Evaluation of Concept Importance in Concept Maps Mined from Lecture Notes
Computer Vs Human

Thushari Atapattu, Katrina Falkner and Nickolas Falkner
School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia

Keywords: Concept Map Mining, Concept Importance, Lecture Notes, Evaluation Methodology.

Abstract: Concept maps are commonly used tools for organising and representing knowledge in order to assist meaningful learning. Although the process of constructing concept maps improves learners' cognitive structures, novice students typically need substantial assistance from experts. Alternatively, expert-constructed maps may be given to students, which increase the workload of academics. To overcome this issue, automated concept map extraction has been introduced. One of the key limitations is the lack of an evaluation framework to measure the quality of machine-extracted concept maps. At present, researchers in this area utilise human experts' judgement or expert-constructed maps as the gold standard to measure the *relevancy* of extracted knowledge components. However, in the educational context, particularly in course materials, the majority of knowledge presented is relevant to the learner, resulting in a large amount of information that has to be organised. Therefore, this paper introduces a machine-based approach which studies the *relative importance* of knowledge components and organises them hierarchically. We compare machine-extracted maps with human judgment, based on expert knowledge and perception. This paper describes three ranking models to organise domain concepts. The results show that the auto-generated map positively correlates with human judgment (r_s~1) for well-structured courses with rich grammar (*well-fitted* contents).

1 INTRODUCTION

Concept mapping is recognised as a valuable educational visualisation technique, which assists students in organising, sharing and representing knowledge. Concept maps model knowledge so that it can be expressed externally using set of concepts and propositions (Novak and Gowin, 1984). These concepts are organised in a hierarchy with the most general concept at the top and the most specific concepts arranged below (Coffey et al., 2003). Based on the Assimilation theory (Ausubel et al., 1978), this externally expressible concept map is utilised to improve human learning, by integrating newly learned concepts and propositions into existing cognitive structures. Concept maps have been widely used in the educational context, particularly in identifying misconceptions and knowledge gaps, conceptual changes and being utilised as "advance organisers" (Novak and Gowin, 1984), and externalising mental models (Chang, 2007).

However, '*construct-by-self*', where students are responsible for creating their own concept maps, introduces a substantial difficulty for novice students to correctly identify concepts, relations and hence, requires continuous assistance from academic staff. A common alternative is to provide students with maps constructed by human experts (*expert maps)*, placing additional load and intellectual commitment on academic staff.

Although constructing a concept map for a lecture is a one-off process, it needs to be updated continuously, to cope with the changing nature of knowledge. However, due to the lack of human awareness of knowledge representations and a general preference for writing informal sentences over creating network models, concept maps are not yet widely used for learning.

Therefore, recent efforts in this area work toward semi- or fully automated approaches to extract concept maps from text (called *concept map mining*), with the aim of providing useful educational tools with minimal human intervention

Atapattu T., Falkner K. and Falkner N..
Evaluation of Concept Importance in Concept Maps Mined from Lecture Notes - Computer Vs Human.
DOI: 10.5220/0004842300750084
In *Proceedings of the 6th International Conference on Computer Supported Education* (CSEDU-2014), pages 75-84
ISBN: 978-989-758-020-8

(Olney et al., 2012); (Alves et al., 2002); (Chen et al., 2008). However, a significant problem in concept map extraction is the lack of an evaluation framework to measure the quality of machine-extracted concept maps (Villalon and Calvo, 2008). At present researchers rely upon human efforts to evaluate machine-extracted concept maps either through manual judgement or comparison with expert maps.

The majority of works in this area focus on the performance of automated tools using the popular metrics - *precision* and *recall*. These forms of measurement evaluate whether the machine extracted concepts and relations are *relevant*. However, in the educational context, particularly in course materials, the majority of knowledge presented is relevant to the learner, resulting in large part of lectures or textbooks being retrieved and identified for knowledge organisation (Atapattu et al., 2012). But, according to the definition of concept maps, a concept map should be an overview, which organises most important knowledge according to learning objectives (Novak and Gowin, 1984). Hence, the aim of this paper is to discuss a machine-based evaluation technique which studies the *relative importance* of knowledge, focusing beyond the simple measure of *relevancy*.

Current instructional methods widely support verbal learning through linear and sequential learning materials. The literature provides inadequate research to assist transforming linearity of resources into network models such as semantic networks and concept maps. Our approach takes the work that has already been invested in producing legible slides and focus on extracting useful knowledge that are beneficial for both the teacher and the learner. This will be an increasingly important research topic in the decade of Massive Open Online Courses (MOOCs). This paper provides a concise overview of our concept map extraction approach using Natural Language Processing (NLP) algorithms.

In this paper, we hypothesize that the natural presentation layout, linguistic or structural features might influence the human expert's judgement of relative concept importance. We developed three ranking models: 1) Baseline methods which use the natural layout of lecture slides (e.g. titles are the most important, sub-points are the least important); 2) Linguistic features such as grammatical structure of English text; and 3) Structural features such as proximity, number of incoming and outgoing connections, and degree of co-occurrence. We compare each of these models with human judgement using Spearman's ranking correlation coefficient (r_s). According to the results (Section 5), outcome of the structural feature model positively correlates with human judgment. There is a strong correlation ($r_s > 0.7$) for well summarised courses with rich grammar (i.e. *well-fitted* content). The correlation ranges from *well-fitted* to *ill-fitted* proportionally with respect to the quality and structure of the content. Lecture notes with some potential issues, including excessive information, category headings (e.g. key points, chapter 1), confusing visual idioms and ambiguous sentences (i.e. *ill-fitted* content) result in poor machine interpretation and hence, poor correlation with human judgement.

The concept map extraction, particularly from course materials, is beneficial for both students and educators. It organises and represents knowledge scattered throughout multiple topics. These maps can be used as an assessment tool (Villalon and Calvo 2008, Gouli et al., 2004) to identify understanding about concepts and relations. Additionally, these concept maps can be used as an "*intelligent suggester*" to recommend concepts, propositions, and existing concept maps from the web (Leake et al., 2004). In the educational context, these maps can provide scaffolding aid for students to construct their own concept maps. Students learn better when they are encouraged to fill in blank links (relations) rather than blank nodes (concepts) (Maass and Pavlik, 2013). Concept mapping has also been utilised widely in question generation (Olney et al., 2012) and question answering (Dali et al., 2009). The preliminary concept maps extracted from this research can also be extended as an ontology for domain modelling in intelligent systems (Starr and Oliveira, 2013).

This paper includes a background study of various concept map mining evaluation techniques in Section 2. In Section 3 and 4, we discuss about our core research of concept map mining from lecture notes and ranking model respectively. We evaluate our approach with human experts and present results and analysis in Section 5 and our study is concluded in Section 6.

2 RELATED WORK

The evaluation of the quality of machine-extracted knowledge representations is a challenging and tedious task. This can categorised into three dimensions as structural, semantic and comparative evaluation (Zouaq and Nkabou, 2009). In the

concept mapping perspective, assigning scores to extracted elements such as concepts and relations can be classified as structural evaluation. In a traditional scoring system, 1 point is assigned for a valid proposition, 5 points for each level of adopted hierarchies, and 10 points for cross-links (Novak and and Gowin, 1984). Although, the scoring technique provides information about creator's knowledge structure, this technique is time-consuming when assessing large-scale maps (Coffey et al., 2003).

In semantic evaluations, human experts are involved in judging the validity of machine-extracted maps. These types of studies are affected by the subjective judgment of human experts. Therefore, an average agreement among participants (*inter-rater agreement*) is compared with *human-to-machine agreement*. Generally, machine extractions are acceptable when *human-to-machine agreement* is equal or higher than *inter-rater agreement* (Hearst, 2000). Other research utilises expert-constructed maps as a gold standard to compare with machine-extracted maps (Villalon and Calvo, 2008). It is uncertain of the objective behind generating concept maps from computer algorithms in the presence of already constructed expert maps.

In comparative analysis, the machine-extracted concept maps are compared with other tools, which are built for the same purpose and test using the same corpus. TEXT-TO-ONTO is a popular ontology extraction tool. It is compared with TEXCOMON (Text-Concept map-Ontology) that automatically extracts concept maps from text (Zouaq and Nkabou, 2009). In order to use the comparative evaluation, other tools should exist which are built for same purpose. We demonstrate our approach using Microsoft PowerPoint Framework (as a commonly used lecture note format), although our approach is not constrained to PowerPoint but generalises across any common lecture note formats such as OpenOffice, Latex, and Apple Key note with a structured template for headers and text. To the best of our knowledge, there are no existing tools which do this.

However, despite the benefits to the educational context, state of art studies focused on *concept existence*, and not their *relative importance*. Our work adapts several structural features (e.g. proximity, incoming and outgoing links) (Leake et al., 2004) and graph-based metrics (e.g. degree) (Zouaq et al., 2012) to rank the concepts according to their importance. However, we also use linguistic features, semantic information and the association between terms to mimic the human judgment using machine algorithms. This resolves syntactically and semantically incomplete information in lecture notes which recognised as a key challenge in applying computer algorithms to semi-structured lecture notes.

3 CONCEPT MAP EXTRACTION

Our core research focus is on extracting useful knowledge as concept maps (*concept map mining*) from educational materials, particularly from lecture notes. Current concept map mining techniques rely upon informational retrieval techniques (e.g. vector space model, C-value/ NC-value), linguistic-based approaches (e.g. part-of-speech tagging, language models) or hybrid models (Frantzi et al., 2000). Information retrieval approaches suffer from probable semantic loss. Although linguistic-based techniques address this issue and extract nouns as semantic concepts, nouns may be present that are not semantic concepts in that particular domain. In order to overcome these issues, studies based on linguistic techniques utilise external dictionaries and thesaurus. However, these types of external resources are very limited for specific domains such as Computer Science.

Therefore, our work utilises NLP algorithms to extract concepts, relations using syntactic parsing and part-of-speech tagging. We rank extracted concepts using statistical features such as term frequency, degree of co-occurrence, proximity (see Figure 1).

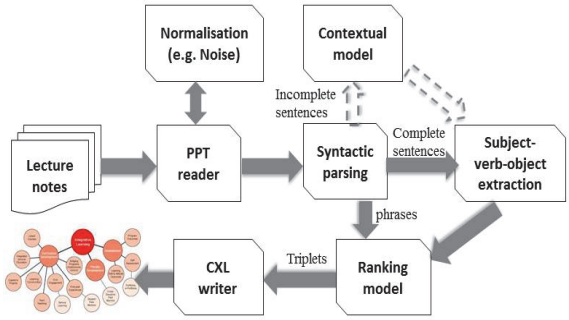

Figure 1: Overview of concept map extraction process.

As shown in Figure 1, our system relies on the use of the lecture notes presented as set of slides. Therefore, it is capable of extracting rich text features such as underline, font color and highlights and type of text such as a title, bullet point, and sub-point. Lecture notes frequently contain noisy data such as course announcements and assignment details that are irrelevant for a knowledge

representation. The system detects and resolves them automatically by utilising *co-occurrence* between domain-related and unrelated topics. For example, if course title is co-occurred with some terms in body text, that pair of terms has strong relation with the domain, and hence recognised as a domain-specific terms.

Lecture slides occasionally contain incomplete and ambiguous English sentences for machine interpretation. Therefore, it is challenging to apply NLP algorithms to extract knowledge from lecture slides. We implemented a contextual model which automatically replaces syntactically and semantically missing entities (e.g. subjects or objects of sentences). Our initial research also focused on resolving pronouns (e.g. *it, their*) and demonstrative determiners (e.g. *these, this*) using a backward search approach (under review).

In contrast to other related works in literature (Chen et al., 2008), which has no relation labels among extracted concepts, our work generates concept-relation-concept triplets by analysing subject-verb-object (SVO) in English sentences. We utilise the link grammar parser developed by CMU[1] to extract SVO in English sentences and applied the greedy approach to the remaining text to identify key terms using part-of-speech tags. The extracted key terms are ranked using the approach discussed in Section 4.

The extracted concepts and relationships are arranged according to their importance, which produces a CXL (Concept map extensible language) file which can be directly exported to IHMC cmap tools[2] [1]for visualisation (Figure 2).

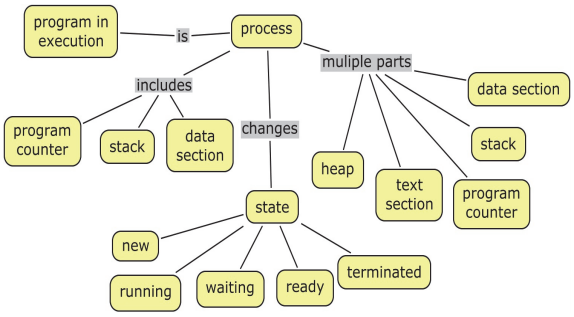

Figure 2: An example of an extracted concept map from 'process' topic of Operating system course.

4 RANKING MODEL

In order to construct a high quality concept map, both domain knowledge and hierarchy are equally significant (Novak and Canas, 2006). This section discusses three candidate models which arrange concepts by their importance.

4.1 Baseline Model

Our knowledge source (i.e. lecture slides) contains a natural layout of presentation title, slide headings, bullet points, and enumerated sub-points. Therefore, one can argue that this layout can directly transfer to a hierarchy. To validate this assumption, we implemented a baseline model by integrating 'text location' in lecture slides (Table 1).

Hypothesis I: *Text location allocated by the natural layout of presentation slides might influence human judgment of which concepts are most important*

Table 1: concept importance by location.

Location	Rank
Title	3
Bullet statement	2
Sub-point	1

However, a concept can occur in multiple locations. In order to select the most suitable location for such concepts, we implemented a "link-distance algorithm" which can be found in our previous work (Atapattu et al, 2012).

4.2 Linguistic Feature Model

First, we used the greedy approach to extract nouns and noun phrases using part-of-speech tags (Atapattu et al., 2012). Although, this approach is efficient for extracting isolated nouns or noun phrases, we found it difficult to extract phrases joined by prepositions (e.g. *of, for,in*) and conjunctions (e.g. *and, or).*Therefore, we developed a new approach using the link grammar parser of CMU[2], which produces syntactic parse trees (Figure 3).

It is straightforward to extract nouns (*leaf* nodes) or noun phrases (*pre-terminal* which is one level above *leaf*). This approach outperforms the first method and hence, solves the preposition and conjunction issue.

Our hypothesis is based on the recommendation of using the smallest number of words for a concept (Novak and Canas, 2006).

[1] http://cmap.ihmc.us/

[2] http://www.link.cs.cmu.edu/link/

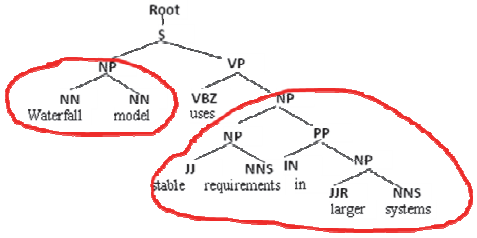

Figure 3: Syntactic parser tree of an example English Sentence.

Hypothesis II: *Simple grammatical structures (nouns, noun phrases) of Lecture slides might have higher influence than complex grammatical structures (nested sentences, dependent clauses, indirect objects) for human judgment of which concepts are most important*

Table 2 shows our ranking based on grammatical structure.

Table 2: Concept importance by grammatical structure; NP: noun phrase, PP: prepositional phrase, S: sentence, VP: verb phrase (see all tags in http://bulba.sdsu.edu/jeanette/thesis/PennTags.html).

Feature	Example grammatical structure	Rank
Noun phrase	(NP (NP (NNP **Advantage**)) (PP (IN **of**) (NP (NN **unit**) (NN **testing**))))	3
Simple sentence	(S (NP (NNP **Process**)) (VP (VBZ **is**) (NP (NP (NN **program**)) (PP (IN **in**) (NP (NN **execution**))))))	2
Complex sentence	(S (NP (DT A) (NN **software**) (NN **process**)) (VP (VBZ **is**) (NP (NP (DT **a**) (NN **set**)) (PP (IN **of**) (NP (NP (ADJP (RB **partially**) (VBN **ordered**)) (NNS **activities**)) (CC **and**) (NP (NP (JJ **associated**) (NNS **results**)) (SBAR (WHNP (WDT **that**)) (S (VP (VBP **produce**) (CC **or**) (VBP **maintain**) (NP (DT **a**) (NN **software**) (NN **product**)))))))))))	1

As shown in Table 2, complex sentences contain nested sentences (S), clauses (SBAR) and conjunctions (CC). Therefore, we assume these sentences contain definitions or elaborations rather than the abstract concepts of a knowledge representation. Verb phrase (VP) is the remaining grammatical structure which is usually nested with a verb (or multiple verbs) and a noun phrase. We usually extract NPs from verb phrases.

4.3 Structural Feature Model

In the third candidate model, we integrate some structural features (e.g. incoming, outgoing links and proximity) which have already been proposed in Zouaq et al., 2012 and Leake et al., 2004 and new distributional features (e.g. typography and co-occurrence) that are unique to presentation framework.

Hypothesis III: *Structural (Incoming and outgoing links, proximity) and distributional (term frequency, degree of co-occurrence, typography) features might influence the human judgment of which concepts are most important*

Log Frequency Weight

The system counts the occurrence of *nouns* or *noun phrases* and normalises the term frequency (t_f) (Atapattu et al., 2012). This value is significant than typical term frequency measure used in information retrieval applications since our 'terms' are restricted to nouns or noun phrases.

$$W_t = log\ (1 + t_f) \quad (1)$$

Incoming and Outgoing Links (I/O links)

We keep track of the number of incoming (n_i) and outgoing (n_o) connections for each node. The 'root' node contains only outgoing links and leaf nodes contain only incoming links. Those that have more outgoing than incoming are identifies as of greater importance.

These metrics are significant to demonstrate disjoint nodes from central concept map. Our system provides this information as a conceptual feedback for teachers. This feedback can be used to reflect on whether their expert structures have been transferred successfully to teaching material. If not, students struggle to organise disjoint information into their knowledge structures since there is no relation between new and existing information (paper under submission).

$$W_o = n_o \quad (2)$$

$$W_i = n_i \quad (3)$$

Degree of Co-occurrence

Our hypothesis is '*if two key terms co-occur in many slides (equals to pages in other documents), it is assumed that those two terms have a strong relation*'

and hence, can be chosen as domain concepts. To measure the degree of co-occurrence, we use the Jaccard coefficient, a statistical measure which compares the similarity of two sample sets.

In order to measure the degree of co-occurrence between term t_1 and term t_2, first calculate the number of slides, that t_1 and t_2 co-occurs. This is denoted as $| n_1 \cap n_2 |$. Then calculate the number of slides the term t_1 ($|n_1|$), t_2 ($|n_2|$) occurs. The degree of co-occurrence of t_1 and t_2 is denoted by J (t_1, t_2) is,

$$J(t_1, t_2) = \frac{|n_1 \cap n_2|}{|n_1 \cup n_2|} = \frac{|n_1 \cap n_2|}{(|n_1| + |n_2| - |n_1 \cap n_2|)} \quad (4)$$

This value is utilised as a key decisive factor for noise detection since key terms such as a*nnouncements, assignments* have low degree of co-occurrence with other domain concepts.

Typography

Lecture slides often contain emphasised texts (e.g. different font color, underline) to illustrate their importance in the given domain. We introduced a probability model to select candidate concepts using their level of emphasis. According to the proposed model, terms which contain infrequent styles are allocated higher weights. More information of this work can be found in Atapattu et al., 2012.

Proximity

We consider the 'lecture topic' as the *root* (or central concept) of concept map. Therefore, we hypothesise the concepts that have a higher proximity to the *root* are expected to be more important than those with lower proximity (Leake et al., 2004). We denote the proximity weight (*Wp)* by calculating the number of nodes (d_n) from root to participating node (inclusive).

$$W_p = \frac{1}{d_n} \quad (5)$$

Generally, a concept map with 15 to 25 nodes is sufficient to assist learning while not providing an overwhelming amount of information (Novak and Canas, 2006). Thus, the aim of introducing a ranking model is to construct a conceptual overview with the most important domain knowledge from the lecture notes.

5 EVALUATION OF CONCEPT IMPORTANCE

We conducted experiments with domain experts (lecturers) to study their judgment of concept importance in their lecture notes. These data are then compared with the machine predictions to assess the accuracy of the auto-generated concept maps.

Data

Seven computer science courses across different Undergraduate levels (1st year, 2nd year, 3rd year and 4th year) were selected. These courses contain a combination of content types such as text, program codes, mathematical notations, tables and images. The seven courses chosen were *Introductory programming (IP), Algorithm design and data structures (ADDS), Object oriented programming (OOP) (level 1); Software Engineering (SE) (level 2); Distributed systems (DS), Operating systems (OS) (level 3); and Software Architecture (SA) (level 4)*. Each participant was provided with approximately 54 slides including one to three topics. Tasks were designed to be completed within 30 to 45 minutes, with the variation due to how recently the lecturer had been teaching the course.

Seven lecturers from the Computer Science School volunteered to assist with the experiments. They are the domain experts of selected topics who have extensive experience in teaching the courses.

Procedure

This study required participants to rate the domain concepts according to their importance. The judgment was expected to reflect personal opinions based on their knowledge and perception. However, we provided a few tips, such as how the importance of a concept can be affected by the learning outcome, course objective, and examination perspective. These instructions did not have any relation with the factors we considered in developing our concept map extraction tool.

We provided colour pens and printed lecture slides to the participants who preferred working in a paper-based environment. The rest used their computers or tablets to highlight the domain concepts. The three rating scale given to the participants consisted of '*most important', 'important',* and *'least important'* using three colours *'red', 'yellow' and 'green'* respectively. Participants tended to rate single concepts as well as noun phrases.

During the experiments, we did not show the machine-extracted concept maps to the participants. They only had access to the course lecture slides. This could prevent any influence arising from structure or layout of concept maps for the human judgement.

Results

We developed a simple program to extract the annotations of participants. A Java API for

Microsoft framework[3] was used to extract highlighted texts. Using this approach, we extracted 678 concepts from 376 lecture slides. The average number of concepts per slide was approximately 2.2 except in IP course. In IP, multiple slides repeated the same content in animations. Therefore, in IP, the average number of concepts per slide is 0.8.

The highlighted texts are categorised and sorted based on their ranks from 3 to 1 (most important to least important). Similarly, our system arranged important concepts according to ranks assigned by each candidate models.

In the baseline model, our ranking algorithm allocated rank 3 for text located in *titles* (see Table 1) and 0 for concepts annotated by human, but not retrieved by machine. The two rankings were compared using ranking correlation coefficient and results are presented in table 4. The correlation (r_s) is close to 0 for the majority of the courses except for ADDS and SA. This implies there is no linear correlation between human judgment of concept importance and the natural layout of presentation software. This causes us to question and reject the original hypothesis that assumes most important, important and least important concepts are located in titles, bullet points and sub points respectively. Therefore, the approach which utilises the natural layout of lecture slides for knowledge organisation does not produce an acceptable outcome (Ono et al., 2011). Further, topic map extraction in Gantayat et al., 2011 and Kinchin, 2006 should focus on fine-grained course contents in addition to lecture headings. The feedback obtained from lecturers regarding concept importance is significant for students. This implies layout of slides is not overlapping with lecturer's judgment of what is more important in the lecture.

However, if we could expand the ranking to a few other levels, we could expect a slightly more positive correlation from the baseline model. This occurs because the ranking model categorises remaining concepts as false positive (rank=0) that have not been ranked by human and false negative (rank =0) that have not been retrieved by machine, but annotated by human.

The linguistic feature model assumes the grammatical structure of text (noun / phrases, simple sentences and complex sentences) has an impact for selecting candidate concepts. Similar to the baseline model, this has assigned higher rank (rank = 3) for noun or noun phrases and lower rank (rank = 1) for complex grammatical structures (see Table 2). However, Table 4 shows the correlation is closer to 0 for all the selected courses. This reveals that, in addition to single terms and brief phrases, simple and complex sentences contain candidate domain concepts. Therefore, a deep analysis of all text contents irrespective of their grammatical complexity is significant to extract the useful knowledge from lecture slides.

In the structural candidate model, we normalise weights of each metrics within the range of 0-1. The influence of each metric (discussed in Section 4.3) is determined by the parameter values (Table 3). For example, terms with higher outgoing links can be more general, thus more important than terms with higher incoming links. We trained our weighting function using previously annotated data for a previous study (Atapattu et al., 2012). The training data contains slides extracted from recommended text books, university course materials and randomly chosen topics from web.

Table 3: Best fit parameter values for Structural features.

Feature	**Best fit parameter values**
Outgoing links	0.923
Proximity	0.853
Typography	0.764
Co-occurrence	0.559
Frequency	0.514
Incoming links	0.281

After obtaining best fit parameter values, we calculated the aggregate weight for each term in the study and sort them in the descending order of weights. Our system defines *upper*, *medium* and *lower* threshold values in order to rank the *most important* (above upper), *important* (in-between upper and medium) and *least important* (in-between medium and lower) domain concepts. These three threshold values vary depending on the number of concepts retrieved. Finally, similar to other two candidate models, we compare the ranks given by participants with machine predication. The results can be found in the last column of Table 4.

The results are interpreted as strong positive or strong negative if r_s close to +1 or -1 respectively. There is no linear correlation when r_s is close to 0 and hence, consider as independent variables.

Since the selected courses contain combinations of content (e.g. text, images, program codes), we claim our data ranges from *well-fitted* (e.g. SE and SA) to *ill-fitted* (e.g. IP and ADDS) contents for 'machine interpretation'.

[3] http:/poi.apache.org/

Table 4: Spearman's ranking correlation (r_s) between candidate models and Computer Science courses.

Model	Baseline (r_s)	Linguistic (r_s)	Structural (r_s)
SE	0.193	0.247	0.805
ADDS	0.436	0.252	0.435
IP	0.113	0.293	0.353
OS	0.325	0.240	0.715
DS	0.183	0.129	0.455
OOP	0.287	0.347	0.521
SA	0.605	0.050	0.806

$$r_s = \frac{1 - 6\sum d_i^2}{n(n^2 - 1)} \quad (6)$$

Figure 4: Distance between human and computer ranking against number of concepts (%) in Software testing topic (r_s=0.813).

In the structural feature model, our results show satisfactory correlation for the majority of the courses and strong positive correlation for SE, SA and OS courses. As an example, in Software Testing topic (Figure 4), 55% of concepts (out of 64) overlap between computer and human (distance = 0) and 39% of concepts indicate one level difference between ranks. This implies 94% of concepts extracted from machine algorithms are closely aligned with human judgement, resulting in a machine extraction of approximate expert maps. Both OS and SE lecture slides are constructed using popular text books written by Sommerville and Silberschatz respectively and SA lecture slides were well written and structured. Therefore, those topics contain rich grammar, good summarisation and emphasise domain concepts. These *well-fitted* contents assist relatively straightforward machine interpretation.

Conversely, the remaining course topics include combinations of category headings (e.g. review, summary, welcome, and today's format), additional text boxes with excessive content, ambiguous texts that are difficult to resolve and repetitive contents in consecutive slides for animations (i.e *ill-fitted* content). These types of content reduce the reliability of machine extraction algorithms. Hence, as a general rule, machine-extracted concept map has a significant correlation with human judgment in *well-fitted* contents.

This study highlights the importance of structural features rather than natural layout or grammatical structures. This implies that important information in the lecture should be emphasised, and recapped. Lecturer should also construct probable links with the central idea of the topic. This ensures that approximately reliable machine extraction of concept maps from algorithms developed in this work.

In this study, we only had a single expert participating for the assessment of each course. Therefore, we cannot measure the *inter-rater agreement* since the author of the material is the only person having an expert knowledge structure of the content.

We received evocative feedback from domain experts during the experiments.

"*I tend to think that summary generally contains things that have already been discussed. But, I found a new concept in the summary which hasn't seen in the lecture note. I read the lecture from the beginning again to locate that concept, but couldn't find it".*

This comment provides an evident that there can be disjoint concepts included in lecture note which are not fitting with students' knowledge structures.

"There are tables which provide comparison between important concepts. How does this handles by the system?"

This is one of our challenges. The data comes from tabular form include useful domain concepts. However, we have not yet implemented a feature to tackle the comparisons in tabular data.

"*Examples are very useful to learn concepts, but they are not concepts. Therefore, I am not sure whether they should be included or not. I have included them in cases where I think they are very useful".*

"In IP, many domain concepts are introduced via analogy. So, are they also be classified?"

We do not have an exact answer for this comment. Examples or analogies can be included into the extracted concept map, if they are strongly correlates with domain or emphasised within the context.

In our future work, we plan to extend the experiments across disciplines to create a general model. The focus of this study is limited to measure the quality of 'concept' ranking according to their

importance. We plan to extend our study to measure the quality of extracted 'relations'. It is difficult for participants to judge relationships from lecture slides since relations are not highly visible like concepts. Therefore, we plan to provide extracted concept maps using IHMC cmap tools to collect feedback on the 'strength of extracted relationships'. Lecturers will also receive conceptual feedback regarding deficiencies in knowledge organisation of their courses. This includes disjoint concepts without any relation to the central concept map and relations without proper labelling. This process should improve the legibility of the materials.

6 CONCLUSIONS

The primary challenge of concept map mining is the lack of a suitable evaluation framework. The existing approaches utilise human experts' judgement or expert maps as the gold standard to measure the quality and validity of machine-extracted maps. However, these studies focus on *concept existence* using IR metrics – *precision* and *recall*, and not the concept ranking according to their importance. Therefore, this paper proposes a machine-based evaluation mechanism to assess mined concept maps in an educational context. We compared the machine-generated maps with human judgment and obtained strong positive correlation (r_s ~1) for *well-fitted* courses.

This work has potential to be utilised as conceptual feedback for lecturers to have an overview of knowledge organisation of their courses. Machine-extracted concept maps require the assistance of domain experts to validate. However, this effort is substantially smaller than that required to construct a concept map manually. In future work, we plan to provide task-adapted concept maps instead of hints in intelligent tutoring environment. This will help students to identify knowledge gaps and to improve their organisation of knowledge. We believe that this will help to improve the depth of meaning that students can extract from their learning.

REFERENCES

Atapattu, T, Falkner, K. and Falkner, N. 2012. Automated extraction of semantic concepts from semi-structured data: supporting computer-based education through analysis of lecture notes. In *proceedings of the 23rd International conference on Database and Expert systems applications*, Vienna, Austria.

Alves, A., Pereira, F and Cardoso, F. 2002. Automatic reading and learning from text. In *International Symposium on Artificial Intelligence.*

Ausubel, D., Novak, J. and Hanesian, H. 1978. *Educational psychology: A cognitive view,* New York.

Chang, S. N. 2007. Externalising students' mental models through concept maps. Journal of Biological Education.

Chen, N., Kinshuk, and Wei, C. 2008. Mining e-learning domain concept map from academic articles, Computer and Education.

Coffey, J., Carnot, M., Feltovich, P., Feltovich, J., Hoffman, R., Canas, A. and Novak, J. 2003. A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support, The Chief of Naval Education and Training.

Dali, L., Rusu, D., Fortuna, B., Mladenic, D., Grobelnik, M. 2009. Question answering based on Semantic graphs. In *Language and Technology Conference.* Poznan, Poland.

Frantzi, K., Ananiadou, S. and Mima, H. 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries.

Gantayat, N and Iyer, S. 2011. Automated building of domain ontologies from lecture notes in courseware. In *Proceedings of the IEEE international conference on Technology for education*, India.

Gouli, E., Gogoulou, A., Papanikolaou, K. and Grigoriadou, M. 2004. COMPASS: An adaptive web-based concept map assessment tool. In *Proceedings of the first international conference on concept mapping.*

Hearst, M. A., 2000. The debate on automated essay grading. Intelligent systems and their Applications.

Kinchin, I. 2006. Developing PowerPoint handouts to support meaningful learning. British Journal of Education technology. 37 (4), 647-650.

Leake, D., Maguitman, A. and Reichherzer, T. 2004. Understanding Knowledge Models: Modelling Assessment of Concept Importance in Concept Maps. In *Proceedings of CogSc*.

Maass, J., Pavlik, P. 2013. Utilising Concept mapping in Intelligent Tutoring Systems. In *Artificial Intelligence in Education.*

Novak, J. and Canas, A. 2006. The theory underlying Concept maps and How to construct and use them. Institute of Human and Machine Cognition.

Novak, J. and Gowin, D. 1984. *Learning how to learn.* Cambridge University Press, New York and Cambridge.

Olney, A. M., Graesser, A. and Person, N. 2012. Question generation from Concept maps. Special issue on Question generation, Dialogue and Discourse.

Ono, M., Harada, F. and Shimakawa, H. 2011. Semantic network to formalise Learning items from Lecture notes. International Journal of Advanced Computer Science.

Starr, R. and Oliveira, J. 2013. Concept maps as the first

step in an ontology construction. Information systems.
Villalon, J. and Calvo, R., 2008. Concept map mining: A definition and a framework for its evaluation. In *International Conference on Web Intelligence and Intelligent Agent Technology*.
Zouaq, A. and Nkabou, R. 2009. Evaluating the generation of domain ontologies in the knowledge puzzle project. IEEE Transactions on Knowledge and Data Engineering.
Zouaq, A., Gasevic, D. and Hatala, M. 2012. Voting theory for concept detection. The Semantic Web: Research and Applications.