Concept Maps: Making Learning Meaningful
Proc. of Fourth Int. Conference on Concept Mapping
Viña del Mar, Chile, 2010

ANALYSIS OF A GOLD STANDARD FOR CONCEPT MAP MINING – HOW HUMANS SUMMARIZE TEXT USING CONCEPT MAPS

Jorge Villalon & Rafael A. Calvo, University of Sydney, Australia

Rodrigo Montenegro, Universidad Adolfo Ibáñez, Chile

Email: jorge.villalon@uai.cl, rafa@ee.usyd.edu.au, rodrigo.montenegro2004@alumnos.uai.cl

Abstract. Essay writing is an important learning activity which involves higher level thinking, the same as concept mapping activities.

1 Introduction

Essay writing is a common requirement for students in higher education due to the higher level thinking required to write essays (Paltridge 2004). Academic writing is an iterative process of writing and revision in which new version improves the final outcome. During revisions the author reflects on his own work and ideas, requiring analysis, argument and independent thought (Emig 1977). However, students can find this reflective thinking activity as hard to engage if it lacks the proper support. Concept mapping also requires higher level thinking from the students for them to build good quality concept maps (Novak and Gowin 1984). A good concept map contains only relevant concepts (a perceived regularity in events or objects, or records of events or objects, designated by a label), connected by linking words into coherent propositions. On deciding what concepts to include in a concept map, and on linking them properly the author's reflection is required (Novak 2007).

Concept maps have been used to support reading and writing activities, what is known as Text Concept Mapping (TCM) (Nurit Nathan 2004). The activities usually consist on summarizing the key ideas in a piece of text, and there are three ways of doing it: Building a concept map from scratch, fixing a previously built concept map and studying a concept map. In the first activity the students build a concept map without any support, in the second activity the teacher builds a map that has some errors and/or missing information that the students have to fix, and in the final activity the students study a concept map built by the teacher which summarizes the text. All activities have been shown to improve the students' understanding on the readings' topics (en Chang, Sung et al. 2002; Hauser, Nuckles et al. 2006).

TCM could be used to support the author's reflection during the writing process, particularly analyzing a concept map of their own work which represents a new visualization of the ideas expressed in their text (Villalon, Kearney et al. 2008). However concept mapping is highly time consuming, therefore having someone to build a concept map from each essay would become a bottleneck that will prevent the feedback to be delivered in a timely fashion, becoming an obstacle during the writing process. This problem could be solved by the automatic extraction of concept maps from student produced text, known as Concept Map Mining (CMM)(Villalon and Calvo 2008).

CMM has been proposed before for several applications such as overcoming the large amount of available data (Valerio and Leake 2006), to facilitate the organization of information in digital libraries (Shen, Richardson et al. 2005), and as a previous step to build domain ontologies (Zouaq and Nkambou 2008). In the context of essay writing, automatic concept maps can be used either to facilitate the construction for teachers or to present them to the author as a raw version they have to fix. In both cases the feedback could be given rapidly so the writing process is not interrupted. However, CMM is still a work in progress, which requires several steps in order to achieve its goal (Villalon and Calvo 2009).

The biggest challenge in CMM is the highly subjective task of evaluating if a concept map correctly summarizes an essay. In order to gain a better understanding of it, first one must have to understand how humans perform the task, so interesting patterns can be analyzed and then inform the design of an automatic process. Such understanding is gai

ned through the construction of a gold standard, which corresponds to concept maps extracted from essays by at least two human annotators, following the same procedure. This collection of concept maps defines a 'gold standard' for the CMM process, and any automatic procedure should aim to extract maps that are equivalent to those in it (Villalon and Calvo 2008).

This paper presents an analysis of a gold standard constructed to evaluate CMM algorithms. Such a gold standard represents a new visual representation of the knowledge expressed in the text in the form of concept maps, and carries a lot of information on how humans respond to the task of summarizing someone else's text into a concept map, particularly students' essays which have several other complexities (grammatical and spelling errors). The analysis tries to answer the question on "How humans summarize knowledge from text using concept maps?" It is structured as follows: Section 2 presents a summary on what CMM is. Section 3 reports on the process of building the gold standard. Section 4 presents an extensive analysis on the produced maps and section 5 concludes.

2 Concept Map Mining

Concept Map Mining is defined as the automatic extraction of concept maps from text that are useful in educational contexts. Its aim is to provide new ways to visualize the knowledge expressed in the text for human consumption. It focuses on the use of concept maps to support writing activities in educational scenarios, which poses several challenges to the task that are summarized in four requirements for automatic concept maps: Educational utility, simplicity, semi-formality and subjectivity. Educational utility refers to the structure required for an automatic concept map which has to be the same as the one proposed by Novak (Novak 2007), otherwise their benefit as educational tools can be questioned. Simplicity refers to the fact that the automatic concept map has to be a reduced version of the text, otherwise it may require the same effort as reading the text. Semi-formality refers to the computational representation of the concept map, which has to be structured enough so they can be processed as digital concept maps can be. Finally, subjectivity refers to the fact that the knowledge represented in the automatic concept map has to be faithful to the written essay, and not to be influenced by external knowledge.

The CMM process consists on identifying the key concepts in a piece of text and the linking words that connect them. It has three sub-tasks which are: Concept Extraction, Relationship Extraction and Summarization. The first task aims to identify every possible concept in the text, the second aims to find all possible connections between the previous concepts and the third step consists on creating a reduced version of the map that summarizes the content, avoiding redundancy and maximizing coverage. Figure 1 shows a scheme that summarizes the process.

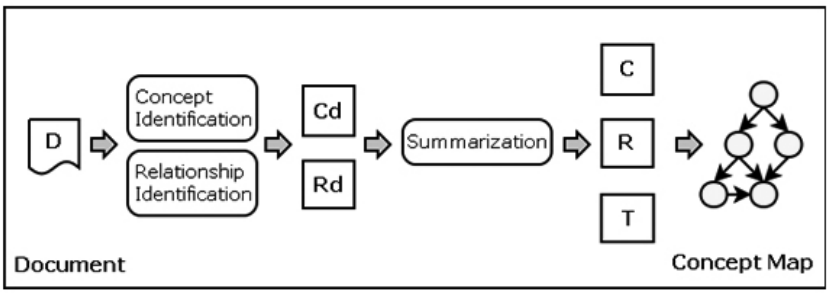

Figure 1. The CMM process

Ontology Learning from Text is an area that is related to CMM in the same way Ontologies are related to concept maps, both have structural similarities however they differ in their goals and therefore in their methods. Ontologies are knowledge representations that also comprise concepts linked by meaningful relationships, making the structurally similar to concept maps. However, ontologies reflect a shared knowledge which is the consensus of a set of experts, while concept maps can reflect the knowledge of any person, particularly non experts which can be highly subjective

and incomplete. As a consequence concept maps can contain information that could hardly be used in an ontology, but at the same time it is highly useful in educational scenarios (identifying misconceptions and/or mistakes).

3 Gold Standard

The annotation of a gold standard is a common requirement for the evaluation of knowledge extraction algorithms (Wiebe, Bruce et al. 1999; Dellschaft and Staab 2006). This is due to the highly subjective nature of knowledge, which requires humans to evaluate what can be considered relevant knowledge and what should not. In other words, the correct answer to a highly subjective question such as "What is the concept map that summarizes a text?" can be answered only by humans. Moreover, it is acknowledged that this subjectivity will not be assessed equally among different human subjects, therefore the answer will be as good as the one shared by most humans (Carletta 1996). In terms of a gold standard for CMM, this means that several human annotators must extract concept maps from the same collection of essays.

Every gold standard requires an annotation procedure that has to be clearly defined, and supported by tools. The procedure used for the gold standard analyzed in this paper was originally written following Novak's suggestions on how to build good concept maps, and then refined by consecutive iterations of the annotators performing it. The procedure was also supported using an annotation tool that checked the consistency of the annotations. Figure 2 shows the interface of the annotation tool, which presents the essay on the left side of the screen, the concepts selected in the previous step (Concept extraction), than can be linked with another concept in the map using linking words that can be freely added.

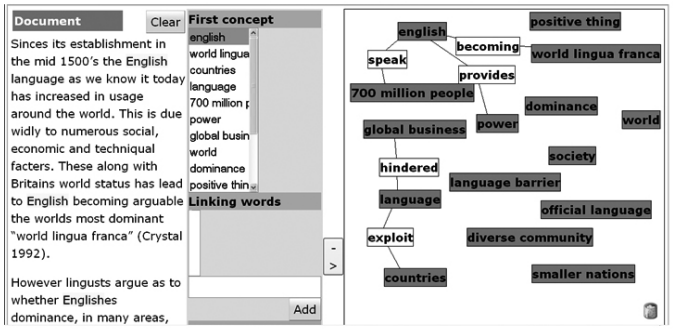

Figure 2. Interface of the annotation tool.

The gold standard was built from a collection of 42 essays written by first year university students. The annotation was done by five human annotators that had previous experience as concept mappers. Three of the annotators had less experience, they were chosen to perform a preliminary annotation to improve the procedure and the last two produced the final version. The first iteration of the annotation process showed that most humans used external information that was not necessarily expressed in the text, although it was common sense to infer the author's intention. This lead to force that the concepts should appear with the linking word within a text passage, in this case a paragraph was defined. The resultant gold standard contains 42 concept maps, totaling 877 concepts and 516 relationships, and took an average of 27 minutes per essay.

4 Analysis

The analysis of a gold standard looks to answer questions that enlighten the understanding on how people perform a highly subjective task such as building concept maps from text. These questions can be classified depending on different aspects of concept maps and text.

- Coverage: How many concepts annotators use to summarize an essay? How many relationships? Is it related to the length of the text? Is it related to its quality?
- Concepts: Are they chosen according to linguistic characteristics such as their part of speech? Are they chosen due to statistical characteristics such as the frequency with which they appear?
- Relationships: Are there common linguistic characteristics to linking words? Are concepts chosen mostly within sentences or paragraphs? Does the sequence presented in the concept map reflect the same sequence in the text?

The answers to these questions provide important evidence on the patterns humans use when creating concept maps from text, which represent the core information for the construction of automatic algorithms. On identifying interesting patterns, suitable techniques to identify such patterns should provide the best possibility for automatic extraction.

4.1 Coverage

When asking humans to summarize a piece of text, a reduced version of it is produced. For example in automatic summarization processes a percentage of the original content, measured in words, sentences or paragraphs. But what happens when a concept map is the new representation? Is it related to its length? Is it related to its quality? In order to analyze this, the length of the essays was calculated in words and sentences. The total number of concepts and relationships used by the annotators was also computed and the correlation between these values is shown in Table 1.

Correlations	Average number of Concepts	Average number of Relationships
Length in words	0.62	0.38
Length in sentences	0.52	0.30
Grades	**0.28**	**0.14**

Table 1. Correlation between the size of concept maps, the length of the essays and their grades.

Several conclusions can be made from the table above, firstly that the length of the essay has a positive correlation with the number of concepts in a concept map, longer essays will be summarized using more concepts. However, the number of connections between those concepts has a significantly lower correlation, indicating that a longer essay doesn't necessarily mean a denser concept map. Finally, the last row in the table shows that the correlation between the grades (that represent to some extent the quality of the essay) and the number of concepts and relationships is very low. This last observation means that when the annotators were asked to summarize the content reflecting the ideas expressed in the essay, regardless of their own knowledge or external information on the essay's subject matter, they will select concepts and relationships making sure they cover the information in them.

4.2 Concepts

Novak proposed a definition of concepts: "A perceived regularity in events or objects, or records of events or objects, designated by a label". This definition is closely related to what is known in linguistics as "nouns", therefore it can be expected that the concepts chosen by the annotators to share such characteristic. Linking words on the other hand are supposed to form "meaningful propositions", however this definition is very subjective therefore there are no obvious linguistic characteristics that can be associated to meaning.

To assess the expected notions on the words or phrases chosen by the humans for concepts and relationships, each text was analyzed using a grammatical parser in order to identify each phrase linguistic characteristic, particularly their part of speech. Figure 3 shows the relative frequency of the parts of speech for concepts, the graph shows that more than 80% of them correspond to noun phrases, validating the closely related definitions of nouns and concepts. This is highly relevant for CMM because such characteristic (this is, a word or phrase being a noun or noun phrase) indicates a very high probability that the concept should go into the final concept map.

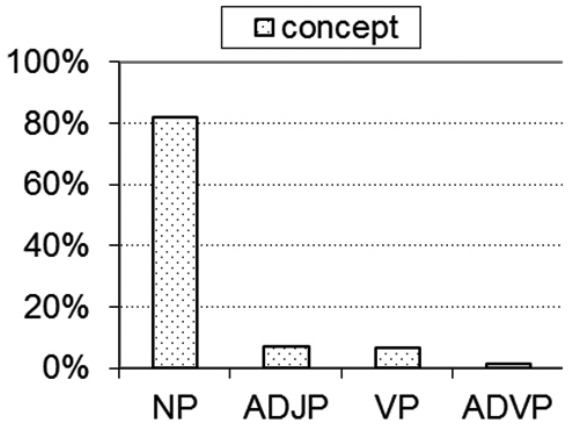

Figure 3. Part of speech for concepts.

Another pattern that can be calculated automatically for phrases and words in documents are their frequencies. In research areas such as keyword extraction and automatic summarization, the frequency of words can be used to identify relevant ideas. In order to validate or refute this idea, the frequency of every concept and relationship within each essay was calculated. Such frequencies represent a distribution that can be highly skewed if a few concepts appear several times while the others appear only a few times, follow a normal distribution if a majority of concepts share an average frequency, or have a uniform distribution if most concepts have a similar frequency.

Figure 4 shows an example of the concepts' distribution for an essay, which is highly skewed with concepts like English or language appearing more than 10 times, while diplomacy or the internet appear only once or twice. The Kurtosis measure of "peakedness" was used to calculate the distributions of concepts and relationships for each concept map, this measure goes between -1.2 for uniform distributions, 0 for normal distributions to 3 or higher for highly skewed distributions. Table 2 shows the results, indicating that both concepts and relationships are highly skewed with Kurtosis measures of 6.06 and 2.07 respectively. This result indicates that only a few concepts appear several times in an essay, but the majority appears only a few times.

One possible explanation for the highly skewed distributions is that some concepts are chosen due to their importance within the essay, but others by a different reason, like forming meaningful propositions with the important concepts. This behavior has been observed before and it is known as the Zipf's law, which indicates that only a few terms count for the most common utterances in any piece of text. This might mean that summarizing text using Concept Maps shows a similar behavior on choosing words as that of writing text from scratch. In order to validate if this behavior is such, a deeper analysis on the relationships is necessary.

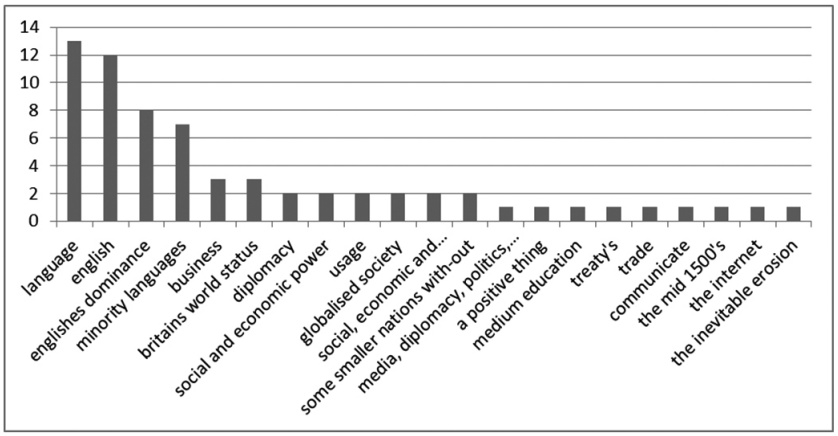

Figure 4. Sample distribution of concepts within an essay.

	Average Kurtosis
Concepts distribution	6.06
Relationships distribution	2.07

Table 2. Average Kurtosis measure for concepts and relationships distributions.

4.3 Relationships

The first analysis on relationships corresponds to identifying the part of speech of linking words, in order to find meaningful patterns. Figure 5 shows the relative distributions of the linking words' parts of speech, with less than 50% for the most common part of speech (verbal phrases), and at least other six relevant categories (40% between noun phrases and adverbs and adjectives), no clear patterns can be defined. This result indicates that the automatic extraction of relationships cannot rely simply on the linguistic characteristics to identify linking words.

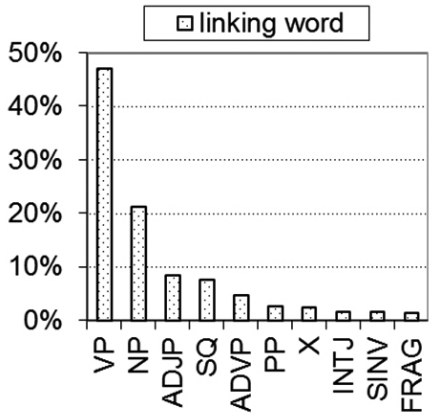

Figure 5. Part of speech frequency for linking words.

A second analysis was done trying to identify if meaningful propositions are obtained linking concepts that appear all together within paragraphs or within sentences. One example is the paragraph: "English, as a predominant and indeed 'global' language is a double edged sword. While easing communication on a global level, it also breeds elitism in countries it has been exported to." Which was represented in the concept map with the proposition "English - breeds -> elitism". An example for the sentence: "Graddol on the other hand points to the social and economic inequality that the dominance of English could lead to." Was represented as "English – lead -> the social and economic inequality".

Table 3 shows the percentage of relationships that appear within a paragraph versus those appearing within a sentence. In this case the results show a clear preference for human annotators to choose two concepts and a linking word from a single sentence. A possible explanation for this behavior could be the internal coherence of the text, indicating a bigger level of coherence within sentences than within paragraphs. This characteristic is highly relevant for CMM because it defines a pattern to identify connections for the most relevant concepts.

A final relevant analysis for relationships is comparing the sequence of a proposition within the concept map versus the essay. In the paragraph example shown before, the proposition chosen to represent the text follows the same sequence of the words in the essay, this is the source concept is followed by the linking word and then by the target concept (not necessarily one next to the other). In the sentence example on the other hand, both concepts appear before the linking word due to the grammatical structure of the sentence.

	% of Relationships
Within a sentence	92.1
Within a paragraph	7.9

Table 3. Percentage of relationships appearing within sentences or paragraphs.

In order to find out if there was a pattern indicating any preference for human annotators, the sequence of each relationship was compared to its sequence in the essay. Table 4 contains the results, showing that there's no clear pattern for choosing one type of sentences over the other. This result indicates that no conclusions can be made from the sequence of words in order to identify linking words for concepts.

	% of Relationships
Same sequence as essay	57.7
Not same sequence as essay	42.3

Table 4. Percentage of relationships following the same sequence as the essay.

5 Summary

This article explained what Concept Map Mining (CMM) is: The automatic extraction of concept maps from essays for educational purposes, and presented the analysis of a gold standard constructed for the purpose of evaluating the algorithms that will implement the task. The main goal of the analysis is to gain an understanding on the characteristics of the concept maps produced by human annotators when asked to create a summary of a piece of text. Such patterns will inform the design of the automatic algorithms that will implement CMM.

Three aspects of the concept maps and the essays were analyzed: Characteristics of the complete essay and concept map (such as the correspondence of their sizes and its correlation with the essay quality), characteristics of the concepts (such as linguistic information and frequency) and characteristics of the relationships (linguistic information, location in text and their sequence). To do this, all essays were parsed identifying grammatical roles such as noun phrases or verbal phrases, all statistics on their sizes were evaluated and each concept map was analyzed against its essay source.

The results of the analysis showed that the size of the concept maps measured in the number of concepts is correlated with the length of the essay measured in both words and sentences. However, the concept maps' connectedness has a lower correlation to their length and none of them correlates with the grades. This indicates that when humans are asked to summarize the content of an essay, they will include information based on the total information in the essay and not in its quality.

A second finding is that linguistic characteristics such as words' part of speech are useful to identify concepts with more than 90% of them sharing only three categories, but not as useful for linking words with less than 50% for the most common category. Frequency on the other hand does not help on identifying neither concepts nor linking words, both present a highly skewed distribution, meaning that only a few concepts in the concept maps appear several times in the essays, while the majority appears only once.

A final interesting finding is that most relationships can be found within a single sentence, meaning that the most relevant propositions in a concept map correspond to important sentences in an essay. This also helps explaining the skewed distribution of concepts, which could be explained by the annotators selecting a few relevant concepts and then trying to form meaningful relationships with those concepts, which were found precisely in the sentences those concepts appeared.

Three aspects of the concept maps and the essays were analyzed: Characteristics of the complete essay and concept map (such as the correspondence of their sizes and its correlation with the essay quality), characteristics of the concepts (such as linguistic information and frequency) and characteristics of the relationships (linguistic information, location in text and their sequence). To do this, all essays were parsed identifying grammatical roles such as noun phrases or verbal phrases, all statistics on their sizes were evaluated and each concept map was analyzed against its essay source.

The results of the analysis showed that the size of the concept maps measured in the number of concepts is correlated with the length of the essay measured in both words and sentences. However, the concept maps' connectedness has a lower correlation to their length and none of them correlates with the grades. This indicates that when humans are asked to summarize the content of an essay, they will include information based on the total information in the essay and not in its quality.

A second finding is that linguistic characteristics such as words' part of speech are useful to identify concepts with more than 90% of them sharing only three categories, but not as useful for linking words with less than 50% for the most common category. Frequency on the other hand does not help on identifying neither concepts nor linking words, both present a highly skewed distribution, meaning that only a few concepts in the concept maps appear several times in the essays, while the majority appears only once.

A final interesting finding is that most relationships can be found within a single sentence, meaning that the most relevant propositions in a concept map correspond to important sentences in an essay. This also helps explaining the skewed distribution of concepts, which could be explained by the annotators selecting a few relevant concepts and then trying to form meaningful relationships with those concepts, which were found precisely in the sentences those concepts appeared.

To conclude, the analysis has shown that several patterns can be found in the way humans summarize text using concept maps, patterns that now have to be exploited by automatic algorithms in order to move towards the possibility of having concept maps extracted from essays.

6 Acknowledgements

This project was supported by Australian Research Council Discovery Project DP0665064, the Becas Chile scholarship program and Universidad Adolfo Ibáñez.

References

Carletta, J. (1996). "Assessing agreement on classification tasks: the kappa statistic." Computational Linguistics 22(2): 249-254.

Dellschaft, K. and S. Staab (2006). On How to Perform a Gold Standard Based Evaluation of Ontology Learning. Proceedings of the 5th International Semantic Web Conference, Athens, GA, USA, Springer.

Emig, J. (1977). "Writing as a Mode of Learning." College Composition and Communication 28: 122-128. en Chang, K., Y.-T. Sung, et al. (2002). "The Effect of Concept Mapping to Enhance Text Comprehension and Summarization." The Journal of Experimental Education 71(1): 5-23.

Hauser, S., M. Nuckles, et al. (2006). Supporting concept mapping for learning from text. Icls '06: Proceedings of the 7th international conference on Learning sciences, International Society of the Learning Sciences.

Novak, J. D. (2007). The theory underlying concept maps and how to construct them, Florida Institute for Human and Machine Cognition.

Novak, J. D. and D. B. Gowin (1984). Learning How To Learn, Cambridge University Press.

Nurit Nathan, E. K. (2004). Text concept mapping: The contribution of mapping characteristics to learning from texts. First International Conference on Concept Mapping, Pamplona, Spain.

Paltridge, B. (2004). "Academic writing." Language Teaching 37: 87-105.

Shen, R., R. Richardson, et al. (2005). Using concept maps in digital libraries as a cross-language resource discovery tool. Proceedings of the 5th Acm/Ieee-Cs joint conference on Digital libraries.

Valerio, A. and D. Leake (2006). Jump-Starting Concept Map Construction With Knowledge Extracted From Documents. Proceedings of the Second International Conference on Concept Mapping.

Villalon, J. and R. A. Calvo (2008). Concept Map Mining: A definition and a framework for its evaluation. Proceedings of the Ieee/Wic/Acm International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia.

Villalon, J. and R. A. Calvo (2009). Concept Extraction from student essays, towards Concept Map Mining. Proceedings of the 9th International Conference on Advanced Learning Technologies, Riga, Latvia.

Villalon, J., P. Kearney, et al. (2008). Glosser: Enhanced feedback for student writing. Proceedings of the International Conference on Advanced Learning Technologies, Santander, Spain.

Wiebe, J., R. Bruce, et al. (1999). Development and use of a gold standard data set for subjectivity classifications. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99).

Zouaq, A. and R. Nkambou (2008). "Building Domain Ontologies from Text for Educational Purposes." Ieee Transactions On Learning Technologies 1: 49-62.