Quantifying Concept Proximity Based on Semantic Measures

1R. Sunitha and 2G. Aghila

1Pondicherry University, Department of Computer Science, Puducherry, India.
2National Institute of Technology Puducherry, Department of Computer Science and Engineering, Karaikal, India.

ABSTRACT

The selection and provision of appropriate learning concepts and the nature of content, matching the self-directed learner’s needs are the fundamental issues in adaptive E-learning systems. Studies mainly focus on qualitative approaches for providing such personalized solutions. Harnessing quantitative measures for building personalized e-learning solutions guarantees promising result. In this work, the knowledge space and hyperspace of educational materials have been symbolized using a three layered structure which in turn is formally represented using ontology. A set of measures espousing semantic relatedness and semantic distance viz. Thematic Distance Measure (TDM) and Thematic Proximity Measure (TPM) using the structural ontology has been proposed. The measures have been used in the learning sequence generation (LSG). The experimental results proves that the approach facilitates quick and coherent learning of the course under study for a self-directed learner.

INTRODUCTION

With the proliferous growth of Internet, e-learning has become a popular web application and number of developments has been made in the field of e-learning to provide better learning experience to the learners. Major works concentrate on providing flexible solutions tailored for each individual learner (Alekseandria et al., 2011; Triantafillou et al., 2002). On the other end, works (Wong et al., 2012; Manolis et al., 2013) that focus on integrating the formal learning theories and teaching strategies in e-learning context, thereby enhancing the technology through a strong theoretical base are also being carried out. Some works (Kontopoulos et al., 2008; Heiyanthuduwage et al., 2006) focus on the management of learning resources, the so-called Learning Object (LO) management providing ontological solutions for representing LOs thereby enabling higher level reasoning about LOs. The development of cognitive tools for supporting e-learning is another dimension of research (Greer et al., 2000). In adaptive e-learning, the significant component of e-learning considered for adaptation is the learning path. Learning path is defined as an organized sequence of learning activities that progresses the learner toward the designated learning goal (Colace et al., 2005). As the learners’ needs and ability varies, mechanisms for adapting learning paths for efficient learning experience have been reported in literature (Colace et al., 2005; Tung-Cheng et al., 2010). These works tailor the learning path statically by acquiring the learner needs before the learning process begins (so-called adaptable e-learning) or adapt the learning path dynamically during the learning activity based on the learners’ performance and other parameters (Stash et al., 2003) (so-called dynamic/context-aware adaptive e-learning). There are two aspects in learning path construction; the former being the selection of learning concepts that forms the learning sequence and the latter is the selection of the appropriate content that suits the individual learner’s preference. Thus the general mechanism for learning path construction involves the selection and sequencing of relevant learning objects which serve as the building blocks of a learning path. In this work we are adopting an incremental strategy based on a set of semantic measures to build a learning sequence with an objective of enabling the self-directed learner with quick and coherent learning of concepts, given the current cognitive state of the learner and his target learning concept. A three layered structure for representing the knowledge and hyperspace of educational contents and their connections has been proposed in this work which is formally represented using ontology. A semantic relatedness measure coined as “Thematic Proximity Measure” (TPM) based on a distance measure...
namely “Thematic Distance Measure” (TDM) has been proposed for the automatic provision of the suitable concept to be learnt in the sequence which aids in achieving the objective of quick and coherent learning by a self-directed learner. The layered structure also facilitates the assessment of the learner’s current learning state.

We have demonstrated that the definition and use of the proposed TPM in an e-learning system provides a better learning experience to the learners. The rest of the paper is organized as follows. The related works have been presented in section two. In section three, the layered structure for concept representation along with the definitions and mathematical formulations of TPM and TDM have been presented. Section four discusses about the proposed measures have been harnessed in the learning sequence generation. Section five presents our conclusions and discusses about future works.

**Related Works:**

In (Greer et al., 2000) a knowledge structure called the concept-topic structure to represent the domain of learning has been proposed. The topic structure facilitates the understanding of the sequence of learning and hence the nodes in this structure are related using prerequisite and temporal links. The concept structure facilitates the understanding of teaching goals. The concepts are related using the abstraction, aggregation, causal, analogy and prerequisite links. The proposed Topic-Concept-Instance structure is a variation of this work which includes additionally the representation of the knowledge about the concepts present in the concept level structure.

This work focuses on quantifying the degree of relatedness between two concepts in a theme using a path based approach. The advantage of path based measures is their simplicity. Many seminal works have been carried out in quantifying the similarity between ontological concepts. In (Wu, Palmer, 1994) the depth of the two concepts and the depth of their Least Common Subsumer (LCS) have been used to calculate the similarity score between the two concepts.

The simplest path based measure is (Rada et al., 1989) in which the concept of semantic distance has been used. The semantic distance is computed as a function of node count and is given by the sum of nodes that lie in the shortest path between two concepts. In (Leacock, Chodorow, 1998), a similarity measure based on the path length and the maximum depth of the taxonomy has been proposed.

In (Hirst, St-Onge, 1998) a relatedness measure that is based on the nature of the path between two concepts has proposed. If the two concepts are connected by a path that is not too long and with minimum changes in direction, then it signifies high degree of relatedness between the concepts.

**Proposed Method:**

In this section, the Topic-Concept-Instance structure has been detailed. Based on the proposed Topic-Concept-Instance knowledge structure we have formulated two quantitative measures viz. Thematic Proximity Measure (TPM), Thematic Distance Measure (TDM) which is presented below.

**The Topic-Concept-Instance structure:**

In e-learning, the domain or theme is populated by enormous learning items aka concepts that vary in granularity. Granularity of a concept is measured in terms of their degree of details addressed by the concept. The concepts in the theme are further related to each other depending upon their semantic and didactic associations. The concepts which are less granular are highly cohesive. These concepts may also have multiple instances (LOs) which vary in terms of the complexity of their content, media in which they are communicated and the coverage or density level of their content. Concepts are paired by various associative relations that dictate their semantic and didactic dependencies. This ideology, as assumed, has been used in formulating a knowledge structure of the learning theme known as the Topic-Concept-Instance structure (see fig. 1). Let \( T \) represents the set of “\( p \)” topics, \( C \) represents the set of “\( m \)” concepts, and \( R \), the set of associative relations defined between pairs of concepts which is given by

\[
T = \{ t_1, t_2, \ldots, t_p \}
\]

\[
C = \{ C_1, C_2, \ldots, C_m \}
\]

\[
R = \{ r_{11}, r_{12}, \ldots, r_{mk} \}
\]

The associative relations are assigned weights intuitively based on their relative importance in contributing to the thematic understanding of the concepts and is given by \( W = \{ w_1, w_2, \ldots, w_m \} \). The items in the concept level have multiple instances given by

\[
CI = \{ \{ CI_{11}, CI_{12}, \ldots, CI_{1j} \}, \{ CI_{21}, CI_{22}, \ldots, CI_{2j} \}, \ldots, \{ CI_{mk}, CI_{m2}, \ldots, CI_{mj} \} \}
\]

with features described as \( F = \{ F_{11}, F_{12}, \ldots, F_{1j} \} \). This representation facilitates efficient provisioning of learning solutions to learners in an e-learning setting.

**Semantic Measures:**

Semantic relatedness between two concepts is defined as the kind of lexical or functional association between classes. In linguistics context, semantic relatedness is defined as a semantic similarity through the lexical relation of synonymy between two words. In e-learning context, a domain represents a collection of
concepts that pertain to a particular subject in this case, a theme. The information exhibited by the relations defined between the different elements in the theme, can be used in quantifying the degree of relatedness between the concepts. Thus a set of path based semantic measures exploiting the structural semantics exhibited by the ontology have been proposed to quantify the semantic relatedness and didactic nearness of the learning items namely the Thematic Distance Measure (TDM) and the Thematic Proximity Measure (TPM).

**Definition 1: Thematic Distance Measure (TDM):**

The Thematic Distance Measure (TDM) is a distance measure that uses the structural semantic of the ontology to compute the semantic distance between the two concepts. It quantifies the amount of conceptual gap that exists between the two concepts in a theme. The calculation of TDM involves the identification of the best path that exhibits the exact thematic link as defined by the associative relations dictating the didactic relation. The identification of the best path is done using the path_weightage parameter. The calculation of path_weightage requires the value of the path_length. The mathematical formulation of Thematic Distance Measure (TDM) between two concepts is given in Eq. 1.

\[
\tau_{C_1C_2} = \beta(\tau_1, \tau_2, ..., \tau_n)
\]

where \(\tau_1, \tau_2, ..., \tau_n\) are the weights of the \(n\) different paths that connects the given concepts \(C_1\) and \(C_2\). The function \(\beta(\tau_1, \tau_2, ..., \tau_n)\) returns the weight of the best path that contributes to the real understanding of the thematic relatedness between \(C_1\) and \(C_2\). The best path helps in the evaluation of the structural semantic significance between the two concepts which in turn is used to evaluate the degree of thematic relatedness between two concepts.

**Definition 2: Thematic Proximity Measure (TPM):**

The Thematic Proximity Measure (TPM) between the two concepts is defined as the measure of degree of didactic nearness between two concepts of a theme. The TPM serves as an indicator to how close a concept is to another while learning so that there is a smooth continuity in the learning process enabling the learner to immediately link and construct his/her own knowledge. The nearness or proximity in general between two entities may be assessed through the distance that separates them. Thus the TPM which indicates the degree of didactic nearness of two concepts in a theme can be measured using the TDM. Proximity, in general is inversely proportional to the distance. Thus \(TPM\) is inversely proportional to the \(TDM\). The mathematical formulation of Thematic Proximity measure \(TPM\) is given in Eq. 3.
As TDM is inversely proportional to TPM, it is understood that if the concepts are separated by a short distance then they are thematically closer or their proximity in turn is very high indicating more didactic nearness. A larger TDM is a cue for the lesser proximity between the concepts.

**Experiment and Result Analysis:**
In this section we will discuss about how the proposed knowledge structure and the semantic measures have been used to generate learning sequences for a course on Data Structures (see fig. 2). The Data Structure concepts as defined by ACM classification have been considered for building the ontology. Fig. 3 illustrates the screen shot of the ontology constructed using Protégé 4.2. The IEEE LOM metadata standard (Standard 2002) has been used to describe the structural aspects of the learning objects and the data type properties of the learning object instances. The relation refinement terms as described by the *kinds* characteristics of the Relations category of the IEEE LOM have been espoused in order to define the structural relations of the learning object. The Relation category of IEEE LOM standard provides a group of data elements to describe the link that refers a related learning concept. The various relation refinement terms are has_version, replaces, requires, has_part, references, has_format, is_based_on and their symmetrical counterparts. Excluding the symmetrical terms, the different relation terms can be classified as “referential” and “semantic” relations (Sanchez, Sicilia 2004). “Semantic” terms refer to information that could be used in automated content selection and delivery. The relations in this category viz. requires, references, is_based_on, has_part have been considered in the concept level structure.

**Algorithm 1: Path Weight Calculation**

**Input:** Two concepts $C_1, C_2$

**Output:** Weight of the all the paths between $C_1, C_2$

Begin
Let $P = \{p_1, p_2, ..., p_n\}$ be the set of “n” paths between $C_1, C_2$

For all paths $p_i \in P$ do

{If $\text{length}(p_i) = 1$ then $S = w_{p_i}(c_1, c_2)$
else
Initialize weight $S$ of $p_i$ to 0
For each node $j$ along the path $p_i$
$S = S + w_{p_i}(c_j, c_{j+1}) / j$; // $c_j$ and $c_{j+1}$ are the successive nodes along the path $p_i$ and $w_{p_i}$ is the edge weight between nodes $c_j$ and $c_{j+1}$
}
$w_i = S$ // $w_i$ is the weight of the path $p_i$

Return $(W = \{w_1, w_2, ..., w_n\})$

End

Fig. 2: Screen Shot of the Data Structure Course Web Page.

TPM uses a path based approach to compute the thematic nearness. The path weight is an important criterion for determining the thematic nearness. The weight of the paths is computed from the weights of the edges representing the relations that lie along the path and the length of the path (see algorithm 1).
Table 1 summarizes the weights allocated intuitively to the refinement terms as specified by the IEEE LOM based on their relative importance in contributing to the thematic understanding of a concept. The larger the weight assigned to the relation, the lower is the indicated significance.

<table>
<thead>
<tr>
<th>Relationship Type</th>
<th>Description of relation</th>
<th>Relationship weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>requires</td>
<td>The described resource needs the referenced resource to support coherence of content</td>
<td>1</td>
</tr>
<tr>
<td>basedon</td>
<td>To make an idea the main part of something</td>
<td>2</td>
</tr>
<tr>
<td>partof</td>
<td>The referenced resource is included in the described resource either physically or logically</td>
<td>3</td>
</tr>
<tr>
<td>references</td>
<td>The described resource references, cites, or otherwise points to the referenced resource</td>
<td>4</td>
</tr>
</tbody>
</table>

Fig. 4 illustrates the concept level structure of the data structures ontology. A sample set of data structures concepts represented using English alphabets and the relationship (weights as assumed in table 1) that exists between them have been illustrated. Table 2 presents the best path, the corresponding TDM and TPM values between the pairs of concepts illustrated in fig. 4.

<table>
<thead>
<tr>
<th>Concept pairs</th>
<th>Concepts along the Best Path</th>
<th>TDM</th>
<th>TPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A,B)</td>
<td>AB</td>
<td>2</td>
<td>0.50</td>
</tr>
<tr>
<td>(A,C)</td>
<td>ABC</td>
<td>2.5</td>
<td>0.40</td>
</tr>
<tr>
<td>(A,D)</td>
<td>ABD</td>
<td>3</td>
<td>0.33</td>
</tr>
<tr>
<td>(A,E)</td>
<td>ABE</td>
<td>3.5</td>
<td>0.29</td>
</tr>
<tr>
<td>(B,C)</td>
<td>BC</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(B,D)</td>
<td>BD</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>(B,E)</td>
<td>BE</td>
<td>3</td>
<td>0.33</td>
</tr>
<tr>
<td>(D,C)</td>
<td>DEC</td>
<td>4.5</td>
<td>0.22</td>
</tr>
<tr>
<td>(E,C)</td>
<td>EC</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(D,E)</td>
<td>DE</td>
<td>4</td>
<td>0.25</td>
</tr>
</tbody>
</table>

The TPM value ranges from 0.22 to 1 representing a simple thematic reference and an absolute dependency in understanding the concept respectively. Given the topic-concept-instance structure representing the concepts, their granularity, the relations among them and the different instances of the concepts, algorithm 2 presents the steps in learning path generation.

Table 3 summarizes the performance of TPM based LSG and the standard LSG based on the formal didactic arrangement. The two approaches are compared in terms of the number of concepts (LO) to be learnt in the sequence, the total time (LT) required learning the concepts in the sequence and the learning efficiency (LE) score computed using time spent in learning and performance score. The results have been compared for 10 randomly selected sequences. The results presented in table 3 indicate that the average LE score is 4.3 in TPM based LSG and 4.48 in standard LSG. The average LE resulted from standard LSG is greater than that of the TPM based LSG, but the difference is trivial. But an increase of 58% on the time spent on learning has been
resulted from the standard LSG when compared to the TPM based LSG. This indicates that a TPM based LSG enables the learner to quickly acquire the knowledge without costing the learning efficiency.

**Algorithm 2: Learning Sequence Generation**

**Input:** Current learning concept X  
**Output:** Learning sequence adapted to the current learning state  
Step 1: Assess Learner’s state C = {c₁, c₂, ..., cₙ}  
Step 2: Repeat until goal satisfied.  
Step 3: Generation of the neighbourhood N = {n₁, n₂, ..., nₘ} of X using the TPM for the current learning concept  
Step 3: Generation of filtered, ranked item set F = {f₁, f₂, ..., fₙ} such that F = N ∩ C  
Step 4: The closest neighbour fᵢ is added as the next learning item in the sequence;  
C = C ∪ fᵢ  
Step 5: Goto step 2

**Table 3:** Comparison of the performance of TPM based LSG and standard LSG.

<table>
<thead>
<tr>
<th>Sequence Number</th>
<th>TPM based LSG</th>
<th>Standard LSG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of LOs</td>
<td>Avg. LE Score</td>
</tr>
<tr>
<td>S1</td>
<td>6</td>
<td>4.3</td>
</tr>
<tr>
<td>S2</td>
<td>5</td>
<td>4.73</td>
</tr>
<tr>
<td>S3</td>
<td>10</td>
<td>3.97</td>
</tr>
<tr>
<td>S4</td>
<td>7</td>
<td>4.52</td>
</tr>
<tr>
<td>S5</td>
<td>10</td>
<td>3.65</td>
</tr>
<tr>
<td>S6</td>
<td>4</td>
<td>4.67</td>
</tr>
<tr>
<td>S7</td>
<td>6</td>
<td>4.5</td>
</tr>
<tr>
<td>S8</td>
<td>8</td>
<td>4.21</td>
</tr>
<tr>
<td>S9</td>
<td>7</td>
<td>4.44</td>
</tr>
<tr>
<td>S10</td>
<td>9</td>
<td>4.01</td>
</tr>
</tbody>
</table>

The efficiency of any metric distance function δ between a concept pair is a non-negative value that should accord the following three axioms (Tversky 1977).

Minimality: δ(A, B) ≥ δ(A, A) = 0  
Symmetry: δ(A, B) = δ(B, A)  
Triangle Inequality: δ(A, B) + δ(B, C) ≥ δ(A, C)

The TDM as a distance measure has been found to accord the Minimality and Triangle Inequality axioms (illustrated in fig. 5). As we have ignored the symmetrical counterparts of relations, the axiom of symmetry is irrelevant in this context.

![Fig. 5: Satisfiability of the triangular inequality axiom by TDM.](image)

**Conclusion:**

In this work, we have proposed a structure to represent the knowledge of concepts in e-learning and two measures namely the “Thematic Proximity Measure” and the “Thematic Distance Measure” based on the proposed structure. The topic-concept-instance structure is efficient in the sense it helps in assessing the learning state of the learner as well as to ascertain the appropriate concepts that facilitate quick and coherent learning. The thematic proximity measure also had been proved efficient in sequencing the learning contents incrementally. The thematic distance measure has been found to satisfy the distance measure axioms. The proposed measures can also be applied in assessing the knowledge gap of the learners and predicting the concepts that that would assist the learners to fill this gap and also in support management.
REFERENCES


Anna Wong, Wayne Leahy, Nadine Marcus, John Sweller, 2012, "Cognitive load theory, the transient information effect and e-learning", Learning and Instruction, 22.


