figure 2. The \( \Sigma \) function transforms the CDG created by \( \mu_L \) into higher level concepts using the concept hierarchy \( \mathcal{C} \) and the CDG \( \Gamma \), iteratively. Therefore, given a program \( P \), the least fixpoint \( lfp(\Sigma(\mu_L(\mathcal{C}), \Gamma)) \) is the final CDG of a program \( P \). Observe that the concept hierarchy \( \mathcal{C} \), the CDG \( \Gamma \), and the summarization function \( \Sigma \) help abstract programs into CDGs and increases the matching likelihood with high level abstract algorithms stored as a reference CDG independent of their lower level implementations.
\[ \Psi(C_a, \text{lp}(\Sigma(\mu_L(P), C, \Gamma)) \approx 1, \text{ where } C_a \text{ is the conceptual CDG of any algorithm } a, \text{ then we assume that the submitted code is acceptable and correct. Conversely, if for any “un-
\] known” program } P, \ \Psi(C_a, \text{lp}(\Sigma(\mu_L(P), C, \Gamma))) \approx 1, \text{ then we can be confident that the unknown program is a candidate
implementation of the abstract and conceptual algorithm } a. \text{ This a significant and powerful method to determine functional
equivalence of unknown codes which is extremely difficult, if not impossible, using PDG based approach due to its inability
to summarize codes functionally.}

V. ASSESSMENT AND TUTORING USING MINDREADER

The high level architecture in figure 4 depicts MindReader’s
two broad subsystems for two distinct but complementary functions – tutoring and grading. In MindReader, all learners have a profile which includes background, past lessons, tests and tutoring activities, known problem areas, and their peer groups. MindReader generates tutorials based on students’
profile and level of programming competence expected along the lines of the systems such as [31] keeping in mind that for computation thinking classes, the challenges primarily involve the difficulties in learning the syntax and understanding the semantics and use of constructs such as loops, conditional statements, and simple algorithms [32]. For the purpose of both grading and tutoring, MindReader assembles the statement structures written by the student into possible CDGs using the concept structures in the Concept Database according to the rules in Concept Construction rule base with the aim of matching the CDG with one of the known templates in the Algorithm Templates. Failure to match CDGs of the student code and the reference template results into a dataflow pattern match using known and random test data of the compiled codes. Failure to match flow patterns forces a diagnostic feedback, but a success indicates a new way of solving a problem unknown to MindReader, and the new CDG is included in the knowledgebase after proper curation.  

VI. LEARNING COMPLEX CONCEPTS

Building concepts hierarchically and generating corresponding
CDGs though intuitive, learning new concepts could be challenging. In MindReader, we assume that it is impossible to enumerate all reference solutions regardless of the complexity. We thus adopt an incremental learning approach with the assistance of a panel of curators or experts in MindReader’s architecture in figure 4. To understand how MindReader learns new concepts, consider an abstract algorithm for bubble sort as shown in algorithm 4. Its C++ implementation \(Q\) as shown below, and its CDG representation \(C_b\) shown in figure 4 as a reference solution. Obviously, the \(\text{lp}(\Sigma(\mu_L(Q), C, \Gamma))\) will not match with \(C_b\), i.e., \(\Psi(C_b, \text{lp}(\Sigma(\mu_L(Q), C, \Gamma))) \ll 1\), since the loop in statement 1 is not a sentinel loop. But a dataflow analysis and random data test comparison will show a match, prompting a curation step and learning the rule that bubble sort can also be performed with an outer counter loop, and a reverse inner counter loop. Note that the blue starred nodes in the CDG in figure 5 will also need to be implemented.

\begin{algorithm}
\caption{Bubble sort}
\begin{algorithmic}
\State \textbf{Input:} A list of \(n\) values in random order
\State \textbf{Output:} Ascending order list
\State \textbf{set sorted} = false;
\While{not sorted}
\State \textbf{set sorted} = true;
\For{element \(i = 1, \ldots, n - 1\) do}
\If{element \(i <\) element \(i + 1\) then}
\State swap elements \(i\) and \(i + 1\);
\EndIf
\EndFor
\EndWhile
\end{algorithmic}
\end{algorithm}

VII. SUMMARY AND FUTURE RESEARCH

In this paper, we reported a late breaking result of a
research focusing on semantic understanding of student codes in an online learning environment. We have demonstrated
that CDG based matching code fragments have a higher
likelihood of detecting semantic and functional equivalence of two programs. The process is complemented by a dataflow
analysis and random testing regime to identify possible valid
solutions and learn new rules. We have also demonstrated that
detecting code clones and plagiarized codes based on PDGs
is fundamentally different from matching two codes functionally
using CDGs. In CDGs we substitute equivalent nodes under the
guidance of a template CDG, and concept hierarchy to
determine semantic similarity essential for grading tasks of
MOOC student assignments. It should be evident that summariza-
tion of concepts in the concept graphs allows for abstract
g礼拜 algorithm development, and it should be possible to actually
write codes in various languages as an inverse function and
develop new languages such as Scratch.

Initial evaluations of MindReader was encouraging and a
more serious performance analysis and comparison with ex-
isting systems is being planned. Once deployed, and students use it for a period of time, we plan to collect a large number of
coding examples and investigate students learning behav-
ior, and effectiveness and do comparative analysis with the
traditional classroom teaching. Identification of problem areas of
learning where a significant number of students are having
difficulty manifested by their inability to solve problems could imply gaps in instruction delivery, course content design or
learning habits warranting a revision, and could help develop
personalized teaching, tutoring and assessment regimes, and
measured for continuous improvement.

REFERENCES

[1] J. J. Kuenzi, “Science, Technology, Engineering, and Mathematics (STEM) Education: Background, Federal Policy, and Legislative Action,” https://tinyurl.com/65x5mz8, 2008, cRS Report for Congress,
Accessed on December 5, 2016.
[2] C. Domínguez, A. J. Elizondo, A. Sánchez, J. M. Blanco, and J. Heras, “A comparative analysis of the consistency and difference among online self-, peer-, external- and instructor-assessments: The competitive effect,” Computers in Human Behavior, vol. 60, pp. 112–120, 2016.

[3] A. Sayapin, “Multiple choice assessments: Evaluation of quality,” in IEEE EDUCON, Berlin, Germany, March 13-15, 2013, pp. 352–355.

[4] R. L. Edwards, J. K. Stewart, and M. Ferati, “Assessing the effectiveness of distributed pair programming for an online informatics curriculum,” Inroads, vol. 1, no. 1, pp. 48–54, 2010.

[5] W. Farag, “Comparing achievement of intended learning outcomes in online programming classes with blended offerings,” in ACM SIGITE, Calgary, AB, Canada, October 11 - 13, 2012, pp. 25–30.

[6] J. Sharp, “Traditional, online, and flipped: A preliminary analysis of instructional approaches on student performance in a c# programming course,” in 22nd AMCIS, San Diego, CA, USA, August 11-14, 2016.

[7] D. Holton, “Two courses that made a difference in student retention,” https://tinyurl.com/kw27vob, February 2016, edTechDev.

[8] T. Tauber, “The dirty little secret of online learning: Students are bored and dropping out,” https://tinyurl.com/cvnj6bv March 2013, quartz.

[9] M. Jazzar, “Online student retention strategies: A baker’s dozen of recommendations,” https://tinyurl.com/ot6egclp December 2012, faculty Focus.

[10] V. J. Marin, T. Pereira, S. Sridharan, and C. R. Rivero, “Automated personalized feedback in introductory java programming moocs,” in 33rd IEEE ICDE, California, USA, April 19-22, 2017, to appear.

[11] H. M. Jamil, “Smart assessment and tutoring of computational thinking mooc assignments using mindreader,” Department of Computer Science, University of Idaho, Moscow, Idaho, Tech. Rep., February 2017.

[12] I. Simonova, “Multiple-choice testing: Knowledge, or random choice?” in IEEE EDUCON, Istanbul, Turkey, April 3-5, 2014, pp. 819–823.

[13] J. M. Azevedo, “e-assessment in mathematics courses with multiple-choice questions tests,” in CSEDU, Volume 2, Lisbon, Portugal, 23-25 May, 2015, pp. 260–266.

[14] S. M. Shahidian, M. Hamilton, and D. J. D’Souza, “Instructor perspectives of multiple-choice questions in summative assessment for novice programmers,” Computer Science Education, vol. 20, no. 3, pp. 229–259, 2010.

[15] J. C. Gluz, F. Penteado, M. Mossmann, L. Gomes, and R. Vicari, “A student model for teaching natural deduction based on a prover that mimics student reasoning,” in ITS, Honolulu, HI, USA, June 5-9, 2014, pp. 482–489.

[16] S. Mehta, “Development and assessment of interactive web-based problem solvers in reinforcing math and physics concepts,” in WWW WebNet, Orlando, Florida, October 25-27, 2001, p. 844.

[17] P. Návrat and J. Tvarozek, “Online programming exercises for summative assessment in university courses,” in 15th CompSysTech, Ruse, Bulgaria, June 27-28, 2014, pp. 341–348.

[18] J. Sorva and T. Sirkkii, “Embedded questions in ebooks on programming: useful for a) summative assessment, b) formative assessment, or c) something else?” in 15th Koli Calling Conference on Computing Education Research, Koli, Finland, November 19-22, 2015, pp. 152–156.

[19] H. B. Hunt III, R. L. Constable, and S. Sahni, “On the computational complexity of program scheme equivalence,” SIAM J. Comput., vol. 9, no. 2, pp. 396–416, 1980.

[20] T. Eiter, M. Fink, H. Tompits, and S. Woltran, “Complexity results for checking equivalence of stratified logic programs,” in 20th IJCAI, Hyderabad, India, January 6-12, 2007, pp. 330–335.

[21] S. Chaudhuri and M. Y. Vardi, “On the complexity of equivalence of distributed pair programming,” in ACM PODS, Minneapolis, Minnesota, USA, May 24-26, 1994, pp. 107–116.

[22] W. Drabant and M. Milkowska, “Proving correctness and completeness of normal programs - a declarative approach,” TPLP, vol. 5, no. 6, pp. 669–711, 2005.

[23] H. Hangar, “Complexity of proving program correctness,” in TACS, Sendai, Japan, September 24-27, 1991, pp. 459–474.

[24] T. Tang, R. Smith, S. Rixner, and J. Warren, “Data-driven test case generation for automated programming assessment,” in ACM ITiCSE, Arequipa, Peru, July 9–13, 2016, pp. 260–265.

[25] G. Li, L. Yu, and H. Sun, “A framework for test data generation of object-oriented programs based on complete testing chain,” in 17th IEEE/ACIS SNPD, Shanghai, China, May 30 - June 1, 2016, pp. 391–397.

[26] T. Görg, “Intereprocedural pdg-based code clone detection,” Softwaretechnik-Trends, vol. 36, no. 2, 2016.

[27] H. Li, H. Kwon, J. Kwon, and H. Lee, “CLORIFI: software vulnerability discovery using code clone verification,” in Concurrence and Computation: Practice and Experience, vol. 28, no. 6, pp. 1900–1917, 2016.

[28] S. Wagner, A. Abdulkhaleq, I. Bogicevic, J. Ostberg, and J. Ramadani, “How are functionally similar code clones syntactically different? an empirical study and a benchmark,” PeerJ Computer Science, vol. 2, e49, 2016.

[29] C. Liu, C. Chen, J. Han, and P. S. Yu, “GPLAG: detection of software plagiarism by program dependence graph analysis,” in ACM SIGKDD, Philadelphia, PA, USA, August 20-23, 2006, pp. 872–881.

[30] F. Zhang, D. Wu, P. Liu, and S. Zhu, “Program logic based software plagiarism detection,” in 25th IEEE ISSRE, Naples, Italy, November 3-6, 2014, pp. 66–77.

[31] S. Boumiza, D. Souilem, and A. Bekiarski, “Workflow approach to design automatic tutor in e-learning environment,” in CoDiT, Saint Julian’s, Malta, April 6-8, 2016, pp. 265–268.

[32] C. Hulls, A. Neale, B. Komalo, V. Petrov, and D. J. Brush, “Interactive online tutorial assistance for a first programming course,” IEEE Trans. Education, vol. 48, no. 4, pp. 719–728, 2005.