Title	Toward a complete e-learning system framework for semantic analysis, concept clustering and learning path optimization
Author(s)	Tam, V; Lam, EYM; Fung, ST
Citation	The IEEE 12th International Conference on Advanced Learning Technologies (ICALT 2012), Rome, Italy, 4-6 July 2012. In Proceedings of the 12th ICALT, 2012, p. 592-596
Issued Date	2012
URL	http://hdl.handle.net/10722/165160
Rights	Creative Commons: Attribution 3.0 Hong Kong License

2012 12th IEEE International Conference on Advanced Learning Technologies

Toward A Complete e-Learning System Framework for Semantic Analysis, Concept Clustering and Learning Path Optimization

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

***Abstract*—Most online e-learning systems often demand the pre-requisite requirements between course modules and/or some relationship measures between involved concepts to be explicitly inputed by the course instructors so that an optimizer can be ultimately used to find an optimal learning sequence of involved concepts or modules for each individual learner after considering his/her past performance, learner's profile, learning style, etc. However, relying solely on the course instructor's input on the relationship among the involved concepts can be imprecise possibly due to the individual biases by human experts. Furthermore, the decision will become more complicated when various instructors hold conflicting views on the relationship among the involved concepts that may hinder any reasonable deduction. Therefore, we propose in this paper a complete system framework that can perform an explicit semantic analysis on the course materials, possibly aided by the relevant Wiki articles for any missing information about the involved concepts, to formulate the individual concepts, and followed by 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 possible advantages of our proposal with many possible directions for future investigation.**

***Keywords*-concept clustering; learning objects; learning styles; learning path optimization.**

I. Introduction

Most online e-learning systems [1], [2] allow the learners or students to specify their own learners' profiles and learning styles. However, these e-learning systems often demand the pre-requisite requirements between course modules and/or some relationship measures between involved concepts to be arbitrarily and explicitly specified by the course instructors such that an optimizer will be ultimately employed to find an optimal learning seqeunce of involved concepts or modules for each individual learner after considering his/her past performance, learner's profile, learning style, and relevant learners' profiles [1]. However, there are several pitfalls in solely relying solely on the course instructor's input on the relationship among the involved concepts possibly due to the individual biases by human experts. Furthermore, the decision will become more complicated when various instructors hold conflicting views on the relationship among the involved concepts that may sometimes contain incomplete information in the form of learning objects [3] for any plausible logical deduction.

There were some previous works that were focused on using statistical or machine learning approaches for the semantic analysis [1] of the relevant course materials and/or the ultimate optimization [3] of learning paths or sequences of involved course concepts or modules. For instance, Wong and Looi [4] 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 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 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 multiple reference paths possibly with conflicting views on the pre-requisite requirements of the course concepts or modules as provided by various human experts. Therefore, we propose in this paper a complete e-learning system framework that can perform an explicit semantic analysis (ESA) [3] 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

978-0-7695-4702-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ICALT.2012.66

prototype, we implemented a prototype of the proposed e-learning system framework. Our empirical evaluation clearly revealed the possible strengths of our proposal. There are many possible directions for future investigation including the application of our proposed complete and comprehensive system framework to develop an effective e-learning system for foreigners to learn about the recognition and/or writing of Chinese characters since most Chinese characters can be broken up into basic structures representing components, concepts or items of daily living. 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 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. Section IV discusses about our rule-based and evolutionary optimization method 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 [5] 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.

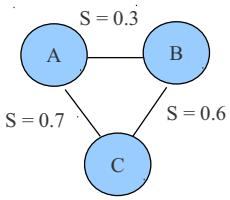

Figure 1. Illustration of multiple knowledge requirement.

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

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

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 [1], [5] 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 inferencing 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 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.

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 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 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

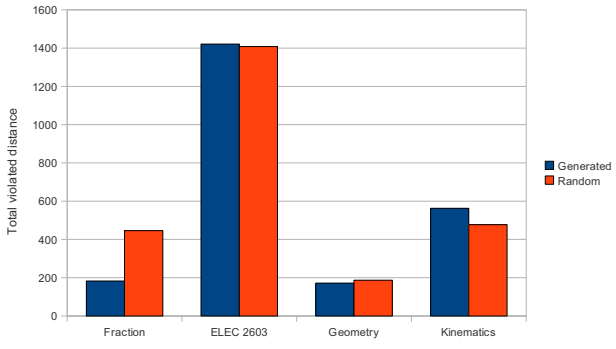

Figure 2. Comparison of learning paths generated by statistical analysis and randomly generated learning paths.

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 category 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 nxn 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 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 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 in the subsequent paragraphs.
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.

IV. A Rule-Based and Evolutionary Optimization Approach

The evolutionary algorithm we adapted 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 evolutionary 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 $i \ldots j$, where $i < j$, will be swapped between the two chromosome $\mathbf{X}$ and $\mathbf{Y}$.
- Mutation operation: the mutation operation randomly selects two indice 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 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

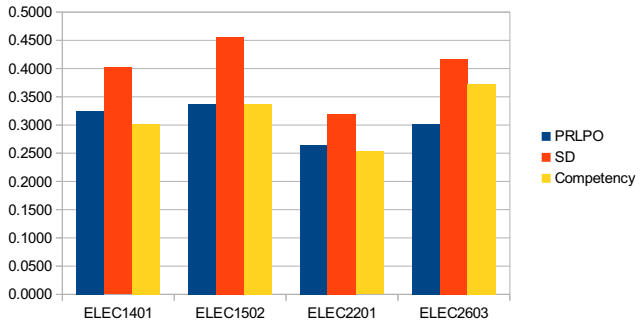

Figure 3. Performance of learning paths generated by our proposal (Ours) against those generated by the shortest distance (SD) and another competancy based approach.

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:

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

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 3 gives the total violated distance of learning paths generated by our proposal against those generated by using the SD and competancy based approach. The results clearly show that our proposal consistently outperforms the SD 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 SD 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 thus 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 e-learning system framework that carefully integrates semantic semantic analysis, concept clustering and learning path optimization. The refined concept correlation information through the concept clustering algorithm 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 analyser enhanced with concept clustering and rule-based optimizer was implemented. Its performance was compared favorably against the benchmarking shortest-distance 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 [2] yet can be easily integrated into such e-learning systems.

Acknowledgment

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

References

[1] 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.

[2] 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.

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

[4] L. Wong and C. Looi, "Adaptable Learning Pathway Generation with Ant Colony Optimization," *Educational Technology and Society*, vol. 12, no. 3, pp. 309–326, 2009.

[5] 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.

[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.