Concept Learning in Text Comprehension

Manas Hardas and Javed Khan

Computer Science Department, Kent State University,
Kent Ohio 44240, USA
{mhardas,javed}@cs.kent.edu

Abstract. This paper presents a mechanism to reverse engineer a cognitive concept association graph (CAG) which is formed by a reader while reading a piece of text. During text comprehension a human reader recognizes some concepts and skips some. The recognized concepts are retained to construct the meaning of the read text while the other concepts are discarded. The concepts which are recognized and discarded vary for every reader because of the differences in the prior knowledge possessed by all readers. We propose a theoretical forward calculation model to predict which concepts are recognized based on the prior knowledge. To demonstrate the truthful existence of this model, we employ a reverse engineered approach to calculate a concept association graph as per the rules defined by the model. An empirical study is conducted of how six readers from an undergraduate class of Computer Networks form a concept association graph given a paragraph of text to read. The model computes a resultant graph which is flexible and can give quantitative insights into the more complex processes involved in human concept learning.

1 Introduction

Text comprehension is a high level cognitive process performed remarkably well by humans. From a computational view point, it is remarkable because humans can understand and comprehend any piece of text they have never seen before and learn completely new concepts from it they never knew before. Human concept learning as defined by Bruner et. al. (1967) is the correct classification of examples into categories. Previous computational models of human concept learning (Tanenbaum 1999, Dietterich et.al. 1997) give very good approximations of this kind of concept learning. However learning concepts from text is unlike learning from examples. Completely new concepts are not learnt by hypotheses induction but by making new associations with prior knowledge. Therefore a cognitive theory of concept construction is needed to explain this process rather than a theory of generalization. Constructivism (Piaget, 1937) is a cognitive theory of learning which explains how concepts are internalization based on previously acquired concepts by assimilation and accommodation. It gives a systematic cognitive model for acquiring new concepts in context of the prior knowledge. Hence we propose that the process of concept learning from text be examined in the light of the cognitive process involved in constructivism.

Text comprehension research in the cognitive sciences domain considers comprehension, from a process model point of view (Kintsch 1988, van den Broek, Risden, Fletcher and Thurlow, 1999, Tzeng, Y., van den Broek, P., Kendeo, P., Lee, C., 2005, Gerrig and McKoon, 1998; Myers and O'Brien, 1998) or from a knowledge point of view (Landauer, Dumais, 1988, 1997). Both approaches to text comprehension necessitate a mathematical model for learning and the involvement of prior knowledge. There is ample evidence for the importance of prior knowledge in learning from texts (Kintsch, E., and Kintsch, W., 1995, McKeown, M. G., et al., 1992, Means, M., et al., 1985, Schneider, W., et al. 1990). As observed by Verhoeven, L. and Perfetti, C. (2008), over the past decade, research on text comprehension has moved towards models in which connectionist memory-based and constructivist aspects of comprehension are more integrated.
There are two main contributions of this paper. The first is a mechanistic model of the cognitive processes involved in concept learning in text comprehension. The process of text comprehension is defined as concept recognition and concept association. During human text comprehension the CAG goes through a series of incremental changes as specific concepts are recognized or discarded depending upon the reader's prior knowledge. This paper proposes a computational model for the two processes.
The second main contribution is a reverse engineered approach to the model for obtaining the concept association graph. When a person reads text, the CAG which is formed cannot be known beforehand. Hence we need a reverse engineered approach to find the CAG from the data generated by the subjects. An empirical study is conducted in which reader drawn CAGs are fed into a constraint satisfaction system which computes the CAG for a reader or a group of readers. This novel method to compute the comprehensive graph can be efficiently used to comment on the state of learning for a reader or a group of readers.

2 Computational Model

2.1 Knowledge Representation

The knowledge representation is a concept association graph (CAG). The graph consists of nodes which represent concepts and the edges between the nodes signify the association between concepts. The strength of the association is given by an association strength. Association strength is positive and can also be negative in some special circumstances. Any CAG, for example T, has a set of concepts and a set of associations represented by tuple $C_T = [c_1, c_2, c_3, ...]$ and $A_T = [l_{c_1,c_2}, l_{c_1,c_3}, l_{c_1,c_4}, ...]$ respectively.

Figure 1. shows an example of a simple concept association graph. From the graph it can be seen that the concept "ethernet" is most strongly associated with "CSMA" because of it high association strength. Similarly "LAN" and "CSMA" are the two most weakly associated concepts. The time line, represented by t=1, t=2 and so on, is the order in which the concepts were acquired. A lower time value means that the concepts were acquired relatively earlier in learning. The

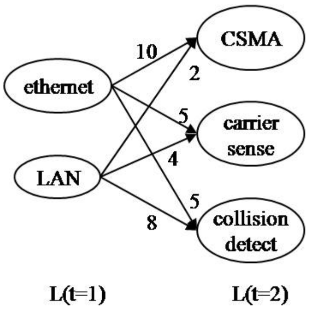


Fig. 1: Example of a simple concept association graph (CAG)

semantic of the association strengths is provided by the theory of constructivism. For the concepts which are acquired at t=2, the concepts acquired at t=1 are considered as previous knowledge. Since according to the theory all new concepts can only be acquired in the context of previous concepts, the strength of the association signifies the importance of the existence of a prior concept to learn a particular new concept. For example, to learn concept “CSMA”, it is more important to learn “ethernet” than it is to learn “LAN” because the association strength between ethernet-CSMA (10) is much greater than that between LAN-CSMA (2).

2.2 Terminology

1. $L(t)$ is defined as the current learned graph at time “t”. It is the graph formed by the reader from episodic memory of the text. The graph goes through a series of changes during text comprehension and is continuously evolving as $L(t \to \infty)$.
2. Z is the graph which represents the global prior knowledge a reader possesses i.e. the non text/domain specific knowledge. It is assumed for computational purposes, that Z has the association information for all new concepts. So whenever a reader is confronted by a totally new concept not present in $L(t)$, the new concept is acquired in $L(t+1)$ by getting the association information from Z.
3. $S(t)$ is a series of graphs each representing the newly introduced concepts in each comprehension episode. Based on the prior knowledge some concepts from this graph are recognized while some are discarded. The concepts which are recognized are then associated with concepts from $L(t)$ to get $L(t+1)$, refer Figure 2.
4. Learning: It is a series of ’comprehension episodes’, each a two step process. Latent concept recognition: From the set of presented concepts, some are recognized while some are not. Latent concept association: From the set of recognized concepts, the concepts are associated with previously known concepts.

3 CAG transition process

In the given model we assume two instances of CAG. The first one is the initially learned CAG namely $L(t = 1)$ represented by concept set $C_{L(t=1)}$ and association set $A_{L(t=1)}$. Learning progresses through a set of learning episodes. It starts with $L(t = 1)$ and then incrementally constructs $L(t = 2, 3, 4, ...)$. In each learning episode a small graph $S(t)$ is presented. $S(t)$ is the graph for every new sentence at time t. It too has a concept set $C_{S(t=1)}$ and an association set $A_{S(t=1)}$. Generally learning episode can range from reading a sentence, or a part of a sentence or simply a word. By the immediacy assumption we consider the smallest unit of a learning episode as a sentence. $S(t = 1)$ includes novel concepts and associations which are not in $L(t = 1)$ and well as known elements.
The second instance of CAG is Z. Z provides a learner with the connections between the current $L(t = 1)$ and the newly acquired concepts from $S(t = 1)$. By definition $L(t = 1)$ cannot have association information that can connect the new concepts in $S(t = 1)$ to that of $L(t = 1)$. Thus the model requires an imaginary CAG namely Z to provide the learner with some basis of computation to discover new concepts and attach them to current $L(t)$. Whenever a new concept is presented to the learner the association information to connect the concept to $L(t = 1)$ is acquired from Z. It is assumed that Z has all the concepts and association information.

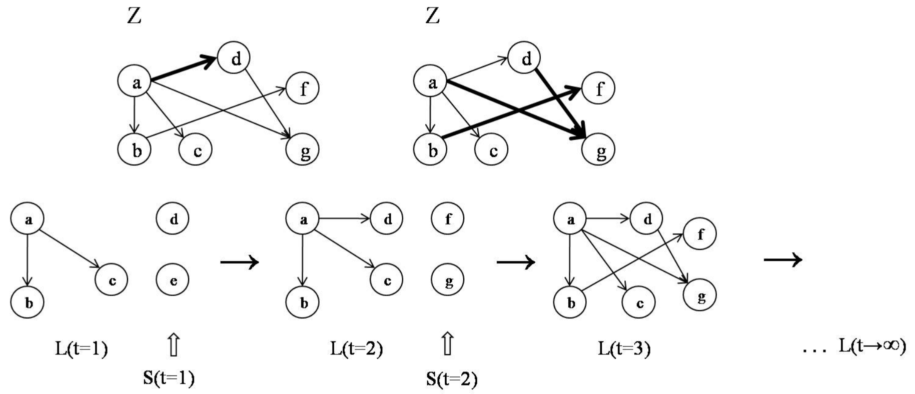


Fig. 2: CAG transition

From the example it can be seen that in the first learning episode when concepts “d” and “e” are presented, the connection between “a” and “d” is obtained from Z. Concept “e” is not present in Z and therefore has no connectivity information. Hence it is discarded while forming the learned graph $L(t = 2)$. In the second episode when concepts “f” and “g” are presented, again the connection information is obtained from Z and $L(t = 3)$ if formed. Fig. 2 shows only the relevant part of Z w.r.t. this example.

4 Processing in a learning episode

As mentioned in the previous section a learning episode is made up of two distinct processes as detailed below.

4.1 Latent concept recognition

The first process is called latent concept recognition. A new concept graph denoted by $S(t)$ is presented out of which some concepts are recognized based on the prior knowledge and some are not. Let the $L(t)$ be the learned CAG denoted by its concept set $C_{L(t)}$ and association set $A_{L(t)}$. A new sentence $S(t)$, presented at step t, has a finite set of discrete concepts $C_{S(t)}$ and association set $A_{S(t)}$. The set may contain already learned (i.e. which are already in $L(t)$) or new concepts (i.e. which are not already in $L(t)$). The new concepts from $S(t)$ which are recognized by the learner are called as "latent concepts" and denoted by the set $C_{lat(t)}$, where $C_{lat(t)} \subset C_{S(t)}$. The latent concept set $C_{lat(t)}$ is formed by evaluating a comprehension function which returns comprehension strength (h_i) for each concept "i" in $S(t)$. The concepts for which h_i exceeds a certain threshold h_T are added to $C_{lat(t)}$ i.e. are recognized.

$$C_{lat(t)} = [i | i \in C_{S(t)}; h_i > h_T] \tag{1}$$

The comprehension strength for a node "i" is function of the association strengths of the links between concept "i" and prior concepts in $L(t)$. It is computed as follows, $h_i = f(l_{s,i}); \forall s \in C_{L(t)}$

We assume a linear relationship between the comprehension strength and the threshold coefficient analogous to linear weighted sum activation function in a simple artificial neural network. It is possible that a nonlinear function holds true, but that is a part of another discussion. Therefore,

$$h_i = \sum_{s \in C_{L(t)}} l_{s,i} \tag{2}$$

An example calculation of the comprehension strength for each concept in $C_{S(t)}$ is shown in Fig. 3. Assume that the association strength between the concepts in $L(t)$ and $S(t)$ are known. Let the threshold coefficient for this particular example be, $h_T = 10$. Calculating the individual comprehension strengths for individual concepts in $C_{S(t=1)}$ we have,

1. $h_{CSMA} = l_{ethernet,CSMA} + l_{LAN,CSMA} = 12 > h_T(10)$
2. $h_{carrier\ sense} = l_{ethernet,carrier\ sense} + l_{LAN,carrier\ sense} = 9 < h_T(10)$
3. $h_{collision\ detect} = l_{ethernet,collision\ detect} + l_{LAN,collision\ detect} = 13 > h_T(10)$

Since comprehension threshold for "carrier sense" is below the threshold it is not recognized and not included in $C_{lat(t=1)}$. After the step of latent concept recognition the set consists of $C_{lat(t=1)} = (CSMA, collision\ detect)$.

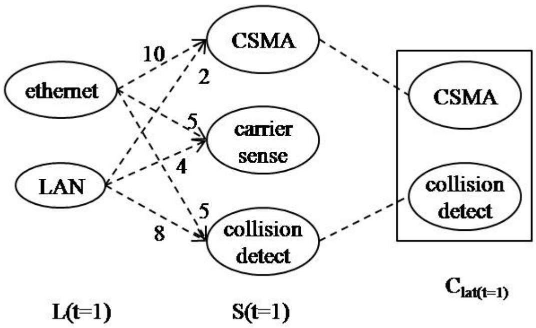


Fig. 3: Latent concept recognition

4.2 Latent concept association

The second process associates the latent concepts in $C_{lat(t)}$ with the concept(s) in the learned set $L(t)$ to form $L(t+1)$. The set of latent associations is denoted by $A_{lat(t)}$, where $A_{lat(t)} \subset A_Z$. The latent association set if formed by evaluating an association function which gives the association strengths $a_{i,j}$ for a concept i and j. The association strength $a_{i,j}$ is simply the scalar link strength of $l_{i,j}$. All the associations with strengths greater than a certain threshold a_T are included.

$$A_{lat(t)} = [l_{i,j}; \forall i \in C_{L(t)}, \forall j \in C_{lat(t)} | a_{i,j} \geq a_T] \tag{3}$$

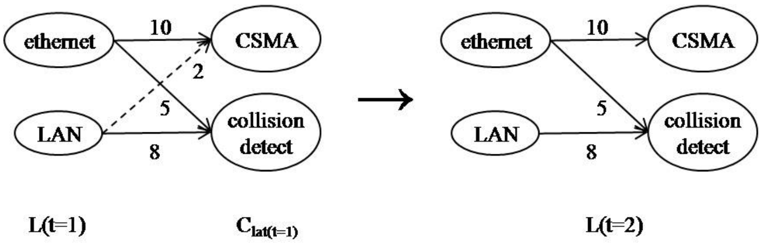


Fig. 4: Latent concept association

If we assume the association threshold $a_T = 5$, then from the figure it can be seen that the association between "LAN" and "CSMA" is dropped because it is less than the association threshold $a_T(5) > a_{LAN,CSMA}(2)$. After the process of recognition and association the concept map evolves from $L(t)$ to $L(t+1)$ represented by concept set and association set computed as follows,

$$C_{L(t+1)} = C_{L(t)} \cup C_{lat(t)} \;\; \& \;\; A_{L(t+1)} = A_{L(t)} \cup A_{lat(t)} \tag{4}$$

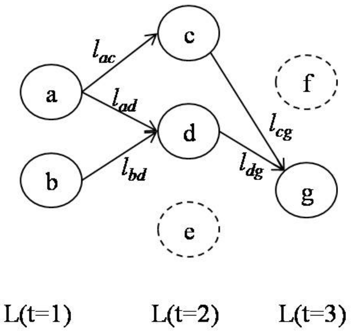


Fig. 5: Example of a learned CAG at time t=3

5 Reverse engineering the association strengths

We present a constraint satisfaction model to calculate the association strengths for the example CAG as shown in Figure 5. Since the concepts (c, d and g) are recognized, it means that the comprehension strengths of each of the nodes is greater than the comprehension threshold. This can be represented by a set of linear equations called as recognition threshold (h_T) equations.

h_T equations	a_T equations
$l_{ac} + l_{bc} \geq h_T$	$l_{ac} \geq a_T, l_{ac} > 0$
$l_{ad} + l_{bd} \geq h_T$	$l_{bc} \leq a_T, l_{bc} > 0$
$l_{ae} + l_{be} \leq h_T$	$l_{ad} \geq a_T, l_{ad} > 0$
$l_{af} + l_{bf} + l_{cf} + l_{df} \leq h_T$	$l_{bd} \geq a_T, l_{bd} > 0$
$l_{ag} + l_{bg} + l_{cg} + l_{dg} \geq h_T$	$l_{ag} \leq a_T, l_{ag} > 0$
	$l_{bg} \leq a_T, l_{bg} > 0$
	$l_{cg} \geq a_T, l_{cg} > 0$
	$l_{dg} \geq a_T, l_{dg} > 0$

It is seen from the graph that concepts “e” and “f” are not recognized. Therefore no associations exist for them. Also, the links l_{bc}, l_{ag} and l_{bg} are less than the threshold, therefore these links are not present in the learned graph. In this discussion we do not consider associations between concepts recognized at the same time. So we form the association threshold (a_T) equations. The h_T equations are the ones that constrain the recognition of concept while the a_T equations are the ones that constrain the association of concepts. It may happen that the association strengths between concepts in $L(t)$ and $C_{S(t)} - C_{lat(t)}$ maybe greater than the association threshold for example, maybe $l_{ae} > a_T$. But since the summation of l_{ae} and l_{be} is not greater than h_T, concept “e” is not recognized. And since concept “e” is not recognized we do not put a_T constraints on it associations. All links are constrained to be greater than 0.

The matrix representation for the equations as a linear programming problem is as follows; min f*x subject to constraints $A * x \leq b$ where x is the vector of variables and f, A and b are the coefficient matrices for the objective function, equation set and the result. Since we are not trying to optimize an objective function all the members of f are set to 0. An important observation here is that, the recognition and association thresholds (h_T and a_T) are variable and factored into the coefficient matrix for the equation set. Fig.6 shows an example matrix representation. Solving this gives the association strengths for all the associations for a given CAG.

$$\min \quad f*x \quad subject \quad to: \begin{bmatrix} & l_{aa}\ l_{ab}\ l_{ac}\ l_{bb}\ \cdots\ h_T\ a_T \\ h_T\ eqn & \left\{\begin{matrix}0\ 0\ 1\ 0 \ldots 1\ 0 \\ 0\ 1\ 1\ 0 \ldots 1\ 0\end{matrix}\right\} \\ a_T\ eqn & \left\{\begin{matrix}1\ 0\ 1\ 0 \ldots 0\ 1 \\ 1\ 0\ 1\ 0 \ldots 0\ 1 \\ \ldots\end{matrix}\right\} \end{bmatrix}_{m\times n} \times \begin{bmatrix} l_{aa} \\ l_{ab} \\ l_{ac} \\ \cdots \\ h_T \\ a_T \end{bmatrix} = \overset{b}{\begin{bmatrix} 0 \\ 0 \\ 0 \\ \ldots \end{bmatrix}}_{m\times 1}$$

Fig. 6: Matrix representation

6 Finding and analyzing the solution CAG

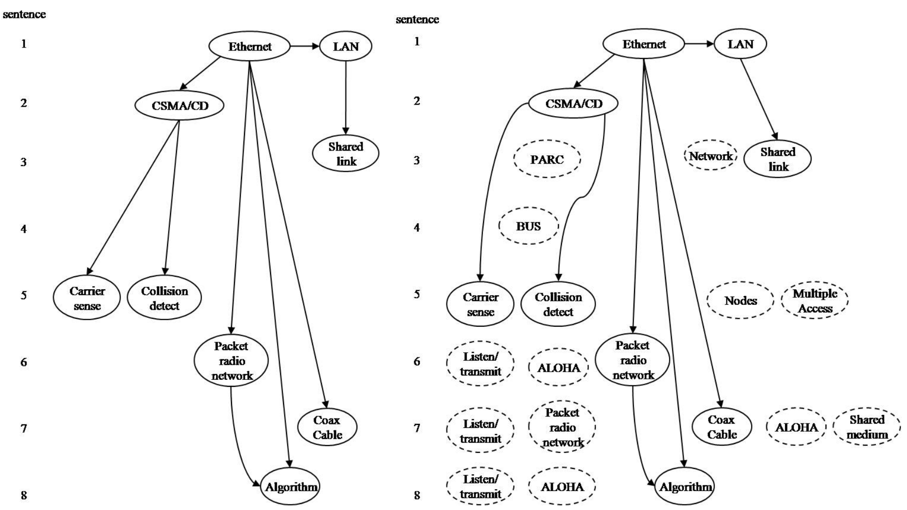
6.1 Experiment setup

To find the comprehensive complex graph that can explain the concept learning for a particular example text an experiment is conducted in the classroom setting with a group of six students in the undergraduate “Computer Networks” class. Subjects were given a paragraph of text about the concept “Ethernet” from the standard text book prescribed for that class and were asked to go through each sentence in the paragraph and simultaneously identify each concept in the sentence and progressively draw CAGs. The paragraph contained 8 sentences, so the concept learning activity was divided into 8 learning episodes. By the end of the eighth episode the students had drawn CAGs from t=1-8 using the concepts from the text.

6.2 Observations

Figure 7 (a) shows the concept graph drawn by one of the students.

The student drawn CAGs are used to reconstruct CAGs for all students indicating the concepts which were recognized and those which were not. To construct these graphs we first have to find $C_{S(t)}$ for t=1-8. This is done by collecting concepts in t=1 to 8 for every student. For example, at t=3 the possible set of recognized concepts which covers all students is, $C_{S(t=3)} = (PARC, Network, Shared\ link)$. Out of these student recognized only the concept “Shared link”. Therefore,



(a) Student drawn CAG (b) Reconstructed CAG

Fig. 7: Reconstructing student drawn CAG

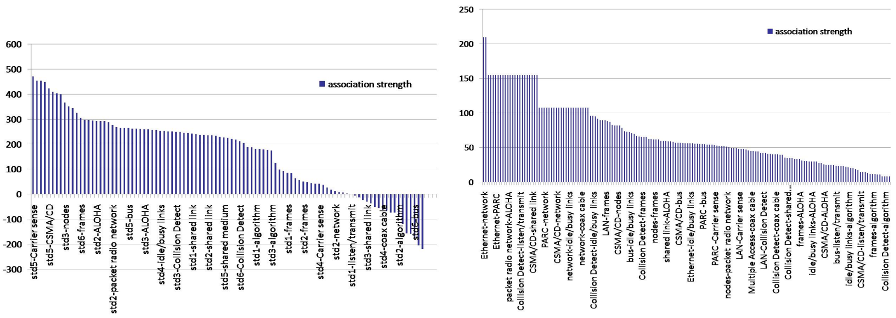
$C_{lat(t=3)} = (Shared\ link)$ and $C_{S(t=3)} - C_{lat(t=3)} = (PARC, Network)$. Figure 7(b) shows the reconstruction of $C_{S(t=3)}$ and $C_{lat(t=3)}$ for a student. After determining $C_{S(t)}$ and $C_{lat(t)}$ for every sentence for every student the concept maps are reconstructed to include the recognized as well as the unrecognized concepts.

Once CAGs for all students are reconstructed, they are converted into a set of linear equations and solved to obtain the values of the association strengths which satisfy all the constraints. The result is a fully connected CAG called as solution CAG which can mathematically explain the concept learning for all students according to the laws of concept recognition and concept association as specified before.

6.3 Analysis of solution CAG

The solution CAG is a fully connected graph between concepts from the text and special “hidden” nodes. These types of nodes are introduced at the stage of reconstructing a particular student CAG. A single hidden node is inserted at time t=1 for every student. This hidden node signifies the background knowledge of a particular student. For an experiment like this, it is impossible to actually construct a graph of a student’s entire background knowledge. There exists no method which can accurately hypothesize a person’s concept knowledge graph. Hence we assume that all the background concept knowledge possessed by a student is encompassed in a single hidden node namely “std1” for student 1 and so on. Thus the resultant CAG contains six such hidden nodes, one for each student. The hidden nodes have connections to all the rest of the concept nodes in the CAG. The associations from a hidden node can have positive as well as negative strengths. This can be intuitively explained as, sometimes the student's background knowledge helps in learning new concepts whereas sometimes it is found to be an obstacle. If the association from the hidden node to a concept node is positive then it implies that the hidden node is beneficial in learning the new concept whereas negative association strength implies that the hidden node is actually detrimental to learning the new concept. A zero strength association from a hidden node implies it's neither beneficial nor detrimental to learning. The existence of hidden nodes also helps in solving another known problem of learning the XOR function using this model since the hidden node associations are allowed to take negative values. The solution CAG is actually the imaginary CAG namely Z, which we assumed to have contained all the connectivity and association strength information. Z can thus be calculated by reverse engineering the observed student drawn CAGs.

Association strength distribution In this section we analyze the distribution of association strengths. Figure 8(a) shows the sorted association strengths between hidden nodes and concept nodes and Figure 8(b) shows just between concepts. Some of the links between hidden nodes and concepts have negative strength link but most have positive strength indicating that more often than not background knowledge helps in learning new concepts. This plot can be used to determine which of the associations are most important and need reinforcement.

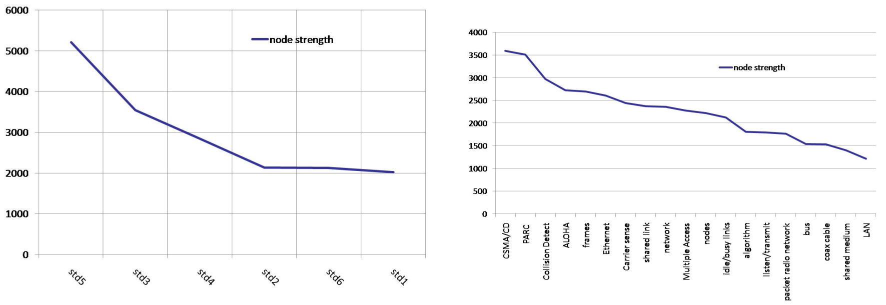


(a) Between hidden and concept nodes

(b) Between concept nodes

Fig. 8: Association strength distribution

Node strength distribution The node strength of a node is calculated by summing up all the association strengths to that particular node. Figure 9(a) shows the sorted node strengths for the hidden nodes. It is seen that std5 hidden node has highest node strength.



(a) For hidden nodes

(b) For concept nodes

Fig. 9: Nodes strength distribution

Figure 9(b) shows the sorted node strengths for the actual concepts. “CSMA/CD” has the highest node strength signifying it importance in comprehension of this particular paragraph of text. This graph gives us an idea about which concepts are central in comprehending this paragraph of text. As seen from the figure the concepts “CSMA/CD”, “Collision detect”, “Aloha”, “frames” etc. are much more central in comprehending the concept of “ethernet” than say, “bus”, “coax cable” or “shared medium”.

Correlation with h_T Each student is assumed to have a variable h_T and a_T. These variables are factored into the problem while constructing the equations and coefficient matrices. The variable h_T for each student signifies the difficulty or ease of a student in comprehending the particular paragraph of text. A lower value of h_T implies that the student possibly has a lower threshold for learning new concepts. Meaning the student is more easily capable of learning new concepts than one with a high threshold.

To observe this we simply plot the correlation between the threshold h_T for each of the students and the number of concepts recognized (n) by the student. Figure 10 table shows the exact values h_T against n and the plot. As expected there is negative correlation between the two variable equal to -0.172.

7 Conclusion and potential directions

In this paper we proposed a computational model for computing the concept associating graph which is formed during human text comprehension. A study is conducted to explain concept learning for a group of six readers, which can be extrapolated to any number of subjects. We perform simple graph analysis on the obtained CAG to find peculiar characteristics a cognitive concept graph might have. Some of the questions we are able to answer are, which associations are important than others, what distribution do association strengths have, which concept is central in comprehending a particular topic, which student has the

student	h_T	n
1	554.027	17
2	582.764	16
3	622.879	11
4	631.583	10
5	775.794	13
6	574.077	9

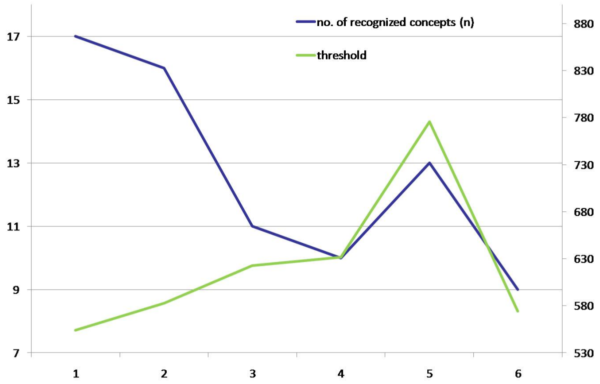


Fig. 10: Correlation between h_T and n for six students is -0.172

maximum chance of learning new concepts and what is the significance of the threshold coefficient in learning new concepts. The CAG can be subjected to rigorous complex network analysis to derive other interesting inferences. From the theory it is clear that prior concepts play an important role in learning new concepts. As a potential direction we plan to study how the sequence of concepts affects concept learning.

References

1. Dietterich, T., Lathrop, R., & Lozano-Perez, T. (1997). Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence 89(1-2), 31-71.
2. Joshua B. Tenenbaum, Bayesian modeling of human concept learning, Proceedings of the 1998 conference on Advances in neural information processing systems II, p.59-65, July 1999.
3. Kintsch, Walter (2001). Predication. Cognitive Science: A Multidisciplinary Journal, 25 (2), 173-202.
4. Kintsch, Walter. The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model. Psychological Review, v95 n2 p163-82 Apr 1988.
5. Kintsch, Walter. Text Comprehension, Memory, and Learning. American Psychologist, v49 n4 p294-303 Apr 1994.
6. Kintsch, Walter; Van Dijk, Teun A. Toward a Model of Text Comprehension and Production. Psychological Review, v85 n5 p363-94 Sep 1978.
7. Nick Chater, Christopher D. Manning, Probabilistic Models of Language Processing and Acquisition. Trends in Cognitive Sciences In Special issue: Probabilistic models of cognition, Vol. 10, No. 7. (July 2006), pp. 335-344.
8. Thomas K. Landauer , Darrell Laham , Peter Foltz, Learning human-like knowledge by singular value decomposition: a progress report, Proceedings of the 1997 conference on Advances in neural information processing systems 10, p.45-51, July 1998, Denver, Colorado, United States.