

A Study of Problem Difficulty Evaluation for Semantic Network Ontology Based Intelligent Courseware Sharing

Javed I. Khan, Manas Hardas, Yongbin Ma

Networking and Media Communication Research Labs

Computer Science Department, Kent State University, Kent, Ohio, 44240

javed|mhardas|yma@cs.kent.edu

Abstract

Testing and evaluation is an integral part of the learning process. Educators have often tried to devise methods for design of test-ware. Intelligent design and compilation of test-ware is a very interesting problem with immense applications. This research aims at automating the process of intelligent design of test-ware by providing qualitative assessment of questions. In this attempt, we provide some synthetic parameters for the evaluation of question in its concept space. The parameters are tested in some real world scenarios and intuitive inferences are deduced predicting the performance of the parameters. It is observed that the difficulty of a question is often a function of the concepts it tests. Concept knowledge can be represented in the form of linked concepts in semantic nets, the links representing the relationships between the concepts. If this directed graph is known, complexity of a question can be computed by synthetic means.

Keywords: test-ware digital library, automatic composition, semantic web.

1. Introduction

At the core of any intelligent design and sharing activity lies a system for assessment of seminal attributes and properties of the elements of design. A design process is guided by these assessments towards a specific goal. Various designers may choose to compose differently to meet his/her creative urge- but an essential element of any design process is an assessment system. In human act of design such as composition of poetry, or painting these assessments are considered to be result of extremely sophisticated cognitive ability. Engineering design uses assessment systems which are little more formal and are based on scientific properties. More recently the goal has shifted for machine design using information as elements. Without a machine computable assessment system it is impossible to program an automated designer system. Currently the web is huge repository of assorted digital resources without much reusability. A problem/question is one type of sensitive test-ware resource. Most educational content is scattered, replicated and not linked to each other by any

kind of relationship. Unfortunately, such scattered information is not easy to use even humanly and particularly useless for automatic use. The recent development in semantic web technology and standards like RDF and OWL is now gradually changing the scenario. It seems that finally the technology is available to encode the knowledge required to make design possible. In this backdrop, in this paper we present an interesting research which now attempts qualitative assessment of problem with a particular objective towards facilitating automatic test composition.

It is interesting to note, that there have been previous attempts to quantify the complexity of problems [2,3,4] though not from automatic design perspective. Some tried to figure out cognition based solution to the “question complexity” problem and understanding student view of question complexity [1,2]. Whilst these works propose fine methods to evaluate complexity of a question, either they are rendered incomputable due to over-generalization of representation or due to over-dependence on external factors. We propose an assessment system which is based on semantic knowledge space, applicable universally and above all which is machine computable.

2. Proposed Approach

Digital test-ware resources are now abundant the web. Complex questions can be decomposed into smaller simpler basic questions. These simple questions always connect to a few concepts from the course ontology. It is important to see what are the criteria against which a problem is selected for test design? A question should not only be comprehensive and diverse but also be relevant to the topics taught. It should be capable of testing varying student populations. Unfortunately, currently it is very difficult to assess any of the above suitability going into test-ware collections- even by human educators. Why it is so? The answer actually is that the questions are not accompanied by the conceptual space they are composed in. On the web questions are singular elements with not much meaning. Thus the qualitative assessment of question in their concept space is a very important step in making online testing, e-learning or web based pedagogy even remotely effective. Semantic Web Standards like RDF

and OWL give a convenient platform for integration and sharing of metadata. RDF and OWL provide means for greater machine interpretability of content. This creates a real opportunity to design systems that can exchange and use complex design resources such as test-ware. In our approach we assume a semantic network mechanism which allows one to represent the context in a standard and sharable way is given. We then isolate the main pedagogical challenge as finding few measurable quantities that can provide guidance in the process of automatic design. We have recently experimented with few such guidance parameters. We have also tested few of these parameters in one of the real world courses being offered on students. The performance of the students is observed and used in determining the effectiveness of these parameters in predicting the difficulty of a problem. In this paper we share evaluation of two such parameters.

3. Semantic Knowledge Space

Topic Dependency Graph $T(C, L)$ is a projection of a semantic net for a course, with vertices C and links L where each vertex represents a concept and each link with weight $l(i,j)$ represents the semantics that c_j is a prerequisite for learning c_i , where $(c_i, c_j) \in C$.

A TDG is further associated with a weight system. The self-weight $W_s(i)$ represents the relative semantic importance of the root topic itself with respect to all other prerequisites. The prerequisite weights $W_p(i)$ represent relative semantic importance among the prerequisite topics. A TDG with root A is represented as T(A) in Fig-1. The semantic computability is built on the following properties of TDG. The prerequisite dependency is transitive, asymmetric, and inverse of *post-requisite*. For any node

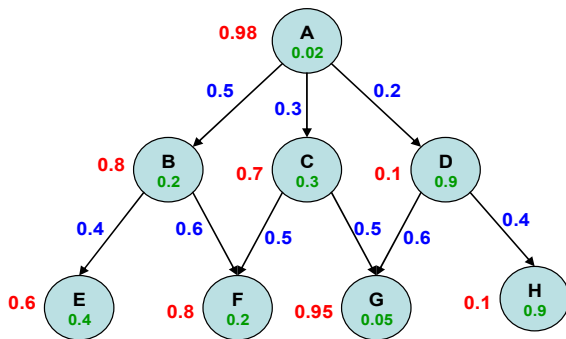


Figure 1. TDG rooted at node A

in the TDG, the sum of self-weight and prerequisite weights and the sum of the prerequisite link weights to its child node set are both always 1. A generalized TDG can be vast. Therefore we define a pruned sub-graph called as Projection graph which cuts the computation based on a limit on propagated semantic significance called as *threshold coefficient* (λ). As the Projection

graph is a sub-graph of the TDG, it is necessary to design the TDG such that the leaf nodes have pre-requisite weights. Flexibility for optional pre-requisite weights for the leaf nodes allows the TDG to be extensible and easily extractable.

Projection Graph (P): Given a root concept x_0 and a projection threshold coefficient λ , and a TDG, $T(C,L)$, a projection graph $P(x_0, \lambda)$ is defined as a sub graph of T with root x_0 and all nodes x_i where there is at least one path from x_0 to x_i in T such that node path weights $\eta(x_0, x_i)$ satisfies the condition: $\eta(x_0, x_i) \geq \lambda$

Node path weight is the product of the link weights and the prerequisite weights for all in nodes in a path between the two subject nodes, including self-weight value of the subject node. The concept of projection allows working on a subset computable graph of the TDG with desired semantic depth and significance. By varying the threshold coefficient the size of the workable graph, i.e. projection, can be changed.

4. Test Design and Evaluation

In the general direction of automating the process of design, composition and evaluation, development of fully automated expert system can be a useful application. E-rater at ETS has experimented with automated evaluation of answer [5]. Bulk of current research however have been performed without the advantage of underlying knowledge space and are based on external attributes such as surface syntactic analysis or psychological reaction parameters such as perceived difficulty ratings provided by students, time to answer etc. Unfortunately many of these do not help in design. It seems that the augmentation of a semantic network and incorporation with the test and problems can open up a whole new dimension and opportunity in computer assisted evaluation and testing.

The TDG gives the layout of the course in the concept space also specifying the course organization, involved concepts and the relations between the concepts. The basic unit of a test can be identified as a problem. The problem can be of various types and each and every problem connects to a specific set of concepts from the semantic net ontology. It means that, to answer a problem a certain set of concepts from the semantic net are required. Thus the evaluation model's basic unit is a problem which connects to a concept set from the ontological representation of the course. This is called as question to concept mapping. The concepts which are must requirement are linked by "AND" relation while those which are not imperatively required are shown by "OR" relationships. Our challenge is to quantify the amount to which the problem tests the concepts individually and with respect to the ontology root.

4.1 Problem Evaluation Parameters

A basic question in machine computable test and evaluation design is the problem complexity estimation-what constitutes to the complexity of a problem? Given the semantic connection of the problems we propose some parameters for guiding the evaluation process. The concept set is the input to evaluate the problem evaluation parameters; hence the concept set has to be precise and methodically selected.

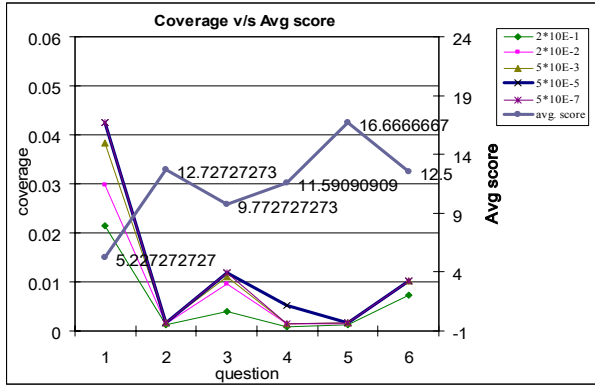


Figure2. Coverage plot for each question for varying λ

Coverage (α): Coverage of a node is the sum of the node path weights of all the nodes in its projection graph, when projected to the ontology root. It is an approximation of the number of prerequisite concepts involved to understand and answer a concept and the extent to which they are important to the root nodes understanding.

Coverage of a node x_0 with respect to the root node r is defined as the product of the sum of the node path weights of all nodes in its projection set $P(x_0, r)$ given by $[x_0, x_1, \dots, x_n]$; and the incident path weight $\gamma(r, x_0)$ from the root r .

$$\alpha(x_0) = \gamma(r, x_0) * \sum_{m=0}^n \eta(x_0, x_m) \quad \dots(1)$$

Where node path weight for a node to itself is its self-weight $\eta(x_0, x_0) = W_s(x_0)$; and $\gamma(r, x_0)$ is the incident path weight from root r to node x_0 , which is same as node path weight excluding the factor of self-weight of the subject node.

Diversity (Δ): Diversity tests the breadth of knowledge domain required to answer particular question. A question is attributed high diversity value if the concepts it tests are distinct in the context of knowledge space. Diversity is factor of the uncommon concepts rather than the common concepts because the disparate concepts are the ones that ascribe diversity to question. If the projections of some of the concepts overlap with each other, i.e. they have some concepts in common;

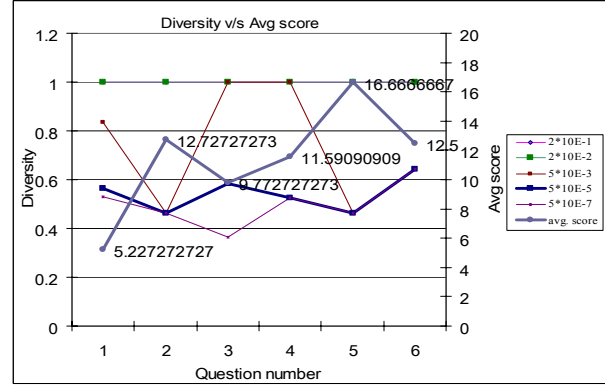


Figure3. Diversity plot for each question for varying λ values.

means that they are less diverse as both indirectly depend on some common ground for their complete understanding. Whereas when no two concepts are common it means that, the question has high diversity.

Diversity is formally defined as “the ratio of summation of node path weights of all nodes in the non-overlapping set to their respective roots, and the sum of the summation of node path weights of all nodes in the overlap set and summation of node path weights of all nodes in the non-overlap set.”

$$\Delta = \frac{\sum_{m=1}^p \eta(i, N_m^i)}{\sum_{m=1}^q \eta(j, O_m^j) + \sum_{m=1}^p \eta(i, N_m^i)} \quad \dots(2)$$

where $\forall i, j \in \text{Concept set}$; N is the non overlapping set while O is the overlapping set and p & q are their respective number of set members. Concepts common to two or more projections are considered in the summation of the node path weights of overlapping set while the un-common ones are considered in the non overlapping set.

5. Analysis

The problem evaluation parameters are tested by subjecting them to a real world scenario of test evaluation. Based on course ontology a test is generated by concept extraction from the TDG and the performance of the students for each of the questions is recorded. The average score per question is the subjective variable along with the parametric calculations. Consequently it is observed that for a more difficult question, i.e. one with high coverage and diversity, the average points scored are less.

5.1 Performance Analysis v/s Average Score

Coverage and diversity analysis is shown in figure 2, 3. Both parameters show exact inverse relationship with average score. As the value of parameters increases the average score decreases and vice versa. Both the parameters are functions of the threshold coefficient. As

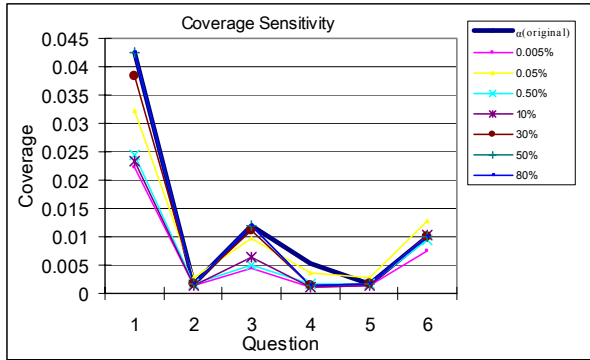


Figure4. coverage sensitivity analysis for varying link weights

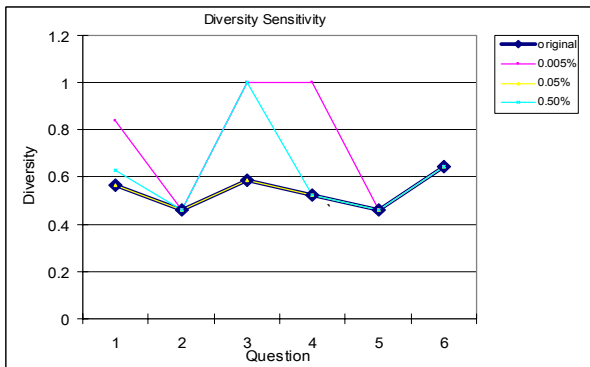


Figure5. Diversity sensitivity analysis for varying link weights

λ decreases, the projection graphs for the concepts increases and so does the coverage. However for diversity, increase in the projection doesn't always mean increase in its value because as projection increases, more concepts begin to overlap thus decreasing the diversity or simply more uncommon concepts are added to non overlap set, consequently decreasing the diversity.

Nonetheless coverage and diversity both follow their exact inverse relationship with average score for a particular middle range of λ value, above or below of which the graphs don't show exactly inverse behavior.

5.2 Sensitivity Analysis

Semantic network systems are generally prone to subjectivity in various parts of its design. The subjectivity in the proposed system lie in its link weights assignments. Course ontology can be envisioned as a collaboratively developed knowledge repository in the future. Many authors and educators can contribute to the global ontology through their individual inputs and comments. This leads to a cumulative knowledge map, which is continuously amended and edited by various authors. In this view the evaluation parameters are subject to change and thus need to be elastic enough to resist radical discontinuity and instability. We put to test the stability of the systems against random yet gradual change in the subjective parameters- the link weight values. Whenever a node is added or deleted from course ontology the link weights change, thus varying

link weights are varied randomly beyond a certain percentage by the equation,

$$l_{(i,j)}^{new} = l_{(i,j)}^{old} * (1 + M.R) \dots(3)$$

Where, M is the strength of the disturbance, and R is the randomness induced above the percentage. The performance of the parameters for link weight sensitivity tests are shown in Figure 4 and 5. the parameter calculations. The

6. Conclusions

Courseware design and sharing is a complex process requiring cognitive precision presently only capable of human mind. Semantic web standards present a way to represent meta information about a knowledge repository which makes it possible to take productive steps towards achieving this process automatically and intelligently. Automatic intelligent test design is subset of this process. We propose few synthetic parameters for evaluation of a problem in a test using the semantic knowledge associated with the problem. The parameters are also analyzed for performance by creating a synthetic test from a pool of Meta knowledge space, and recording the performance of student on that test. It is observed that the parameters perform very well under certain conditions. The research has been pursued as a side project of a NSF funded research Grant #0333520, under its Digital Library Initiative.

7. 5. References

- [1] Croteau, E., Heffernan, N. T. & Koedinger, K. R. (2004) Why Are Algebra Word Problems Difficult? Using Tutorial Log Files and the Power Law of Learning to Select the Best Fitting Cognitive Model. 7th Annual Intelligent Tutoring Systems Conference, Maceio, Brazil.
- [2] Kuo R., Lien, W-P., Chang, M., Heh, J-S, Difficulty Analysis for Learners in Problem Solving Process Based on the Knowledge Map. International Conference on Advanced Learning Technologies, 2003, 386-387.
- [3] Li, T and S E Sambasivam. Question Difficulty Assessment in Intelligent Tutor Systems for Computer Architecture. In *The Proceedings of ISECON 2003*, v 20 (San Diego): §4112. ISSN: 1542-7382.
- [4] Lee, F.-L, Heyworth, R., Problem complexity: A measure of problem difficulty in algebra by using computer. Education Journal Vol 28, No.1, 2000.
- [5] Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. AI Magazine, 25(3), 27-36.