

# Enhancing Learning Paths with Concept Clustering and Rule-Based Optimization

S.T. Fung, Vincent Tam and Edmund Y. Lam  
*Department of Electrical and Electronic Engineering*  
*The University of Hong Kong*  
*Pokfulam, Hong Kong.*  
*Email: {stfung, vtam, elam}@eee.hku.hk*

**Abstract**—Finding a good learning path with respect to existing reference paths of closely related concepts is very challenging yet important for effective course teaching and especially adaptive e-learning systems. There are various approaches including ontology analysis to extract the key concepts which could then be correlated to one another using an implicit or explicit knowledge structure for relevant courses. With the available correlation information, an effective optimizer can ultimately return a good learning path according to its predefined objective function. In this paper, we propose to obtain more thorough correlation information through concept clustering, which will then be passed to our rule-based genetic algorithm to search for better learning path(s). To demonstrate the feasibility of our proposal, a prototype of our ontology analyser enhanced with concept clustering and rule-based optimizer was implemented. Its performance was thoroughly studied and compared favorably against the benchmarking shortest-path optimizer on actual courses. More importantly, our proposal can be easily integrated into existing e-learning systems, and has significant impacts for adaptive or personalized e-learning systems through enhanced ontology analysis.

**Keywords**—concept clustering, learning path, ontology analysis, rule-based optimization.

## I. INTRODUCTION

With respect to a set of existing reference learning paths as often provided by domain experts or experienced instructors, finding an alternative and reasonably good learning path as another fixed sequence of closely related concepts is very challenging yet important for effective course teaching and particularly sophisticated e-learning systems [1]. It is worth noting that when solely compared to the original set of reference learning paths, the alternative learning path obtained will never excel any of the reference learning paths. However, in real-life applications, such alternative learning path may sometimes offer a more appropriate learning sequence of relevant concepts for individual classes or students.

To extract key concepts/topics from relevant course materials or exercises, there are many different approaches including ontology analysis [2], [3], [4] or statistical methods such as the support vector machine (SVM) [5]. After being extracted, these key concepts should then be correlated to one another using an implicit or explicit knowledge structure for relevant courses. For instance, based on the implicit knowledge structures of relevant courses, Chen [3] proposed

to infer the correlation between two concepts when most students gave wrong answers to both questions referring to the two involved concepts simultaneously. The proposal is simple and attractive. However, it obviously requires all students to give answers sensibly in the quiz. In case all students give their answers randomly during the quiz, the deduced correlation information will be of no use for further analysis or the ultimate planning of learning path(s). On the other hand, there are many other statistical approaches such as the statistical keyword extraction algorithm [6] based on the co-occurrence of keywords across various chapters/modules of course materials. In general, the higher the occurrence of identified keywords across the different modules of course materials, the more likely the relevant keywords will be extracted. Accordingly, the greater the co-occurrence of “relevant keywords” in the two concerned course modules denoting related concepts, the higher the similarity of these two course modules, therefore the larger their correlation values. With the obtained correlation factors from the ontology analysis, an effective optimizer can ultimately return a good learning path according to its predefined objective function.

In this paper, we propose to obtain more thorough correlation information through concept clustering, which will then be passed to our rule-based genetic algorithm to search for better learning path(s). To demonstrate the feasibility of our proposal, a prototype of our ontology analyser enhanced with concept clustering and rule-based optimizer was implemented. Its performance was thoroughly studied and compared favorably against the benchmarking shortest-path optimizer on actual courses. More importantly, our proposal can be easily integrated into existing e-learning systems, and has significant impacts for adaptive or personalized e-learning systems through enhanced ontology analysis.

This paper is organised as follows. Section II describes the preliminaries that are important for our subsequent discussion, and our previous findings on related works. Section III considers our proposal of enhancing the ontology analysis through concept clustering, that is essentially systematic grouping of closely related concepts. Section IV discusses about our rule-based and evolutionary optimization methods with the refined rules extracted from the enhanced ontology analysis. Section V gives a thorough comparison

of our implemented prototype against that of the benchmarking shortest-path optimizer on real engineering courses offered in the University of Hong Kong. Lastly, Section VI summarises this work and shed light on various possible directions for future investigation.

## II. PREVIOUS WORK

The goal of most previous work to find a good learning path is essentially to search for a fixed sequence of all the relevant concepts while satisfying the knowledge or pre-requisite requirement of such concepts behind each course module. A systematic approach is to construct a graph over all the involved concepts extracted from the course modules and then find the overall shortest path trying to optimize the correlation values between the associated concepts along the learning path. Initially, there is no edge between any course material, through the process of concept correlation, we try to link up relevant concepts by adding edges between them in the underlying concept graph/map. Constructing a concept map and its associated edges for any course module without the prior knowledge of the course structure is definitely a very challenging task.

A possible way is to perform ontology analysis [3] or statistical algorithms [6] to extract keywords that may possibly denote key concepts in relevant course modules as based on its frequency of occurrence in the course materials and/or other reasonable factors. Among such keyword extraction algorithm, the quickest approach is to extract words or phrases from relevant course materials of the course modules as based on a simple scoring scheme directly dependent on its frequency of occurrence. Each time a word appears simultaneously in any two course modules, it score one point. Clearly, the higher the score of a keyword representing a particular concept, the higher the similarity measure of the two course modules involving that specific concept.

Let  $\mathbf{M}$  be a set of  $n$  course modules, and  $S(i, j)$  is the similarity measure of two course module  $i$  and  $j$ . Basically, finding the shortest path  $\mathbf{P}$  to link up all the relevant course modules is trying to maximize the sum  $D$  of similarity measures of all consecutive course modules in the path  $\mathbf{P}$  such that

$$D = \sum_{i=1}^{n-1} S(P_i, P_{i+1})$$

However, this efficient search method may not necessarily yield a good learning path since

- 1) In most of the learning paths generated by the shortest path approach, consecutive course modules along the learning paths are closely related to each other locally. However, it lacks a global consideration of their linking/relationship across the whole course framework that may unavoidably generate learning paths of relatively lower quality in practical applications.

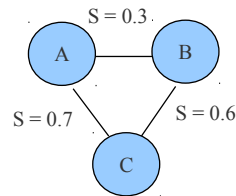


Figure 1. Illustration of multiple knowledge requirement.

- 2) The greedy shortest-path approach cannot address the problem of multiple knowledge requirements. For example, a course module  $C$  requires both course module  $A$  and  $B$  as its pre-requisite. Figure 1 shows this illustrative example of knowledge requirement involving the 3 courses.

The shortest path can be the specific learning path:  $A-C-B$  which violates the knowledge requirement of having the course module  $B$  as the prerequisite of course module  $C$ . In many real-life applications, it is very common that a high-level course module would require more than one course as its knowledge requirements.

Beside ontology analysis, there is another alternative approach to link up relevant modules with edges in the underlying graph by applying statistical methods. Among these methods, Chen [2], [3] proposed to use the students' answers in a quiz to deduce the implicit knowledge structure of the concerned course.

The motive of using students' quiz answers for inferring is based on the assumption that if most student simultaneously gave wrong answers to any two questions covering concept  $A$  and concept  $B$  respectively, we may then deduce that concept  $A$  and concept  $B$  may have some association. This method is simple and efficient since it only requires students' answers to construct the underlying concept graph and its linking edges, and students answers could be easily collected in most e-learning system. However, to extract meaningful correlation of concepts/modules from student's answers through this approach, all the students should sensibly give their answers in the quiz, which is an extremely difficult task and there is no vigorous way to detect whether the students are behaving sensibly or not during the quiz.

Figure 2 shows the quality of the learning paths as generated by Chen's approach on the four sets of data when compared to some randomly generated learning path. The quality of each generated learning path is measured in term of the sum of the violated distance defined as the variation of the generated learning path with respect to the reference learning path. Basically, the lower the sum of the violated distance, the better the learning path.

It is worth noting that only one of the 4 generated learning paths among all the test cases in Figure 2 shows some significant improvement in quality over those of the

randomly generated learning paths. One of the major reasons is the noisy input data sets that will significantly degrade the performance of Chen’s approach since it has previously mentioned that Chen’s approach requires all students’ sensible answers during the quiz. In case the students’ answers are not sensible or the original data set is somehow corrupted, the overall performance of Chen’s statistical approach will be drastically deteriorated. All in all, this statistical approach is very much dependent on input data set and therefore possibly not desirable for many practical applications. Nevertheless, Chen [3] proposed to consider the learning path optimization problem as a discrete constrained optimization problem [7].

### III. OUR PROPOSAL

Accordingly, a good learning path is essentially a sequence of course modules arranged in a way that can satisfy most/all the knowledge requirements of the involved course modules. For instance, the course concept/module as “*Summation*” is often considered as a knowledge requirement of “*Multiplication*” in an elementary course of Mathematics. On the other hand, it was observed in some other cases that the courses structure is mostly a flat structure consisted of several course modules across different domain knowledges, that may not be closely related to each other.

Table I shows the topics of course modules as extracted from the course materials of a Year-1 core course ELEC1401 about computer organization offered in the Department of Electrical and Electronic Engineering, the University of Hong Kong. It is interesting to note that the course concept/module “*Binary Arithmetic*” may not be closely related to “*IEEE Floating-point format*” yet it is a pre-requisite requirement of another course module “*IEEE Floating-point addition/subtraction*” that is under the same domain knowledge.

Basically, after employing the statistical keyword extraction method to extract out all relevant topics for the concerned course modules, ontology analysis is applied to build

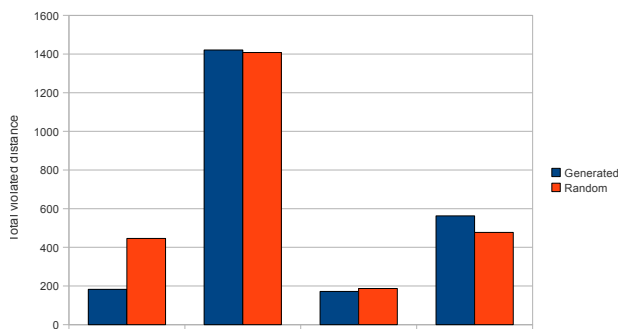


Figure 2. Comparison of learning paths generated by statistical analysis and randomly generated learning paths.

Group ID	Course module
1	Representation of numbers
1	Base/Radix conversion
1	Codes (Binary, Alphanumeric or Parity)
2	Binary Arithmetic
3	2’s complement for binary integer system
3	Sign bit with 2’s complement system
3	2’s complement subtraction
2	Fractional parts
2	Binary multiplication and division
4	IEEE floating point format
4	IEEE floating point addition/subtraction
5	Summary

Table I  
COURSE MODULES AND KNOWLEDGE GROUP ID’S FOR ELEC1401.

the module graph and edges between the course modules. However, instead of directly searching for the shortest path from the course module graph, our proposal works to extract precedence constraints/rules as  $precede(A, B)$  denoting that the course module  $A$  should precede the module  $B$  in the ultimate sequence of course modules for a learning path, which clearly define the knowledge requirement(s) of all involved course modules. Hence, our major objective is essentially to search for a feasible learning path that can satisfy all the precedence constraints. In addition, to facilitate the systematic and thorough analysis on the relationship among all the involved course concepts/modules, we propose to enhance the ontology analysis through the concept clustering technique that will categorize each course module into different predefined knowledge/subject group. Accordingly, each course module will be assigned with a knowledge group ID, that is deduced from our proposed concept clustering algorithm. Then, our enhanced ontology analyser will work to deduce precedence rules inside each knowledge group as according to their assigned knowledge group ID and also across different knowledge groups. Intrinsically, our concept clustering algorithm is applied to group contextually more similar concepts into the same knowledge group. Therefore, there can still exist some knowledge requirements occurred as indirect associations within or across the knowledge groups, that may preserve such kinds of knowledge requirements to a certain extent.

Given a set of  $n$  course modules and their corresponding course materials, we can obtain a feasible learning path through the following procedure.

- 1) **Preprocessing:** extract concept title and description from the concerned course materials.
- 2) **Keyword extraction:** deduce the importance of keywords through a document classification technique in which the co-occurrence statistical information based keyword extraction algorithm proposed by [6] is adapted in our proposal.
- 3) **Parameterize:** assume that there are  $M$  keywords

and key phrases in total extracted from all  $n$  sets of course materials, a  $M$ -dimensional Euclidean space can be constructed accordingly. Each course module is parameterized as a keyword vector which represents the corresponding concept in the  $M$ -dimensional euclidean space, with each dimension representing the importance of a keyword in the relevant course.

- 4) **Computing the correlation coefficient matrix:** the correlation coefficient matrix  $R$  is an  $n \times n$  matrix, and  $R_{ij}$  is the similarity measure of the course module  $i$  and  $j$ .

$$R_{ij} = \cos(\theta) = \frac{w_i \cdot w_j}{\|w_i\| \|w_j\|}$$

where  $w_i$  and  $w_j$  are the keyword vectors of the course module  $i$  and  $j$  respectively.

- 5) **Concept clustering:** all course modules are clustered by using the K-means clustering algorithm on their similarity measures to categorise the course modules into different knowledge groups. Each cluster of course module is treated as an individual knowledge domain under the course.
- 6) **Rule extraction:** precedence rules, as in the form of constraints, are extracted within each cluster and across the various clusters. The details of the rule extraction process will be discussed later.
- 7) **Learning path optimization:** the genetic algorithm is used to optimize the learning path/sequence of relevant course modules. Its detail will be thoroughly considered in Section IV.

Basically, the rules represent the prior/posterior knowledge requirements as extracted from the course modules of different knowledge groups as formed by our concept clustering technique. For instance, the precedence constraint  $precede(i, j)$ , or equivalently expressed as the rule  $\langle i, j \rangle$ , formally specifies that the course module  $i$  should be taught before the course module  $j$  in the designated learning path. The rule set is extracted in the following steps:

- 1) **Extraction of intra-domain knowledge rules:** we consider only knowledge groups (or clusters) containing more than one course module. For each knowledge cluster  $C_i$ , extract  $n - 1$  rules from the best  $n - 1$  course module tuples that attain the highest similarity values.
- 2) **Extraction of inter-domain knowledge rules:** assume that there are  $n$  knowledge groups,  $n - 1$  inter-domain knowledge rules can be extracted by considering the  $n - 1$  pairs of knowledge groups that have the highest inter-domain knowledge correlation coefficients.

#### IV. A RULE-BASED AND EVOLUTIONARY OPTIMIZATION APPROACH

The genetic algorithm we adopted in our proposal to optimise a learning path with respect to the extracted rule

set is fairly standard. The chromosome string is defined as a sequence of  $n$  integers ranged from 1 to  $n$  in which each integer denotes the corresponding course module. The whole chromosome string represents the learning path/sequence of the course modules covered in the whole course.

The detail of the genetic algorithm is given as follows.

- **Fitness function :** the fitness function is a performance indicator used to determine the quality of the generated learning path as measured by the number of precedence rules violated by the learning path itself. Basically, the more rules the generated learning path is violated, the worse the quality of the generated learning path.
- **Reproduction:** reproduction is the operation to generate new chromosomes by manipulating the “parent chromosomes”. It is an important operation that can greatly influence the overall performance of the genetic algorithm. Essentially, reproduction includes the crossover, mutation and random generation operation. In our adopted genetic algorithm to optimise for the learning paths directly denoted by individual chromosomes, the size of the reproduction pool is 100 chromosomes. After each iteration, the best 5 chromosomes will be carried to the next iteration with 80 new chromosomes generated by the crossover operator, 10 chromosomes generated by the mutation operator and the last 5 generated randomly.
- **Crossover operation:** in each crossover operation, two chromosomes ( $\mathbf{X}$  and  $\mathbf{Y}$ ) will be selected by roulette-wheel selection to perform their crossover. In order to avoid illogical learning path, that is having multiple occurrence of the same integers inside the chromosome string, and also retaining the basic sequential order of both chromosomes, a special segment-based crossover scheme is used in which the randomly selected segment ranging from  $i \dots j$ , where  $i < j$ , will be swapped between the two chromosome  $\mathbf{X}$  and  $\mathbf{Y}$ .
- **Mutation operation:** the mutation operation randomly selects two indices and swap the serial numbers in the involved chromosome.

#### V. AN EMPIRICAL EVALUATION

To demonstrate the effectiveness of our proposal, a prototype of our enhanced ontology based analyser integrated with the rule-based genetic algorithm was implemented and evaluated on 3 undergraduate Engineering courses including the ENGG1002 - Computer programming and application, ELEC1401 - Computer organization and microprocessor and ELEC2201 - Signals and linear systems as the actual test cases. For a more thorough investigation, two different schemes for defining the concerned course concepts are adopted. For ENGG1002 and ELEC2201, each individual lecture note is treated as one course concept so that each course concept may contain a relatively larger amount of

information/knowledge. However, for ELEC1401, each lecture note is further broken down into many smaller topics, and each topic is treated as 1 course concept such that each course concept is relatively simpler, therefore more likely to have a larger number of course concepts to be highly correlated to each other.

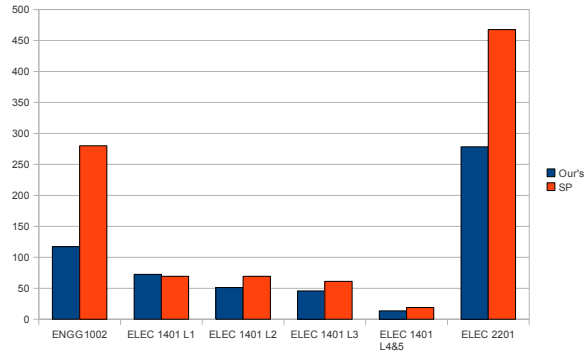


Figure 3. Performance of learning paths generated by our proposal (Ours) against those generated by the shortest-path (SP) approach.

To evaluate the performance of our proposal, the original teaching sequences of the above courses are used as the reference learning path for comparison. The original teaching sequence can be represented by a set of reference prior/posterior rules. For example, the teaching sequence of 5, 1, 4, 6, 3, 2 can be represented by the reference rule set  $\langle 5, 1 \rangle$ ,  $\langle 1, 4 \rangle$ ,  $\langle 4, 6 \rangle$ ,  $\langle 6, 3 \rangle$  and  $\langle 3, 2 \rangle$ . The total violated distance of any generated learning path is defined as below:

$$\gamma = \sum_{i=1}^{n-1} \sum_{j=(i+1)}^n \text{MAX}(P_i - P_j, 0)$$

where  $\gamma$  denotes the total violated distances of the generated learning path,  $n$  represents the total number of course concepts in a course module,  $P_i$  and  $P_j$  are the corresponding position index values in the generated learning path that violates the reference rule set. Figure 3 gives the total violated distance of learning paths generated by our proposal against those generated by using the shortest-path approach. The results clearly show that our proposal outperforms the shortest-path approach by returning better learning paths in most cases. It is worth noting that the performance difference in ENGG1002 and ELEC2201 is significantly larger. This is probably due to the fact that the amount of knowledge encapsulated in a course concept is larger in ENGG1002 and ELEC2201 as we regard each set of lecture notes as one single concept. In such case, most of the concepts are loosely correlated and having more course modules requiring course modules as knowledge requirement, thus making it difficult for the shortest-path approach to search for a reasonably good learning path. On the other hand, using our constraint based approach can effectively determine the

prior and posterior sequence of pairs of course modules during the more thorough and systematic process of rule extraction as enhanced by our concept clustering technique, and therefore significantly minimizing the search difficulty for the rule-based optimization as later performed by the genetic algorithm in such cases.

## VI. CONCLUDING REMARKS

In this paper, we propose to conduct a more thorough and systematic ontology analysis through concept clustering. The refined concept correlation information will then be passed to the rule-based genetic algorithm to optimise for better learning path(s). To demonstrate the feasibility of our proposal, a prototype of our ontology analyser enhanced with concept clustering and rule-based optimizer was implemented. Its performance was compared favorably against the benchmarking shortest-path optimizer on various actual courses. More importantly, our proposal clearly demonstrates the importance of enhanced ontology analysis for the overall performance of adaptive or personalized e-learning systems [1] yet can be easily integrated into such e-learning systems.

## ACKNOWLEDGMENT

The authors would like to thank Dr. Kinshuk and Dr. Daniel Churchil for their fruitful discussions.

## REFERENCES

- [1] W.-S. Lo, I.-C. Chung, and H.-J. Hsu, "Using ontological engineering for computer education on online e-learning community system," in *Proceedings of the International Conference on Education Technology and Computer (ICETC)*. IEEE, April 2009, p. 167.
- [2] C.-M. Chen, C.-J. Peng, and J.-Y. Shiue, "Ontology-based concept map for planning personalized learning path," in *Proceedings of the IEEE Cybernetics and Intelligent Systems*, Chengdu, November 2008, pp. 1337–1342, ISBN: 978-1-4244-1673-8.
- [3] C. Chen, "Ontology-based concept map for planning a personalised learning path," *British Journal of Educational Technology*, vol. 40, no. 6, pp. 1028–1058, 2009.
- [4] N. Guarino and C. Welty, "Conceptual modeling and ontological analysis," in *Proceedings of AAAI-2000*, 2000, url at <http://www.cs.vassar.edu/weltyc/aaai-2000/>.
- [5] C. J. BURGESS, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.
- [6] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 1, pp. 157–170, 2004.
- [7] E. Tsang, *Foundations of Constraint Satisfaction*. University of Essex, 1993.