J. Comput. Educ. (2014) 1(4):335–352
DOI 10.1007/s40692-014-0016-8

A new framework of concept clustering and learning path optimization to develop the next-generation e-learning systems

Vincent Tam · Edmund Y. Lam · S. T. Fung

Received: 25 August 2014 / Revised: 14 September 2014 / Accepted: 18 September 2014 /
Published online: 25 October 2014
© Beijing Normal University 2014

Abstract Most existing e-learning systems strictly require the course instructors to explicitly input the pre-requisite requirements and/or some relationship measures between the involved concepts/modules such that an optimal learning path as a sequence of the involved concepts can be determined for a class or an individual after considering the student's academic performance, educational background, learning interests, learner profile, learning styles, etc. In some cases, the learning path is determined solely by human experts. Since human can be biased, the course instructor's views on the relations of the involved concepts/modules can be imprecise or even contradictory, thus prohibiting any logical deduction of an optimal learning path. Besides, human experts may ignore or possibly be confused by contradictory requirements in the real-world applications. Therefore, we propose a new and systematic framework to develop the next-generation e-learning systems that will perform an explicit semantic analysis on the course materials to extract the individual concepts, and then grouped by a heuristic-based concept clustering algorithm to compute the relationship measures as the basis for extracting the pre-requisite requirements/constraints between the involved concepts. Lastly, an evolutionary optimizer will be invoked to return the optimal learning sequence after considering multiple experts' recommended learning sequences which may contain conflicting views in different cases. It is worth noting that our proposed and structured framework with the seamless integration of concept clustering and learning path optimization uniquely represents the first attempt to facilitate the course designers/instructors in providing more personalized and 'systematic' advice

V. Tam (✉) · E. Y. Lam · S. T. Fung
Department of E.E.E., The University of Hong Kong, Pokfulam Road, Hong Kong, China
e-mail: vtam@eee.hku.hk

E. Y. Lam
e-mail: elam@eee.hku.hk

S. T. Fung
e-mail: stfung@eee.hku.hk

through optimizing the learning path(s) for each individual class/learner. To demonstrate the feasibility of our prototype, we implemented a prototype of the proposed e-learning system framework, and also enhanced the original optimizer with the hill-climbing heuristic. Our empirical evaluation clearly revealed the many possible advantages of our proposal with interesting directions for future investigation.

Keywords Concept clustering · Learning objects · Learning styles · Learning path optimization

Introduction

Most online e-learning systems (Chen et al. 2008; Lo et al. 2009) allow the learners or students to specify their own learners' profiles and learning styles. Yet these e-learning systems (Baosteel Talent Development Institute, 2013; CityU 2009; Monroe 2012) usually also demand the pre-requisite requirements between course modules and/or some relationship measures between the involved concepts to be explicitly specified by the course designers/instructors so that an optimizer will be ultimately employed to return an optimal learning path/sequence (Acampora et al. 2008; Wong 2009; Tam et al. 2012) of involved concepts or modules for each individual learner after considering his/her academic performance, educational background, learning interests, learner's profile, learning style, and possibly other learners' profiles of relevant groups/levels (Chen et al. 2008). However, there are various pitfalls in solely relying on individual course instructor's input on the relationship among the involved concepts since human experts can be biased. Moreover, the decision will become more complicated when various instructors hold conflicting views on the relations of the involved concepts that may sometimes contain incomplete information in the form of learning objects (Fung et al. 2011) for any plausible and logical deduction of an optimal learning path as the best possible sequence of involved concepts recommended for an individual class/learner to study in a specific course module.

There were some previous works focused on using statistical or machine learning approaches (Burges 1998) for the semantic analysis (Chen et al. 2008) of the relevant course materials and/or the ultimate optimization (Fung et al. 2011) of learning paths or sequences of involved course concepts or modules. For instance, (Wong and Looi 2009) proposed an adaptive learning pathway generation approach using the ant colony optimization method. Furthermore, we studied in a previous work about the use of a heuristic based concept clustering algorithm (Malmberg and Maskell 2002; Fung et al. 2011) for the clustering of relevant concepts and their possible relationship, and lastly optimize the resulting learning path of involved concepts for a group of learners based on a predefined objective function and a reference path of concepts provided by a course instructor. *However, **none** of these earlier proposals consider a complete e-learning system framework that tries to automate or at least semi-automate the whole process from searching for any*

possible relationship among involved concepts or modules to the ultimate generation and optimization of the resulting learning paths, especially in the presence of incomplete information stored in the form of learning objects to describe the involved concepts and also possibly with multiple reference paths containing conflicting views on the prerequisite requirements of the course concepts or modules as provided by different human experts. Therefore, we propose in this paper a complete e-learning system framework that can perform an explicit semantic analysis (ESA) (Fung et al. 2011; Tam et al. 2012) on the course materials, possibly aided by the relevant Wiki articles for any missing information about the involved concepts, to formulate the set of individual concepts. This is followed by using a heuristic-based concept clustering algorithm to group relevant concepts before finding their relationship measures. Lastly, an evolutionary optimizer will be used to return the optimal learning sequence after considering multiple experts' recommended learning sequences possibly containing conflicting views. To demonstrate the feasibility of our prototype, we implemented a prototype of the proposed e-learning system framework. Our empirical evaluation clearly revealed the many possible advantages including the flexibility and extensibility of our proposal. The key contributions of our work are at least four-folded. First, by formally stating the pre-requisite relations between the involved concepts explicitly as "constraints", our work clearly suggests to the conventional view of formulating the learning sequences of course modules as a subjective and unstructured process, based solely on human experts' experience and opinions, a more structured framework of decisions in which designing a learning path for any course module can be regarded as a constrained optimization process subject to various criteria for more critical reviews and thorough analyses in some later stage of system development. Second, it is worth noting that our proposed and structured framework with the seamless integration of concept clustering and learning path optimization uniquely represents the first attempt to assist the course designers/instructors in providing more personalized and 'systematic' advice through optimizing the learning path(s) as a more suitable and viable solution for each individual class/learner. For example, when an individual student or group of students is found to have some difficulty in understanding a specific concept through certain assessment, the course instructor may adopt a different and possibly longer learning path of concepts that may help to achieve better academic performance for this individual or group of students. Third, there are many possible directions for the future investigation of our proposal including the application of our proposed system framework to develop an effective e-learning system (CU 2013; PPDH 2011) for learning languages, especially for foreigners to learn about the writing and understanding of Chinese characters/phases since most Chinese characters/phases can be readily broken up into some basic units representing physical components or abstract concepts in our daily living. Last but not least, our proposal can be easily integrated into existing e-learning systems, and has significant impacts for those e-learning systems to evolve into adaptive and personalized e-learning systems (CityU 2009) through our enhanced ontology analysis (Guarino and Welty 2000).

This manuscript is organized as follows. The second section describes the preliminaries that are important for our subsequent discussion, and our previous

findings on related works. The third section considers our proposal of performing an explicit semantic analysis, followed by enhancing the ontology analysis through concept clustering, that is essentially systematic grouping of closely related concepts, and lastly applying an evolutionary optimizer to find an optimal learning path of involved concepts or modules. The fourth section details our rule-based and evolutionary optimizer with the refined rules extracted from the enhanced ontology analysis while the fifth section reveals on how we try to improve the original evolutionary optimizer with the hill-climbing heuristic. The second last section gives a thorough comparison of our original prototype against that of the benchmarking shortest path optimizer, and also that of the enhanced optimizer on various Engineering courses offered in the University of Hong Kong. Last but not least, the last section summarizes this work and shed light on many possible directions for future investigation.

Related work

The goal of most previous work to find a good learning path (Acampora et al. 2008; Wong and Looi 2009; Tam et al. 2012) is essentially to search for a fixed sequence of all the relevant concepts while satisfying the knowledge or prerequisite 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 2 concepts as extracted from the underlying 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 (Chen 2009) or statistical algorithms (Matsuo and Ishizuka 2004) 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 scores one point.

Clearly, the higher the score of a keyword to represent a particular concept, the higher the similarity measure of the two course modules involving that specific concept. Let 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 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 P such that

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

As shown in Fig. 1, the shortest learning path can be: *A–C–B* which may violate the knowledge requirement of having the course module *B* as the prerequisite of course module *C* as stated in the underlying curriculum or subject domain. 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.

Besides the ontology analysis, there is another alternative approach to link up relevant modules with edges in the underlying concept graph by applying statistical methods. Among these methods, (Chen et al. 2008, 2009) 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 the logical inferencing is based on the assumption that if most students 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 yet efficient since it only requires students' answers to construct the underlying concept graph and its edges, and students' answers could be easily collected in most e-learning systems. 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 4 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 distances as defined by the differences between the generated learning paths and the reference learning paths. Basically, the lower the sum of the violated distances, the better the learning path. It is worth noting that only 1 of the 4 generated learning paths among all the test cases in Fig. 2 shows some significant improvement in the 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 been previously mentioned that Chen's approach requires all students to answer sensibly during the quiz.

Our proposed e-learning framework

Accordingly, a learning path of good quality is essentially a sequence of course modules arranged in such 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 prior knowledge requirement or pre-requisite 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

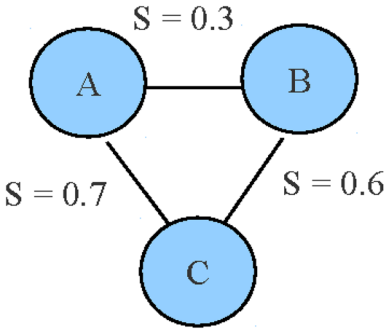

Fig. 1 An illustration of the knowledge requirements of multiple concepts

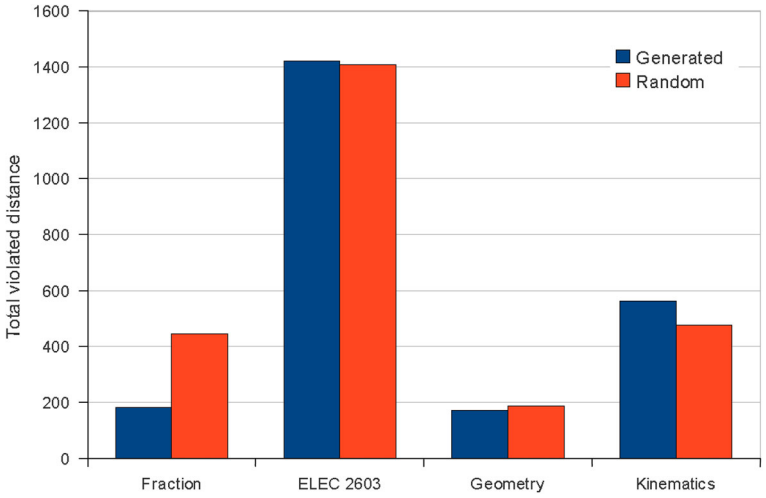

Fig. 2 A comparison of learning paths generated by the statistical analysis and randomly generated learning paths

flat structure consisted of several course modules across different domains of knowledge that may not be closely related to each other.

Table 1 shows the topics of course modules as extracted from the online course material of an existing Year-1 core course ELEC1401 on Computer Organization and Microprocessors 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 of knowledge about computer architecture or systems. Basically, after applying the statistical keyword extraction method to extract out all relevant topics for the concerned course modules, an ontology analysis (Guarino and Welty 2000) is applied to build the concept/module graph and edges between the course concepts/modules. However, instead of directly searching for the shortest path from the course

Table 1 Modules and Knowledge Group ID's of ELEC1401

Group id	Course module
1	Representation of binary 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

module graph, our proposed framework works to extract the 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 the 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 category each course module into different predefined knowledge/subject group. Accordingly, each course module will be assigned with a knowledge group identification number (ID), that is deduced from our proposed concept clustering algorithm. Then, our enhanced ontology analyzer 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 may still be some knowledge requirements occurred as indirect associations within or across the knowledge groups, aiming to preserve such kinds of knowledge requirements to a certain extent.

Essentially, our proposed and complete framework of e-learning systems relies on the formal definitions of the learning path optimization problem (LEAPOP) with its feasible and optimized, possibly sub-optimal, solution(s) as the learning path(s) obtained by the underlying search method. The LEAPOP is intrinsically a constrained optimization problem (COP) that can be considered as a variant of the job-shop scheduling problems containing some 'soft' precedence constraints/rules and various objective criteria as set by different human experts to maximize the learning interests and/or minimize the learning time, etc.

The formal definitions of the constraint satisfaction problem (CSP) as the basis of the COP (Tsang, 1993), and the COP are given as follows.

Definition 1 A CSP is defined by a triple $P = (X, D, C)$ where X is a n-tuple of variables $(x_1, x_2,\ldots, x_n)$, D is the corresponding n-tuple of domains, i.e. $(D_1, D_2,\ldots, D_n)$ such that $x_i \in D_i$, and D_i is a set of values, for $i = 1,\ldots, n$. C is a t-tuple of constraints $(c_1,c_2,\ldots, c_t)$.

Definition 2 A constrained optimization problem (COP) is defined as an optimization problem under the framework of the constraint satisfaction which is defined by a 4-tuple $P' = (X, D, C, O)$, where (X, D, C) forms a CSP and $O$is an objective function corresponding to the given criterion to be minimized or maximized.

It is worth noting that in practical applications, the objective function$O$can also be expressed as the weighted sum of various objective functions for minimization or maximization, i.e. $O = \varphi_1 f_1 + \varphi_2 f_2 + \cdots + \varphi_k f_k$ where φ_j is the corresponding weighting value assigned to each objective function f_j for $j = 1,\ldots, k$ in order to reflect its relative importance in the overall objective function O.

Based on the above definition of the COP, the LEAPOP can be formally defined as below.

Definition 3 A LEAPOP is defined as a variant $\hat{P} = \left(\tilde{X}, \tilde{D}, \tilde{C}, \tilde{O}\right)$ of the COP customized in the area of learning analytics such that $\tilde{X}$ is a m-tuple of variables $(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m)$ for the m concepts/modules involved in a specific subject domain, $\tilde{D}$ is the corresponding m-tuple of domains, i.e. $\left(\tilde{D}_1, \tilde{D}_2, \ldots, \tilde{D}_m\right)$ such that $\tilde{x}_l \in \tilde{D}_l$, and $\tilde{D}_l$ is a set of sequence numbers for assigning to each individual concept, for $l = 1,\ldots, m$. $\tilde{C}$ is a s-tuple of precedence constraints $(\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_s)$ and $\tilde{O}$ is the weighted objective function corresponding to the weighted sum of a set of criteria to be minimized or maximized.

It should be noted that the set $\tilde{C}$ of precedence constraints may contain both hard and soft precedence constraints as depending on the nature of the underlying subject domain. All the hard precedence constraints must be strictly satisfied for any feasible learning path p while soft precedence constraints can be violated in some cases according to the specific threshold values set by the concerned course designers or instructors.

Given a set of m course modules and their corresponding course materials, a feasible learning path p that satisfies all the hard precedence constraints and possibly violates some soft precedence constraints in the set $\tilde{C}$ and then optimized by some iterative search method with respect to the weighted objective function $\tilde{O}$ of the underlying LEAPOP can be obtained through the following *7* steps as illustrated in Fig. 3.

The detailed procedure or operation of each of the 7 steps proposed in Fig. 3 is clearly explained as below.

(1) *Preprocessing* to extract the concept title and description from the concerned course materials.

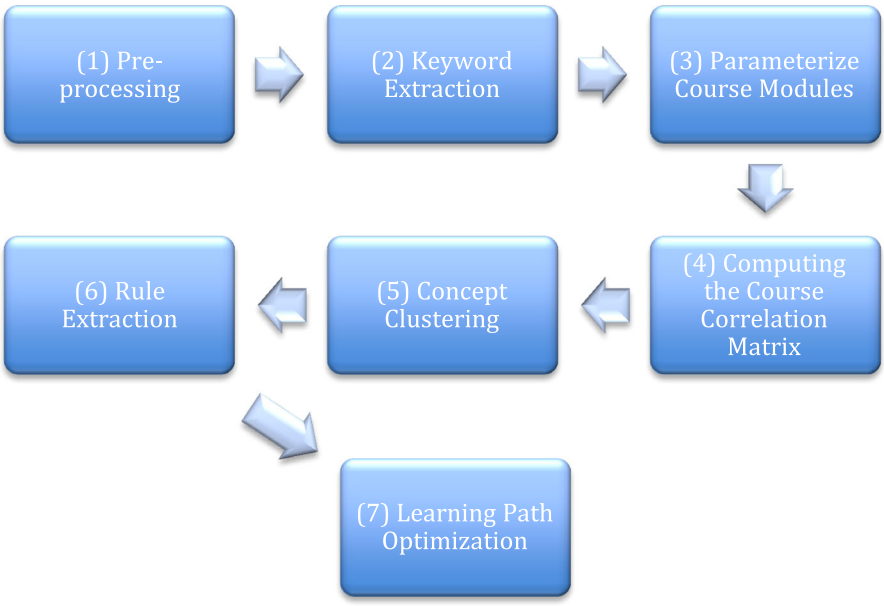

Fig. 3 The detailed steps of our proposed e-learning framework for learning path optimization

(2) *Keyword Extraction* to deduce the importance of keywords through a document classification technique in which the co-occurrence statistical information based keyword extraction algorithm proposed by (Matsuo and Ishizuka 2004) is adapted in our proposal.

(3) *Parameterize Course Modules* assuming that there are n keywords and key phrases in total extracted from all m sets of course materials, a n-dimensional Euclidean space can be constructed accordingly. Each course module is parameterized as a vector of keywords that represents the corresponding concept in the n-dimensional Euclidean space, with each dimension representing the importance of a keyword in the relevant course.

(4) *Computing the Course Correlation 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.

$$R_{ij} = \cos(\theta) = \frac{\omega_i \cdot \omega_j}{\|\omega_i\|\|\omega_i\|}$$

where ω_i and ω_j is the keyword vector 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 categorize the course modules into different knowledge groups. The K-means clustering algorithm

is chosen due to its efficiency and simplicity to handle the potentially large number of concepts and their possible relations that may be reviewed during concept clustering for each course module. When we have n concepts, the possible number of relations can be n^2 for each cluster, thus with the overall computational complexity as up to n^3 in the search to build up each cluster. Hence, we need a simple yet efficient method such as the K-means clustering algorithm to perform the iterative search process. Essentially, each cluster of course module is treated as an individual knowledge domain under the course module.

(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 in the subsequent paragraphs.

(7) *Learning Path Optimization* an evolutionary algorithm is used to optimize the learning path/sequence of relevant course modules. In some previous work, heuristic-based optimizers such as the ant optimization algorithm (Wong and Looi 2009) or other memetic algorithm (Acampora et al. 2008) have been used to find the optimal learning path/sequence with respect to the predefined objective function for a group of learners. In this work, we proposed to use an efficient evolutionary algorithm to return the optimal learning path based on an objective function to minimize on the learning path distance with respect to a set of experts' recommended paths. The detail of our evolutionary learning path optimizer will be thoroughly considered in the next section.

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 set of precedence rules is extracted as in the following steps:

i. *Extraction of the intra-domain precedence rules* we consider only knowledge clusters (or groups) containing more than one course module. For each knowledge cluster C_i, our algorithm will extract $(n - 1)$ rules from the best $(n - 1)$ course module tuples that attain the highest similarity values.

ii. *Extraction of the inter-domain precedence rules* assuming that there are n knowledge groups, there can be $(n - 1)$ inter-domain knowledge rules extracted by considering the $(n - 1)$ pairs of knowledge groups with the highest inter-domain knowledge correlation coefficients.

A rule-based evolutionary optimizer

The evolutionary algorithm we adapted in our proposal to optimize 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. Figure 4 shows the basic steps of our evolutionary optimizer used for learning path optimization.

The detail of the evolutionary algorithm as highlighted in Fig. 4 is clearly explained as below.

- *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 optimize for the learning paths directly denoted by individual chromosomes, the size of the reproduction pool is 100 chromosomes. In 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 (***X*** and ***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 $i \cdots j$, where $i < j$, will be swapped between the two chromosome ***X*** and ***Y***.
- *Mutation operation* the mutation operation randomly selects two indices and swaps the serial numbers in the involved chromosome.

Improving the evolutionary optimizer

Figure 5 shows the basic steps of our evolutionary optimizer enhanced by the hill climbing (HC) heuristic (Tsang 1993) to try to search for learning paths of even better quality for a group or individual learners.

Our enhanced evolutionary algorithm will iterate to apply both crossover and mutation operators to evolve new learning paths until the solution is converged, or the maximum number of generations preset as *10,000* is exceeded. In each iteration, after the new learning paths are generated, our enhanced evolutionary optimizer will invoke the hill-climbing heuristic function to perform the HC on the fitness function to check whether the quality of the new learning paths is better than that of the

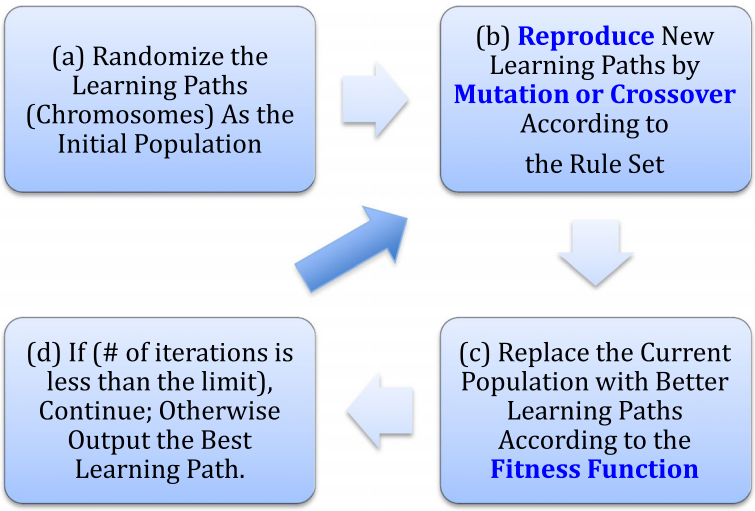

Fig. 4 Our evolutionary optimizer for learning path optimization

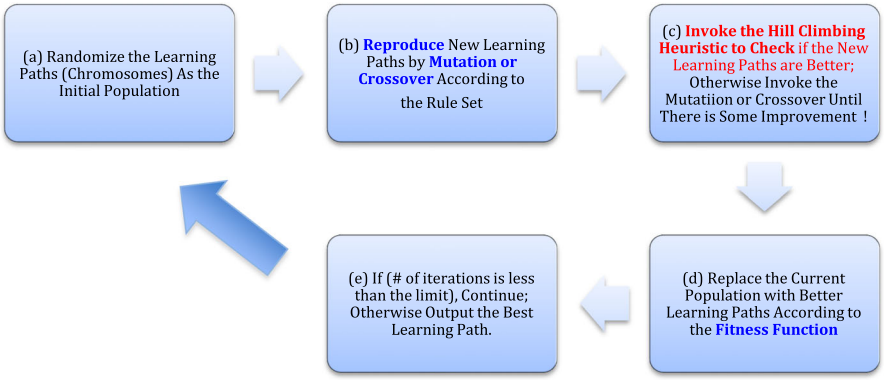

Fig. 5 Our evolutionary optimizer enhanced by the hill climbing

parent chromosome. In case not, the hill-climbing function will call the crossover and mutation operator continuously until the resulting quality of the generated learning path is improved.

Our empirical evaluation

To demonstrate the effectiveness of our proposal, a prototype of our enhanced ontology based analyzer integrated with the rule-based genetic algorithm was implemented and evaluated on 4 undergraduate Engineering courses including the ELEC1401—Computer organization and microprocessor, ELEC1502—Object

oriented programming and data structures, ELEC2201—Signals and linear systems and ELEC2603—Systems and network programming as the actual test cases. For a more thorough investigation, two different schemes for defining the concerned course concepts are adopted. For ELEC1502, ELEC2201 and ELEC2603, 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.

To evaluate the performance of our proposal, the original and other domain experts' recommended teaching sequences of the above courses are used as the reference learning paths for comparison. The original teaching sequence can be represented by a set of reference prior/posterior rules. For instance, the original 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:

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

where γ 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 6 gives the total violated distance of learning paths generated by our proposal against those generated by using the shortest-path (SP) approach firstly on

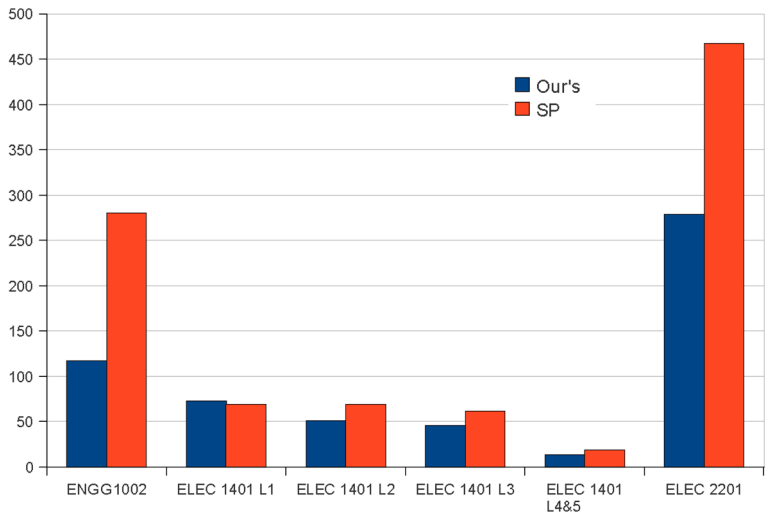

Fig. 6 Performance of our proposal against the shortest path approach

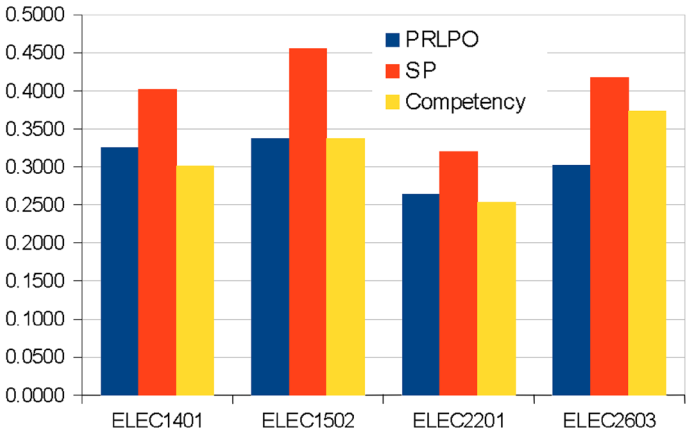

Fig. 7 Performance of our proposal against the shortest path (sp) and competency based approach

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. The results clearly show that our proposal outperforms the SP 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 SP 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.

Figure 7 gives the total violated distance of learning paths generated by our proposal against those generated by using the SP and competency based approaches. The results clearly show that our proposal consistently outperforms the SP approach by returning better learning paths in all cases. It is worth noting that the performance difference in ELEC1502 and ELEC2603 is significantly larger. This is probably due to the fact that the amount of knowledge encapsulated in a course concept is larger in ELEC1502 and ELEC2603 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 SP 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

Table 2 Performance with or without HC on ELEC1401

Cases	With HC	Without HC	% of Improvement
Fitness	0.9874	0.8721	13.22
Common rules	43.1	44	−2.05
Recomm. path # 1	2078.1	2151.9	−3.43
Recomm. path # 2	2087.2	2184.8	−4.47
Recomm. path # 3	2147.8	1847.7	16.24
Recomm. path # 4	930.7	914.4	1.78

systematic process of rule extraction as enhanced by our concept clustering technique, and thus significantly minimizing the search difficulty for the rule-based optimization as later performed by the genetic algorithm in such cases.

With conflicting views as opposite rules occurring in different experts' recommended learning sequences, our evolutionary optimizer may apply some filtering mechanism such as formulating a common set of rules among all recommended sequences or simply using the hill-climber (HC) to determine the quality of each individual recommendation to the overall search before biasing toward some recommended sequences.

Table 2 shows the averaged qualities of the learning paths generated by the HC over 10 runs on the fitness values, common rules or any of the 4 experts' recommended paths for ELEC1401. When working solely to optimize on the fitness values of the obtained learning paths with the HC, the percentage of improvement on the ultimate quality of the best learning path returned can be *13.22 %* that is fairly good when compared to that of the original evolutionary optimizer without the HC. In working according to the common set of precedence rules among all the recommended paths, the limited set of common rules may limit the search space of the HC heuristic, thus slightly degrading the overall performance possibly due to the randomized searches performed by the mutation or crossover operation when the total number of iterations exceeds the preset limit of *10,000*. Similarly, in following the precedence constraints/rules (Tsang, 1993) as extracted from the recommended learning path #1 or #2, which is mainly obtained from some external and online sources of course material, the overall performance of our evolutionary optimizer is also slightly degraded roughly by 3–4 % possibly due to the limited search spaces. On the other hand, when working according to the relatively less "restrictive" set of precedence constraints (Tsang, 1993; Marriott and Stuckey 1998) as extracted from the recommended learning path #3 or #4, that is essentially formulated from our lecture notes or the textbook adopted for ELEC1401, the opportunistic search as performed by the HC heuristic helps to improve the overall performance of our evolutionary optimizer to successfully return even more optimized learning paths of higher quality when compared to those of the original optimizer without the HC. Clearly, the largest gain as 16.24 % can be attained by the HC on the learning path #3 as formulated from our lecture notes of ELEC1401. To summarize, the improvement as possibly attained by the HC heuristic is apparently dependent on the 'quality' of the recommended learning paths with their precedence constraints/

rules to guide the overall search space of our evolutionary optimizer. In using the HC heuristic to try to enhance the evolutionary optimizer, it is always wiser to use several individual's recommended learning paths to avoid any bias into some limited search spaces for comparison, or simply focusing on optimizing the fitness values by the HC heuristic without referring to any recommended learning path in case of uncertainty on the 'quality' of the recommended learning paths.

Conclusions

In this paper, we propose to conduct a more complete e-learning system framework that carefully integrates semantic analysis (Chen et al. 2008), concept clustering (Guarino and Welty 2000) and learning path optimization. The refined concept correlation information through the concept clustering algorithm will then be passed to the rule-based evolutionary algorithm to optimize for better learning path(s). Moreover, we consider to improve the original evolutionary algorithm with the HC heuristic. Besides promoting the fitness values of newly generated learning paths during the evolutionary search, the hill climber (Tsang, 1993) can also be used to carefully select the more valuable reference path out of experts' recommendation.

To demonstrate the feasibility of our proposals, a prototype of our analyzer integrated with the concept clustering and the original or enhanced rule-based optimizer was implemented. The performance of both evolutionary optimizers was compared favorably against that of the benchmarking shortest-distance optimizer on various actual courses, thus showing the effectiveness of our proposed and complete e-learning system framework for learning path optimization. More importantly, our proposal clearly demonstrates the importance of enhanced ontology analysis, utilizing the appropriate heuristic and rule selection for the overall performance of adaptive or personalized e-learning systems (Lo et al. 2009) for learning analytics.

There are many possible directions for future exploration. First, it is attractive to investigate on how our proposed and complete e-learning framework of semantic analysis, concept clustering and learning path optimization can be easily integrated into existing e-learning systems to make them more adaptive for learning analytics. Essentially, after the key concepts/phrases can be extracted from each course module stored in the databases of existing e-learning systems possibly through utilizing some publicly available keyword extraction algorithms, our proposed framework of concept clustering, rule extractions and learning path optimization techniques can then be readily applied. Second, the possible uses of other heuristics and/or carefully designed rule selection schemes to improve the overall performance of the evolutionary optimizer is worth investigating. Last but not least, it is definitely very important and meaningful to conduct a thorough investigation on the pedagogic theory to support our proposed framework for learning path optimization, the potential educational applications and careful evaluations of their pedagogical impacts of our proposed e-learning system framework in various real-life classroom scenarios, subject areas and/or school settings that will make this work more complete, thus providing many practical and useful guidelines for the next-generation e-learning systems.

Acknowledgments The authors would like to thank Professor Kinshuk, Dr. Daniel Churchill, Professor Ricky Kwok and Professor Yi Shang for their fruitful discussions.

References

Acampora, G., Gaeta, M., Loia, V., Ritrovato, P. & Salerno, S. (2008). Optimizing learning path selection through memetic algorithms. In Proceedings *of the IEEE International Joint Conference on Neural Networks (IJCNN) 2008* (pp. 3869–3875), ISSN: 1098-7576, Hong Kong, China.

Baosteel Talent Development Institute. (2013). Baosteel e-learning system. Retrieved March 8, 2013, from http://www.cuhk.edu.hk/eLearning/c_systems/elearn/implementation.html.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*, 121–167.

CityU Enterprise Ltd. (2009). Personalized education network. Retrieved March 2, 2010, from http://pen.actin-education.hk//login/login.jsp.

Chen, C. M. (2009). Ontology-based concept map for planning a personalized learning path. *British Journal of Educational Technology, 40*(6), 1028–1058.

Chen, C.M., Peng, C.J. & Shiue, J.Y. (2008). Ontology-based concept map for planning personalized learning path. In Proceedings of the *IEEE Cybernetics and Intelligent Systems* (pp. 1337–1342). ISBN: 978-1-4244-1673-8, Chengdu, China: The IEEE Computer Society.

CU eLearning Development Team. (2013). CU e-learning system. Retrieved March 7, 2013, from http://www.cuhk.edu.hk/eLearning/c_systems/elearn/implementation.html.

Fung, S.T., Tam, V. & Lam, E.Y. (2011). Enhancing learning paths with concept clustering and rule-based optimization. In Proceedings of the *IEEE International Conference on Advanced Learning Technologies (ICALT)*, (pp. 167–171).

Guarino, N. & Welty, C. (2000). Conceptual modeling and ontological analysis. In Proceedings of AAAI-2000. Retrieved August 10, 2000, from http://www.cs.vassar.edu/weltyc/aaai-2000/.

Lo, W.S., Chung, I.C., & Hsu, H.J. (2009). *Using ontological engineering for computer education on online e-learning community system.* In *Proceedings of the IEEE International Conference on Education Technology and Computer (ICETC)* (pp. 167–170). The IEEE Computer Society.

Malmberg, A., & Maskell, P. (2002). The elusive concept of localization economies: towards a knowledge-based theory of spatial clustering. *Environment and Planning A, 34*(3), 429–449. The Pion Publication.

Marriott, K., & Stuckey, P. J. (1998). *Programming with Constraints: an Introduction*. Cambridge: The MIT Press.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools, 13*(1), 157–170. The World Scientific Publishing Company.

Monroe College eLearning Development Team. (2012). Monroe College eLearning Mobile System. Retrieved May 8, 2012, from http://elearning.monroecollege.edu/default.asp.

PPDH eLearning Development Team. (2011). PPDH eLearning System. Retrieved April 20, 2011, from http://iportal.ppdh.org.hk/.

Tam, V., Lam, E.Y. & Fung, S.T. (2012). Toward A Complete e-Learning System Framework for Semantic Analysis, Concept Clustering and Learning Path Optimization. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT)*, (pp. 592–596).

Tsang, E. (1993). *Foundations of Constraint Satisfaction*. Amsterdam: University of Essex. The Academic Press Publishing.

Wong, L. & Looi, C. (2009). Adaptable Learning Pathway Generation with Ant Colony Optimization. *Journal of Educational Technology and Society*, 12(3), pp. 309–326. Belmont, CA: Wadsworth Publishing Company.

Vincent Tam completed his Ph.D. in computer science from the University of Melbourne, Australia in 1998. Dr. Tam was the winner of Innovative Teaching Award (2000) in the School of Computing, National University of Singapore, and also the Best Teacher Award (2011) in the Faculty of Engineering, University of Hong Kong. He is a Principal Lecturer and Honorary Assistant Professor in the University

of Hong Kong, the part-time deputy director of an EMB-funded project on e-learning packages, and also the principal investigators of various teaching development projects.

Edmund Y. Lam received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University. He is now a Professor in Electrical and Electronic Engineering at The University of Hong Kong, and the Director of its Computer Engineering program.

S. T. Fung completed his M.Phil.(Eng) degree from the University of Hong Kong in 2009–2011. He is interested in the use of semantic analysis for e-learning systems.