

A COMPUTATIONAL MIMICRY OF THE KNOWLEDGE AUGMENTATION  
PROCESS IN COMPREHENSION BASED LEARNING

A dissertation submitted  
to Kent State University in partial  
fulfillment of the requirements for the  
degree of Doctor of Philosophy

by

Amal Babour

November 2017

Dissertation written by

Amal Babour

B.Sc., King Abdulaziz University, Jeddah, Saudi Arabia, 2002

M.S., Cairo University, Cairo, Egypt, 2007

PhD., Kent State University, USA, 2017

Approved by

\_\_\_\_\_, Chair, Doctoral Dissertation Committee

\_\_\_\_\_, Members, Doctoral Dissertation Committee

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

Accepted by

\_\_\_\_\_, Chair, Department of Computer Science

\_\_\_\_\_, Dean, College of Arts and Sciences

## TABLE OF CONTENTS

<b>LIST OF FIGURES .....</b>	<b>VI</b>
<b>LIST OF TABLES .....</b>	<b>IX</b>
<b>DEDICATION.....</b>	<b>XI</b>
<b>ACKNOWLEDGEMENTS .....</b>	<b>1</b>
<b>CHAPTER 1 INTRODUCTION .....</b>	<b>4</b>
1.1 Text comprehension and Prose comprehension .....	4
1.2 Dissertation contributions .....	5
1.3 Thesis organization .....	8
<b>CHAPTER 2 KNOWLEDGE INDUCTION PROCESS OF PROSE</b>	
<b>COMPREHENSION .....</b>	<b>10</b>
2.1 Concept representation generation .....	14
2.2 Reference Consultation .....	17
2.2.1 Extracting the highest familiarity knowledge from a reference text.....	17
2.2.2 Extracting knowledge from an Ontology Engine.....	22
2.3 Summary .....	28
<b>CHAPTER 3 KNOWLEDGE DISTILLATION PROCESS OF PROSE</b>	
<b>COMPREHENSION .....</b>	<b>30</b>
3.1 Goodness Paths(s) notion .....	33
3.2 Equivalent Electrical Circuit (EEC).....	35
3.3 Goodness Function Measurement .....	40
3.4 Grading Process.....	41

3.4.1	Alpha Label graph extraction.....	42
3.4.2	Current Calculation .....	44
3.4.3	Skimmed Process .....	48
3.4.4	Skimmed Illuminated Knowledge Graph SIKG generator .....	50
3.5	Summary .....	51
<b>CHAPTER 4 COMPUTATIONAL MODEL EVALUATION, EXPERIMENTS</b>		
<b>AND RESULTS.....</b>		<b>53</b>
4.1	Comprehension Model Evaluation.....	53
4.1.1	Information content.....	54
4.1.2	Knowledge Organization.....	55
4.1.3	Experiment .....	56
4.1.4	Comprehension Efficiency .....	69
4.1.5	Human experiments Analysis.....	77
4.2	Summary .....	83
<b>CHAPTER 5 CONCLUSIONS.....</b>		<b>85</b>
5.1	Does the proposed comprehension engine help in improving the quality of comprehension? .....	85
5.2	Does the proposed comprehension engine effect in saving time of learning? .....	87
5.3	Criticize and Challenges.....	87
5.4	Future work .....	88
5.5	Summary .....	89
<b>APPENDIX A EXAMPLE OF THE DATA USED IN THE EXPERIMENT .....</b>		<b>91</b>

**REFERENCES..... 118**

## LIST OF FIGURES

Figure 2.1. (a) Example of an Illuminated Knowledge Graph, (b), (c), (d) and (e) are examples of Geometric path. (b), (c), and (e) are examples of Knowledge Paths. ...	13
Figure 2.2. Example of different types of Steiner trees. ....	18
Figure 2.3. Terminal to Terminal Steiner Tree algorithm. ....	20
Figure 2.4. Example of a Terminal to Terminal Steiner Tree.....	21
Figure 2.5. Discovering OE- Knowledge-Paths algorithm.....	23
Figure 2.6. Example of an OE-Knowledge-Path. ....	25
Figure 2.7. Comprehension Engine. ....	27
Figure 2.8. The process of connecting the prose concepts $C_L$ using reference texts and ontology engine. (a) a set of prose concepts $C_L$ . (b) Knowledge path K from the prose LTX. (c) Knowledge path K using RTX1. (d) Knowledge path K using Ontology Engine OE. (e) Knowledge path K using RTX2. (f) Knowledge path K using RTX3.....	28
Figure 3.1. Examples of different types of paths between two concepts.....	31
Figure 3.2. An Example of multiple knowledge paths between ethane and carbon-carbon. ....	32
Figure 3.3. Two simple graphs where both (a) shortest-hubs path and (b) network flow fail for discovering “Alpha Knowledge Pathway” between two concepts”.....	35
Figure 3.4. Equivalent Electrical circuits EEC graphs. ....	37
Figure 3.5. Pathological Cases in Random Walk Interpretation.....	39
Figure 3.6. Comprehension Engine. ....	42

Figure 3.7. Alpha Label graph $g$ extraction Algorithm. ....	43
Figure 3.8. An example of a knowledge path. ....	44
Figure 3.9. Example showing resistance, voltage, and current in three different Alpha Label graphs $g$ . ....	46
Figure 3.10. Example showing the “Alpha Knowledge Pathway” $K$ ’ in the three Alpha Label graphs $g$ in Figure 3.9. ....	46
Figure 3.11. Example showing the “Alpha Knowledge Pathway” $K$ ’ in Example 1 based on sentences budget. ....	50
Figure 3.12. Skimmed Illuminated Knowledge Graph SIKG of Figure 3.10. ....	51
Figure 4.1. Prose Concepts connectivity per the knowledge graphs. ....	60
Figure 4.2. Information growth rate $\lambda$ per the knowledge graphs. ....	61
Figure 4.3. Information overload rate $\gamma$ per the knowledge graphs. ....	62
Figure 4.4. Entropy $\delta$ per the knowledge graphs. ....	63
Figure 4.5. Break down of the Entropy $\delta$ per the knowledge graphs for the three LTX. ....	64
Figure 4.6. Cluster Coefficient $\beta$ per the knowledge graphs. ....	65
Figure 4.7. Density $\rho$ per the knowledge graphs. ....	66
Figure 4.8. Example of the variance of the illuminating value $h_i$ in the learning process over three phases. ....	72
Figure 4.9. Illustration values per phases for concept carbonization in (a) Knowledge Induction Process (Illuminated Knowledge Graph IKG) and (b) Knowledge Distillation Process (Skimmed Illuminated Knowledge Graph SIKG). ....	73

Figure 4.10. Correlation among phases $\Theta_i$ and the concept illumination value $h_i$ for LTX1, LTX2, and LTX3 in the knowledge graphs.....	75
Figure 4.11. Illustration values per phases for LTX1, LTX2, and LTX3 (a) Knowledge Induction Process (Illuminated Knowledge Graph IKG) and (b) Knowledge Distillation Process (Skimmed Illuminated Knowledge Graph SIKG).....	76
Figure 4.12. The average of the recognized concepts and the recognized relations in the three proses.....	81
Figure 4.13. The average of the incremental enhancement of the recognized concepts and the recognized relations in the three proses.....	82
Figure 4.14. The distribution of the knowledge paths length to the correct recognized relations in the three proses.....	83



## LIST OF TABLES

Table 2.1. Relation Structure between any pair of Concepts.....	16
Table 3.1. The Delivered Current for All the Knowledge Paths in Figure 3.9.....	49
Table 3.2. The Delivered Current for All the Knowledge Paths in Figure 3.8 sorting in descending order.....	50
Table 4.1. List of the proses used in the experiment .....	57
Table 4.2. Break Down of the readable reference texts in each LTX.....	58
Table 4.3. The average time for reading the prose, the references and finding the highest familiarity sentential relations connecting the prose concepts in each reference in (h:m:s).....	59
Table 4.4. Break Down of the total number of sentential relation and concepts in the three proses .....	60
Table 4.5. The average time for reading the prose, the references, finding the highest familiarity sentential relations connecting the prose concepts in each reference and finding the Alpha Knowledge Pathway between each pair in (h:m:s) .....	67
Table 4.6. Break Down of the total number of the sentential relations and concepts in the three proses .....	67
Table 4.7. Basic graph metrics analysis for the prose knowledge graph $G_0$ and the Skimmed Illuminated Knowledge Graph SIKG.....	68
Table 4.8. Diameter of the Knowledge Graphs .....	69

Table 4.9. Rubric for incremental enhancement for knowledge comprehension ..... 78

## **DEDICATION**

Optional dedication page.

## **ACKNOWLEDGEMENTS**

Acknowledge those who helped or supported you in finishing this dissertation/thesis.

Your Name

Defense Date, Kent, Ohio

## Notations

Notation	Meaning
LTX	The learnable prose
$G_{LTX}/G_0$	The knowledge graph of the prose
$IKG_i, 0 < i \leq n, n=final$	Illuminated Knowledge Graph i
$IKG = \{IKG_1, IKG_2, \dots, IKG_{final}\}$	A set of Illuminated Knowledge Graph
$c_i$	Concept i
$C_L = \{c_i, \dots, c_n\}$	List of learnable noun concepts in the prose
$ C $	The number of concepts in the knowledge graph
$s_{i,x}$	The $x^{th}$ sense for concept i
$E = \{e_1, e_2, \dots\}$	A set of edges in the knowledge graph, each represents a sentential relation among a pair of concepts.
$ E $	The number of edges in the knowledge graph
$RTX_i$	A reference text i
$RTX = \{RTX_1, RTX_2, \dots, RTX_n\}$	A set of reference texts
OE	Ontology Engine
$Z = \{z_1, z_2, \dots\}$	A set of geometric paths
K	A sequence of edges constructing a Knowledge Path, where $K \in Z$
KG	Syntactical Explicit Graph generator function
$G_{Ri}$	A reference knowledge graph i
$t_b, b=1,2,\dots, n$	A set of words between concept $c_i$ and $c_j$
L	The maximum allowed distance between concept $c_i$ and $c_j$ and in a sentence
$w_{ij}$	The familiarity value of the of sentential relation between concept $c_i$ and $c_j$
$f_{i,j} / f_i$	the frequency of the relation type between concept $c_i$ and $c_j$ or the frequency of concept $c_i$ extracted from "Gutenberg Project"
MST	Minimum Steiner Tree
TST	Traffic Steiner Tree
TTST	Terminal to Terminal Steiner Tree
$G_{Ui}$	The name of the graph/tree extracted by TTST( )
$G_{temp}$	Temporary knowledge graph
$KP_{OE}()$	OE-knowledge-paths algorithm
R	A dictionary of all relations between concepts in the Ontology Engine OE
$\alpha$	The maximum length of an OE-knowledge path
$G_{Wi}$	The name of the graph created by $KP_{OE}()$
K'	Alpha Knowledge Pathway
SIKG	Skimmed Illuminated Knowledge Graph
g	Alpha Label graph

ER	Effective Resistance
RW	Random Walk
$\Omega_i$	The total resistance of the edges incident from concept $c_i$ in the equivalent electrical circuit
$\Omega_{i,j}$	The resistance of the edges incident from concept $c_i$ to $c_j$ in the equivalent electrical circuit
$V_i$	The voltage of concepts $c_i$ in the equivalent electrical circuit
$I_{i,j}$	The current flow between each two concepts $c_i$ and $c_j$ in the equivalent electrical circuit
ST	Sentetext generator
$\lambda$	Knowledge growth
$ G $	The size of the knowledge graph
$\gamma$	Knowledge overload
$\delta$	Graph entropy
$p_i$	The probability of the concept $c_i$ .
$d_i$	The number of the sentential relation of concept $c_i$
$\beta$	Cluster Coefficient
$NIC_i$	The neighbors' interconnections coefficient of concept $c_i$
$\rho$	Graph density
$\Theta_i$	Phase transition
$h(\Theta_i)$	A concept illumination value at phase $i$
$H = \{h_1, h_2, \dots, h_n\}$	Vector of all the concept illumination values
$ H $	The summation of $h$ for each concept $c_i$ in the list of prose concepts $C_L$
$a_{i,j}$	The value of the sentential relation strength (weight) between $c_i$ and $c_j$ .
A	A matrix with $a_{i,j}$ elements
u	The budget of the number of sentences

## CHAPTER 1

### Introduction

#### 1.1 Text comprehension and Prose comprehension

Text comprehension is a very distinguished form of knowledge learning acquisition due to the property of text as a highly-structured method of knowledge delivery. A well written text certainly has cues for building up a coherent representation of the given text and for playing an important role in the process of text comprehension. Thus, *Text comprehension* can be thought of as the process of acquiring knowledge represented in a set of concepts and a set of associations among them derived from the text itself. However, sophisticated text is often rich with complex concepts covering very specialized meanings and associations that can be subtle, difficult to comprehend from the text, and that need the use of external sources to achieve the goal of comprehension. In the dissertation, we identify this type of text as prose.

For a prose text, the reader is not expected to comprehend it based on the knowledge in prose alone. Therefore, significant external knowledge, often called prior knowledge, is needed (Al Madi, 2014; Kintsch, 2004). Further, prior knowledge is different from one to another. Sometimes, people may not have the minimal prior knowledge about a specific topic. So, they need to acquire the knowledge they lack prior to reading by utilizing resources such as lexicons and external references (Moravcsik & Kintsch, 1993). By this, it can be stated that in order to fill the gaps left by the prose and achieve the goal of prose comprehension, the reader should make a connection between the prose's contents and his or her prior knowledge which is not inferred from the presented data in the prose. In other

words, prose comprehension involves the flexibility of using different resources of information and integrating them into the text. One example is the integration between an encyclopedia and linguistic information/ dictionaries. However, the choice of the right reference in many areas is still a big challenge. Not only that, but finding and reading the relevant parts from the reference can become too time-consuming (Bergenholtz & Gouws, 2010; Fathi, 2014). What is needed is a cognitive comprehension model that would enhance prose comprehension by reading relevant parts from incremental external references related to the given prose and finding meaningful knowledge associating the prose concepts. As this is indeed a very common phenomenon in human prose comprehension.

Knowledge can be defined as “*the process of knowing, a reflexive process that takes data and information, in a social context ... and generates new data, information, and/or knowledge*” (Sasser, 2004). Furthermore, the incremental use of external references may result in too much knowledge or an overload problem of familiar/unfamiliar, easy/complex, and/or repeated information which may further cause a misleading in prose comprehension. Thus, there is a need to control knowledge overload and provide the reader with the best related familiar knowledge that is easy to understand to enhance prose comprehension. So, it is necessary to decide which information to keep and which to remove.

## **1.2 Dissertation contributions**

In this research, we study text comprehension requesting external consultation. Although the process of comprehension has long been studied, no algorithm level model for comprehension is available. So, the main contribution of this research is to study the algorithms behind prose comprehension. An initial version of algorithms that can be used



to enhance the comprehension is proposed. The result product is a comprehension engine consists of two processes; the Knowledge Induction Process and the Knowledge Distillation Process. The Knowledge Induction Process seeks to increase knowledge by reading incremental external reference texts and finding the highest familiarity knowledge associations among prose concepts. We suggest using steps of algorithms to find the most familiar knowledge with fewest associations among prose concepts using no or the minimum number of external concepts. Therefore, the Knowledge Distillation Process grades all the knowledge associations between each pair of the prose concepts and selects the one which has the most familiar, easiest-to-understand knowledge that can be delivered to the reader. The effective resistance measurement is suggested for grading knowledge associations and selecting the one which has the highest delivered current between the two concepts.

In current psychology, it is understood that there are two types of memories: short-term memory and long-term memory. Long-term memory stores decontextualized knowledge such as alphabets, words, grammar rules and elementary aspects of language. Short-term memory stores the knowledge which is actively being processed. It is usually thought that prior knowledge is stored in long-term memory and during comprehension this knowledge is called upon to make interconnections with the knowledge in short-term memory (obtained from reading the text) to generate a cohesive state of comprehension (Hardas, 2012).

The dissertation answers the following question; can the comprehension engine improve the quality of the comprehension and save time efficiency?

Although there is an extensive number of external references that help the reader to increase his or her knowledge, he or she may struggle in reading a large amount of external references and keep up with the new knowledge read. This back to that the human cognitive abilities have definite limits in several dimensions. For example, Miller found that a person's short-term memory has a limited capacity (Miller, 1956), which means that the reader can only hold a certain number of pieces of knowledge at a time (Hao, 2016). Therefore, he or she may not effectively interconnect the newly-read knowledge stored in the short-term memory with prior knowledge stored in long-term memory (Maria & MacGinitie, 1980; Spiro, 1980). He or she may not organize the knowledge into a mental image while reading (Gagné & Memory, 1978; Levin, 1973; Witte, 1978). Thus, the level of comprehension during the act of reading may not be properly monitored (Baker, 1979; Di Vesta, Hayward, & Orlando, 1979; Owings, Petersen, Bransford, Morris, & Stein, 1980; Paris & Myers, 1981) (Davey, 1983). The proposed comprehension engine can handle these problems and improve the quality of learning by reading an incremental number of external references and finding the best related highest familiarity knowledge that can help the reader to increase knowledge and enhance comprehension.

Human reading rate has been an area of research that has considerable interest (Carver, 1992; Just & Carpenter, 1980; Huey, 1908). Carver's reading model (Carver, 1992) explores the relationship model between reading and comprehension. He theorized that reading involves five basic processes: memorizing, learning, rauding, skimming, and scanning. An individual's reading rate may vary depending on the difficulty level of the material and purpose of the reader (Carver, 1992, 1984). For example, a reader may use

the skimming process when only an overview of the material is necessary, while learning processes that require rereading are utilized when a more thorough comprehension of the material is required. Based on Carver's reading model, the rate for the process of memorization is around 138 words per minute (Wpm), the learning process is around 200 Wpm, the reading process is around 300 Wpm, the skimming process is around 450 Wpm, and the scanning process is around 600 Wpm. For example, if a reader needs to read and comprehend a text consisting of 2000 words, he or she needs around 10 minutes to read it for the purpose of learning. How about reading long texts or more than one text? One of the significant advantages of the proposed comprehension engine is its ability for saving time efficiency. By reading several reference texts and finding the highest familiarity knowledge, this method can aid the reader in increasing his or her knowledge and enhancing comprehension in a short amount of time.

### **1.3 Thesis organization**

Chapter 2 gives the definitions used in the Knowledge Induction Process. In addition, the model used for representing the knowledge is explained, along with the iterative process used for augmenting the highest familiarity knowledge associations that connect the prose concepts from external sources. Chapter 3 presents in detail the definitions used in the Knowledge Distillation Process, the method used for grading the augmented knowledge and selecting the best familiar easy to understand knowledge between each pair of prose concepts. Chapter 4 describes the model used for evaluating the comprehension. Also presented are the experiment, the content material, and the analysis of the results gained by the Knowledge Induction Process and the Knowledge Distillation

Process. In addition, a description of the design and the implementation of our human study experiment and the analysis of its results is included. Chapter 5 summarizes the work, in addition to some of the interesting suggestions for the future work.

## CHAPTER 2

### **Knowledge Induction Process of Prose Comprehension**

This chapter serves to introduce the Knowledge Induction Process's algorithms specifically designed to enhance prose comprehension. The chapter starts by elaborating the definitions used in the Knowledge Induction Process. It then explains the two fundamental techniques of the process. In addition, it explains each crucial step to achieve the task of the Knowledge Induction Process.

Sophisticated prose is often rich with specialized concepts and terminologies that are sensitive and difficult for inexperienced readers to comprehend the knowledge associations among them from the prose itself. Such difficulty can also be observed in readings of many domains, such as science and technology. Additionally, it is believed that the process of prose comprehension involves the integration of concepts with significant external knowledge, which is often called prior knowledge (Babour, Khan, & Nafa, 2016; Khan & Hardas, 2013). However, readers have different levels of prior knowledge; depending on the reader, sometimes they might not even have the minimal prior knowledge about a specific topic. Therefore, they need help through external knowledge sources such as reference texts, ontology engines, dictionaries, papers, or conversations with experts that allow them to compensate for the lack of prior knowledge (Moravcsik & Kintsch, 1993). However, the extensive number of knowledge sources might itself become an obstacle to text comprehension. For example, if the reader decides to read reference texts to enhance

his or her comprehension, he or she might struggle to keep up with the type and the large amounts of reference texts, which can easily be disturbing. Additionally, searching for the relevant needed parts in the reference texts can be extensive and time-consuming. For that reason, the Knowledge Induction Process is designed to substitute the lack of knowledge in the prose by augmenting knowledge associations among the prose concepts by using reference texts and ontology engine as external knowledge sources. We assume the concepts are known and we focus on finding the knowledge associations among them. The process reads the most appropriate parts from relevant reference texts to find the highest familiarity knowledge associations that connect the prose concepts and uses an ontology engine to find lexical knowledge associations among each pair of concepts in order to enhance prose comprehension.

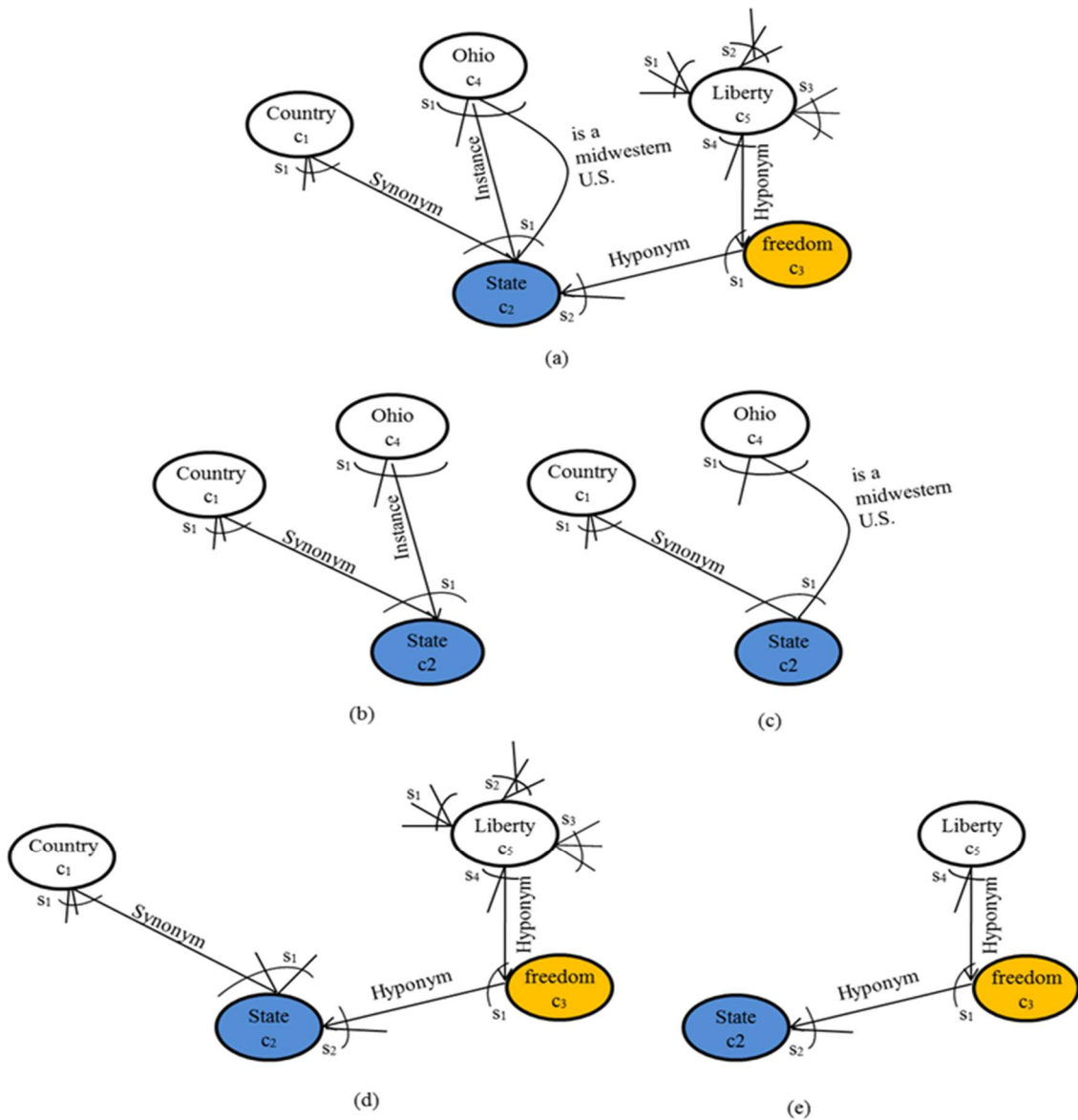
The problem is represented more formally as follows:

**Given** a specified prose  $LTX$  for comprehension, a list of prose concepts  $C_L = \{c_1, \dots, c_n\}$  from  $LTX$ , a set of reference texts  $RTX_i$ , and an ontology engine  $OE$ . **Find** the Illuminated Knowledge Graph  $IKG$  representing the knowledge associations among the prose concepts  $C_L$ .

The *Illuminated Knowledge Graph*  $IKG$  is defined as a graph that provides a capture of the current state of the learning progress by showing the prose concepts  $C_L$  and the relationships between them found by reading the prose  $LTX$ , the relevant parts from reference texts  $RTX_i$ , and the Ontology Engine  $OE$ . It is represented as a directed graph  $IKG = (C, E)$ , where  $C$  is a set of concepts and  $E$  is a set of edges. Each concept  $c_i$  can have one or more senses  $(s_{i,1}, s_{i,2}, \dots, s_{i,x})$  where  $i$  is the concept number and  $x$  is the sense number.

Each edge connects two concepts via a specific sense of each concept and has a type of the following types: Synonym, Hyponym, Hypernym, Meronym, Holonym, Instance or Syntactic explicit. The concept is either in *LTX*, *RTX*, or *OE*, while the edge between any two concepts represents a sentential relation between them. Figure 2.1 (a) shows an example of the Illuminated Knowledge Graph *IKG* with five concepts; *Country*, *State*, *Ohio*, *freedom*, and *Liberty* where *Country*, *Ohio*, and *Liberty* are the  $C_L$  concepts; *State* belongs to a reference text; and *Freedom* belongs to an Ontology Engine.

A *Knowledge Path K* is defined as a path illuminating the relationship between two concepts. It is represented as a sequence of edges that connects a concept  $c_i$  with a concept  $c_j$  in a preserved sense, where  $c_i$  and  $c_j$  are concepts from *LTX*. The intermediate concepts in the path can be external to the list of prose concepts  $C_L$ . Examples of Geometric paths are  $Z$ , (i)  $c_1 - s_{1,1} : \text{Synonym} : s_{2,1} - c_2 - s_{2,1} : \text{Instance} : s_{4,1} - c_4$ , (ii)  $c_1 - s_{1,1} : \text{Synonym} : s_{2,1} - c_2 - s_{2,1} : \text{syntactic explicit} : s_{4,1} - c_4$ , and (iii)  $c_1 - s_{1,1} : \text{Synonym} : s_{2,1} - c_2 - s_{2,2} : \text{Hyponym} : s_{3,1} - c_3 - s_{3,1} : \text{Hyponym} : s_{5,4} - c_5$ , where  $c_1, c_2 \dots$  etc. refer to the concept number.  $s_{1,1}$  is the first sense of the first concept,  $s_{2,1}$  the first sense of the second concept, etc.  $K$  can be extracted from  $Z$ . For example, (i)  $c_1 - s_{1,1} : \text{Synonym} : s_{2,1} - c_2 - s_{2,1} : \text{Instance} : s_{4,1} - c_4$ , (ii)  $c_1 - s_{1,1} : \text{Synonym} : s_{2,1} - c_2 - s_{2,1} : \text{syntactic explicit} : s_{4,1} - c_4$ , and (iii)  $c_2 - s_{2,2} : \text{Hyponym} : s_{3,1} - c_3 - s_{3,1} : \text{Hyponym} : s_{5,4} - c_5$  are considered knowledge paths because the incoming and the outgoing senses for each concept are preserved. Figure 2.1 (b) (c) (d), and (e) are examples of the different types of paths.



**Figure 2.1. (a) Example of an Illuminated Knowledge Graph, (b), (c), (d) and (e) are examples of Geometric path. (b), (c), and (e) are examples of Knowledge Paths.**

The process uses two fundamental techniques to generate the Illuminated Knowledge Graph *IKG*: Concept representation generation and Reference consultation.



## 2.1 Concept representation generation

When attempting to understand any text, readers always break the text down into concepts and create knowledge associations among them (Wittrock, 1989, 1992, Kintsch, 1988, 1994). Therefore, it is necessary to use a computational representation model on which a computer algorithm can handle the problem of finding new or missing knowledge associations among the prose concepts. This in turn increases comprehension of the relationships among the prose concepts, thus comprehending the prose. A graph is used as a computational representation model, where each node represents a concept; each edge represents a sentential relation (knowledge association/sentence) between two concepts; and each path is a sentence or sequence of sentences between two concepts.

We use *Syntactical Explicit Graph generator function*  $KG$  to convert the prose  $LTX_i$  to a knowledge graph  $G_0$  and a reference text  $RTX_i$  to a reference knowledge graph  $G_{Ri}$ . Formally, given a prose  $LTX$  or a reference text  $RTX_i$ , for each sentence *in*  $LTX/RTX_i$ , the function searches sentence-by-sentence for any pair of concepts  $(c_i, c_j)$  if there is a word or sequence of words  $t_b, b=1,2,\dots,n$  between them in the same sentence, where in  $LTX$ ,  $c_i$  and  $c_j$  belong to the list of prose concepts  $C_L$  and in  $RTX_i$ ,  $c_i$  and  $c_j$  belong to the noun concepts in the reference text, the distance between  $c_i$  and  $c_j$  is less than or equal  $L$ . If so, it saves the triple  $[c_i, t_b, c_j]$  as an edge in the graph represents a syntactical explicit relation between a pair of concepts  $c_i$  and  $c_j$ .

As the purpose of this process is to find the highest familiarity knowledge associations connecting the prose concepts for enhancing the prose comprehension, it is necessary to evaluate the familiarity of these knowledge associations. This can be done by

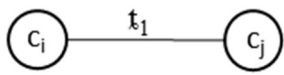
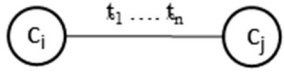
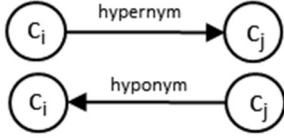
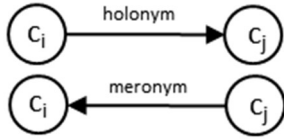
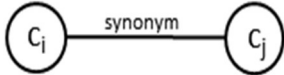
calculating the weight of each edge in the knowledge graph. Calculating the weight (familiarity value) is based on the type of the sentential relation between concepts. Table 2.1 represents different types of sentential relation structures between any pair of concepts. The familiarity value  $w_{i,j}$  is calculated by Equation 2.1 where,  $f_{i,j}$  represents the frequency of the relation type between concept  $c_i$  and  $c_j$  comes from the word frequency in the “Gutenberg Project” (Hart, 1971). This online archive of documents has been used by many researchers (Kaster, Siersdorfer, & Weikum, 2005; Kaluarachchi, Roychoudhury, Varde, & Weikum, 2011; Agrawal & An, 2012; Chandran, Crockett, Mclean, & Bandar, 2013) where "Gutenberg Project" is one of the online resources offers over 53,000 whose copyright has expired in the USA. If there is a sequence of words between the concepts, the familiarity value is based on the lowest weight of the words sequence. Since, word frequencies are in the millions, we find the *log* of a word frequency divided by  $10^9$ . To avoid negative values, we multiply the result by -1. High frequency means high familiarity of the relation type. The relation between  $f$  and  $w$  is an inverse relation; the higher the frequency, the less its weight or the less its cost.

$$w_{i,j} = -1 / \left( \frac{1}{\log\left(\frac{f_{i,j}}{10^9}\right)} \right) \quad (2.1)$$

In Table 2.1, *Case #1* there is a single word  $t_b$  between  $c_i$  and  $c_j$ ,  $b=1$ . (i.e. ***Ethane contains carbon-carbon***),  $c_i = \textit{Ethane}$ ,  $c_j = \textit{carbon-carbon}$ , and  $t_1 = \textit{contains}$ . Therefore, the edge weight is calculated by substituting the frequency of the word *contain* in Equation 2.1. For *Case#2*, more than one word  $t_b$  exists between  $c_i$  and  $c_j$ ,  $b>1$ , (i.e. ***petroleum is heterogeneous composed of hydrocarbon***),  $c_i = \textit{petroleum}$ ,  $c_j = \textit{hydrocarbon}$ ,  $b=2$ , and  $t_1 = \textit{heterogeneous}$ ,  $t_2 = \textit{composed}$ . The edge weight is calculated by calculating the weight of

word *heterogeneous* and *composed* separately, then selecting the minimum of them. In *Case#3: class/sub-class*, if the relation type is either a hypernym or hyponym (i.e. *Fossil fuel* is a hypernym of *petroleum*),  $c_i = \text{Fossil fuel}$ ,  $c_j = \text{petroleum}$  and  $t = \text{hypernym}$ , the edge weight is calculated by substituting the frequency of the word *class* in Equation 2.1. *Case#4: part/sub-part*, if the relation type is either a holonym or meronym (*Atom* is a holonym of *carbon*),  $c_i = \text{Atom}$ ,  $c_j = \text{carbon}$  and  $t = \text{holonym}$ , the edge weight is calculated by substituting the frequency of the word *part* in Equation 2.1.

**Table 2.1. Relation Structure between any pair of Concepts**

Sentential relation type	Sentential relation structure	$w_{i,j}$ value
Syntactical Relation		
	Case#1: <i>single word</i> : $c_i - : s_{i,*} - t_b - s_{j,*} : - c_j ; b=1$	$w_{i,j} = t_1$
	Case#2: <i>sequence of words</i> : $c_i - : s_{i,*} - t_1 t_2 \dots t_n - s_{j,*} : - c_j$ $b=1,2,\dots,n$	$w_{i,j} = \min(t_b)$
OE Relation		
	Case#3: <i>Class/sub-class</i> : $c_i - : s_{i,*} - \text{Hypernym} - s_{j,*} : - c_j$ or $c_i - : s_{i,*} - \text{Hyponym} - s_{j,*} : - c_j$	$w_{i,j} = t_{\text{class}}$
	Case#4: <i>Part/sub-part</i> : $c_i - : s_{i,*} - \text{Holonym} - s_{j,*} : - c_j$ or $c_i - : s_{i,*} - \text{Meronym} - s_{j,*} : - c_j$	$w_{i,j} = t_{\text{part}}$
	Case#5: <i>synonym</i> : $c_i - : s_{i,*} - \text{Synonym} - s_{j,*} : - c_j$	$w_{i,j} = t_{\text{synonym}} = 1$

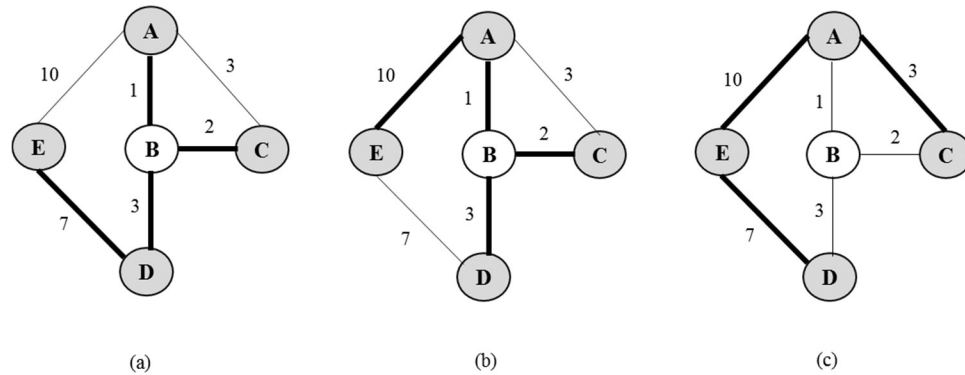
*Case#5: synonym*, if the relation type is synonym (*Petroleum is a synonym of oil*), we suppose the frequency of the synonym relation equals  $1$ , therefore the edge weight is calculated by substituting the frequency by  $1$  in the equation.

## 2.2 Reference Consultation

### 2.2.1 Extracting the highest familiarity knowledge from a reference text

The reader is not expected to find knowledge associations among prose concepts and comprehend the prose based on the knowledge found in the prose alone. To fill the gap left by prose, external text references are used to find knowledge associations among the prose concepts which are not inferred from the sentences presented in the prose. For that reason, a *Terminal to Terminal Steiner Tree function (TTST)* can find missing knowledge associations as well as new knowledge associations among the prose concepts. This process occurs by reading the relevant parts from external reference texts whose subjects are related to the given prose, and discovering the highest familiarity meaningful knowledge associations that connect the prose concepts. This in turn increases knowledge and enhances prose comprehension.

It is therefore relevant to discuss some techniques on graph theory which can be used for finding the knowledge associations among the prose concepts  $C_L$  before presenting the suggested technique. Many versions of Steiner Trees exist that are useful for finding knowledge associations among prose concepts  $C_L$  (Takahashi & Matsuyama, 1980; Kou, Markowsky, & Berman, 1981). Below are some examples which may serve the Knowledge Induction Process.



**Figure 2.2. Example of different types of Steiner trees.**

A Minimum Steiner Tree (MST), a heuristic approach to obtain an approximate solution for MST is based on finding knowledge with the minimum cost among the prose concepts  $C_L$  (Takahashi & Mastsuyama, 1980). Given a connected, undirected graph  $G = (C, E)$ , where  $C$  is a set of concepts,  $E$  is a set of edges representing the relations between pairs of concepts  $c_i$  and  $c_j$ , for each edge  $e \in E$ , the weight  $w_{ij}$  specifying the cost (the familiarity value of the sentential relation between  $c_i$  and  $c_j$ ) and a set of prose concepts  $C_L$ ,  $C_L \subset C$ . MST's cost is calculated by  $\sum(w_{ij})$ , where  $i, j \in C$ ,  $i \neq j$ . MST may contain some concepts that do not belong to the prose concepts  $C_L$  but are used to connect them. The algorithm complexity is  $O(C^2)$ . However, one of the debatable points of MST is the addition of external concepts (not belonging to  $C_L$ ). This contradict the purpose for connecting  $C_L$ , as we are seeking to connect  $C_L$  with the minimum number of external concepts. As increasing the number of external concepts leads to increasing the load of learning their relations. An example of this can be found on the graph shown in Figure 2.2,  $C_L = \{A, C, E, D\}$ . The MST's cost for the graph is 13 as shown in Figure 2.2. (a). It can be seen that there is an external concept  $B$ .

Suppose that there is a traffic  $Tr$  between  $c_i$  and  $c_j$ , where the traffic refers to the amount of comprehension between the two concepts. The purpose of Steiner tree now is to decrease the traffic among  $C_L$ . In this case, the Steiner tree is called a Traffic Steiner Tree (TST). Its cost is calculated by  $\sum (Tr_{ij} \cdot w_{ij})$ , where  $i, j \in C, i \neq j$ . The algorithm complexity is  $O(C^2)$ . Now suppose that the traffic  $Tr_{ij}=1$  and  $Tr_{EA}=100$ . Thus, the TST for the graph is 1006 as shown in Figure 2.2. (b). Again, we see the external concept  $B$  appears in TST.

None of the mentioned Steiner Tree versions could work well here, as a more fitting tree for connecting  $C_L$  should have no or the minimum number of external concepts. A version of Steiner Tree called *Terminal to Terminal Steiner Tree (TTST)* finds knowledge with the minimum associations among the prose concepts  $C_L$  using no or the minimum number of external concepts. An example of this type of Steiner tree can be seen in Figure 2.2 (c).

The algorithm in Figure 2.3 represents *TTST* as follows: The input of the algorithm is a reference knowledge graph  $G_{Ri}$  and  $C_L$ , where the output is  $G_{Ui}$ , which is a tree(s) from  $G_{Ri}$  that presents the highest familiarity knowledge path(s) among  $C_L$ . The search for *TTST* has been implemented as a Breadth-First-Search (BFS). For each component  $comp$  in  $G_{Ri}$ , the algorithm uses a queue data structure *Queue* to temporarily hold each visited concept in  $G_{Ri}$  with its neighbors. It picks any concept from  $C_L$  as the source  $s$  for initializing the *Queue*. Then, it initializes the *cost* between  $s$  and each concept  $c$  in the  $comp$  to *INFINITY* and initializes the previous concept *prev* of each  $c$  to  $-1$ . In the loop iteration, it dequeues the first concept  $c$  in the queue, marks it as visited, and checks if  $c \in C_L$ . If so, it updates its *cost* to  $0$ , adds it to  $M$  where  $M$  holds the found  $C_L$  concepts and removes it from  $C_L$ .

Then, it enqueues all the neighbors  $c_i$ 's of concept  $c$  if they are marked as non-visited, assigns  $prev$  and calculates  $cost$  for each of them. If the current  $cost$  of  $c_i$  is less than its previous  $cost$ , that means a less costly knowledge path to  $c_i$  is found, where less cost means high familiarity. The  $c_i$ 's  $prev$  and  $cost$  are updated to the new lesser values and the process is repeated till the queue is emptied. If all  $comp$  are checked,  $getPaths$  constructs the  $TTST$  from  $M$  and  $prev$ . The returned  $TTST$  are represented in  $G_{U_i}$ .

---

**Def Terminal to Terminal Steiner Tree ( ):**


---

**Input:**  $G_{R_i}, C_L$

**Output:**  $TTST$

```

1. //initialization
2. for each comp in  $G_{R_i}$ :
3.   if  $C_L \neq \phi$  :
4.     for each concept  $c$  in comp
5.        $prev[c] = -1$ 
6.        $cost[c] = INFINITY$ 
7.        $Visited[c] = False$ 
8.    $Queue = \phi$ 
9.    $s =$  pick any member from  $C_L$ 
10.  enqueue( $Queue, s$ )
11.  While  $Queue \neq \phi$ :
12.     $c =$  dequeue( $Queue$ )
13.     $Visited[c] = True$ 
14.    if  $c$  in  $C_L$ :
15.       $cost[c] = 0$ 
16.      add  $c$  to  $M$ 
17.      remove  $c$  from  $C_L$ 
18.    for each neighbor  $c_i$  of  $c$ :
19.      if  $c_i$  not in  $Queue$  and  $Visited[c_i] == False$ :
20.        enqueue( $Queue, c_i$ )
21.         $alt = cost[c] + a_{c_i, c}$ 
22.        if  $alt < cost[c_i]$ 
23.           $prev[c_i] = c$ 
24.          // a less cost knowledge path to  $c_i$  has been found
25.           $cost[c_i] = alt$ 
26.  $TTST = getPaths(M[ ], prev[ ])$ 
27. return  $TTST$ 

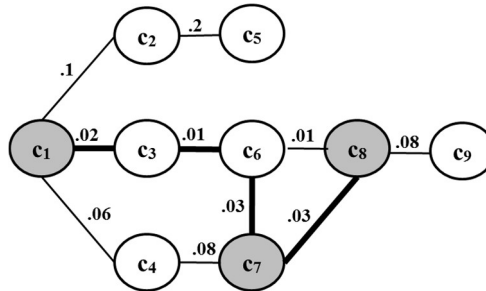
```

---

**Figure 2.3. Terminal to Terminal Steiner Tree algorithm.**

To extract the *TTST*, the algorithm scans  $G_{Ri}$  until all of the concepts in  $C_L$  are visited. In a worst case scenario, the entire  $G_{Ri}$  must be scanned. This means that all of the concepts will then be pushed in the queue and each outgoing edge is scanned once. Thus, the time complexity of the algorithm is  $O(C+E)$ .

In the initialization *line 2 to line 7*, each concept is scanned exactly once which takes  $O(1)$  time, so the total devoted for initializing the whole concept is  $O(C)$ . The outer loop in *line 11* dequeues each concept just once which takes  $O(1)$  for each concept and totals  $O(C)$  for dequeuing the whole concepts. Meanwhile, the inner-loop in *line 18* enqueues each edge from each concept just once which takes  $O(E)$  for scanning the whole edges. Thus, the total running time of *TTST* is  $O(C + E)$ . Consider the reference knowledge graph  $G_{Ri}$  shown in Figure 2.4 and  $C_L = \{c_1, c_7, c_8\}$ . The *TTST* returned by algorithm is  $\{[c_1, c_3, c_6, c_7, c_8]\}$ .



**Figure 2.4. Example of a Terminal to Terminal Steiner Tree.**

The new knowledge associations found using the external reference text need to be assimilated by connecting them to the current comprehend one. So, the knowledge graph is merged with  $G_{Ui}$  and generates  $G_{temp}$  which represents the current state of comprehension after reading  $RTX_i$ .



### 2.2.2 Extracting knowledge from an Ontology Engine

Using an ontology engine is considered a useful source for finding knowledge associations between two concepts. The *OE-Knowledge-Paths function*  $KP_{OE}$  is utilized to provide knowledge about the ontology engine associations between each pair of concepts. This helps in the addition of new types of knowledge that contribute to increasing knowledge and enhancing prose comprehension.

Given a pair of concepts  $s$  and  $t$ , the OE-Knowledge-Path is a sequence of edges that connects a concept  $s$  with a concept  $t$  in a preserved sense, where both  $s$  and  $t \in G_{temp}$ . Each edge in the sequence represents the relation between its ends, where the relation is one of  $R$  relation types which are (*synonym, hyponym, hypernym, meronym, and holonym*). The algorithm in Figure 2.5 searches for OE-Knowledge-Paths of length less than or equal to  $\alpha$  connecting each pair of concepts  $s$  and  $t$ , where  $s$  and  $t$  are respectively the first and last concepts in the path if found using an Ontology Engine  $OE$ . Its input is  $s, t$  where  $s, t \in G_{temp}$ ,  $R$  is a dictionary of all relations between concepts in the Ontology Engine  $OE$ , and *relationalGraph* is a dictionary used to hold concepts that have any type of relations from  $R$  with the last node of the current path. The search for an OE-Knowledge-Path has been implemented as a Breadth-First-Search (BFS). The algorithm searches for and within all senses of  $s$ . For each sense, it searches for OE-Knowledge-Paths from  $s$  concept to  $t$  concept by searching the neighbors of  $s$  that have any type of relations from  $R$  and have the same sense of  $t$ . Then, it searches the neighbors of the neighbors, and so on until it reaches  $t$ . The algorithm does not return the shortest path between the pair of concepts as the knowledge path, because it could be a path with concepts that require the use of multiple senses.

---

**Def discover knowledge-paths ( ):**


---

**Input:** s, t, R,  $\alpha$ 
**Output:** OE-Knowledge-Paths between s and t

```

1. PathQ=[ ]
2. Kapths=[ ]
3. // push the first path into PathQ
4. PathQ.append([s])
5. for sen in s.sense( ):
6.     while PathQ:
7.         // get the first path from the PathQ
8.         NodeQ = PathQ.pop(0)
9.         // get the last node from NodeQ
10.        node= NodeQ[-1]
11.        // path found
12.        if node == t:
13.            Kapths.append(NodeQ)
14.            return Kpaths
15.        else:
16.            If len(NodeQ) <=  $\alpha$ :
17.                sl= list( )
18.                for key, value in R.iteritems( ):
19.                    re= value
20.                    // get all concepts have relations from R with node and have the same
sense of node
21.                    x= re (node, sen, key)
22.                    sl=sl+x
23.                    relationalGraph[node] = sl
24.                    // enumerate all adjacent nodes, construct a new path and push into the
queue
25.                    for adjacent in relationalGraph.get(node,[ ]):
26.                        new_path=list(NodeQ)
27.                        new_path.append (adjacent)
28.                        if len(new_path) <  $\alpha$ :
29.                            PathQ.append(adjacent)
30.                        else:
31.                            break
32.            if !(Kpaths):
33.                return 'Not found'

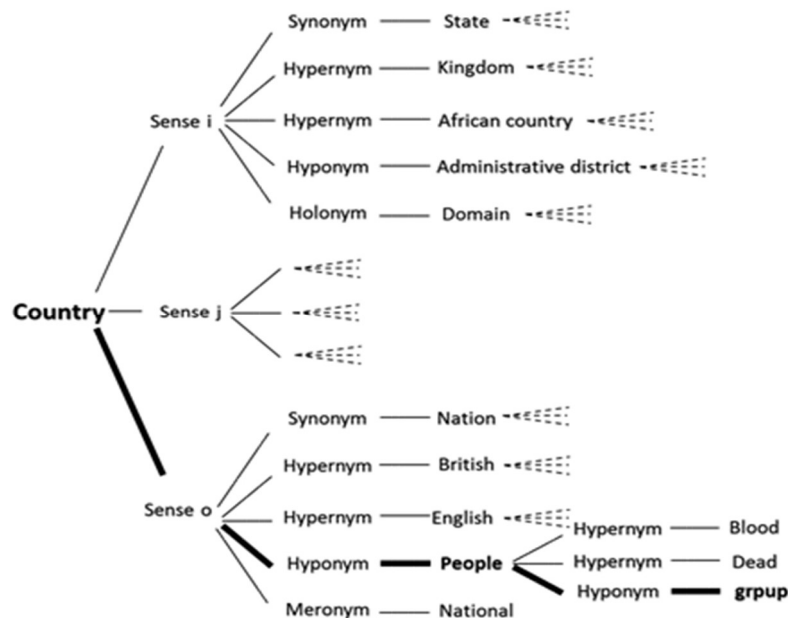
```

---

**Figure 2.5. Discovering OE- Knowledge-Paths algorithm.**

The algorithm uses two queue data structures: *NodeQ* and *PathQ*. The *NodeQ* saves the current path that has the concept to explore the next one. The *PathQ* holds the created

paths until now. The algorithm starts with  $s$  as the current path, while a loop iterates through the paths in  $PathQ$  searching for an OE-Knowledge-Path connecting  $s$  and  $t$ . In each loop iteration, it dequeues the first path in  $PathQ$  and signs it in  $NodeQ$ . It then checks if the last concept in  $NodeQ$  matches  $t$ . If so, it saves the OE-Knowledge-Path in  $Kpaths$ . If not, it checks if the length of the  $NodeQ$  is less than  $\alpha$ . If it is, for the sense of the last concept in  $NodeQ$ , the function gets all of the concepts that have one of the relation types from  $R$ , with the last concept in  $NodeQ$  and adds them to  $relationalGraph$ . A number of paths are created between each concept in the  $relationalGraph$  and the current path. The new created paths are then saved in  $PathQ$ . If all paths in  $PathQ$  are checked and  $Kpaths$  does not exist, the function returns ‘Not found’. The algorithm is performed between each pair of concepts  $s$  and  $t$ ,  $s, t \in G_{temp}$ . The returned paths hold in  $Kpaths$  between each pair of concepts are added to  $G_{wi}$ . Let us consider  $s = \text{“country”}$  and  $t = \text{“group”}$ ,  $\alpha = 4$ , and the following path is the OE-Knowledge-Path returned by the algorithm **country** (hyponym) **people** (hyponym) **group**. Figure 2.6 illustrates the process of discovering the OE-Knowledge-Path between the two concepts “country” and “group”.



**Figure 2.6. Example of an OE-Knowledge-Path.**

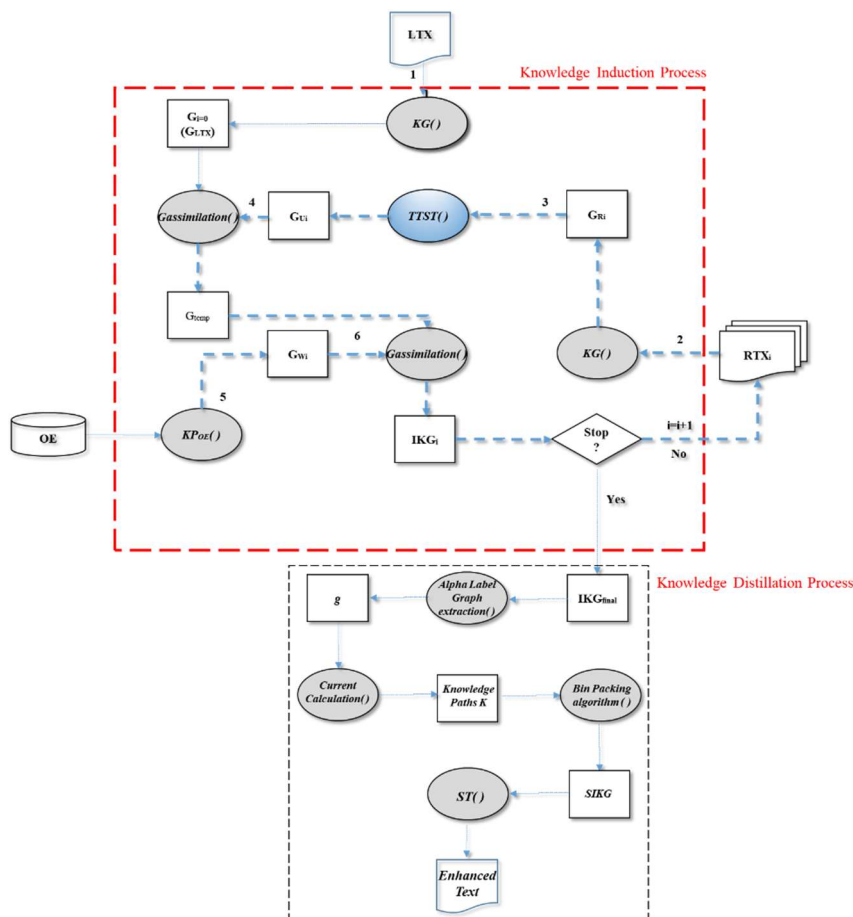
The new knowledge associations found using the ontology engine are need to be assimilated by connecting them to the current comprehended knowledge. So,  $G_{temp}$  is merged with  $G_{Wi}$  to generate  $IKG_i$  which represents the current state of comprehension after reading  $RTX_i$  and the ontology engine.

The Knowledge Induction Process performs the following steps to generate the Illuminated Knowledge Graph  $IKG$ . The sequence of the steps is presented in Figure 2.7.

1. The process uses the *Syntactical Explicit Graph generator function*  $KG$  to convert the given prose  $LTX$  to a prose knowledge graph  $GLTX$  ( $G_{i=0}$ ) representing the syntactical association between each pair of  $C_L$  concepts in  $LTX$ .
2. The process uses the same function *Syntactical Explicit Graph generator function*  $KG$  to convert the reference text  $RTX_i$  to a reference knowledge

graph  $G_{Ri}$  representing the syntactical association between each pair of noun concepts in  $RTX_i$ .

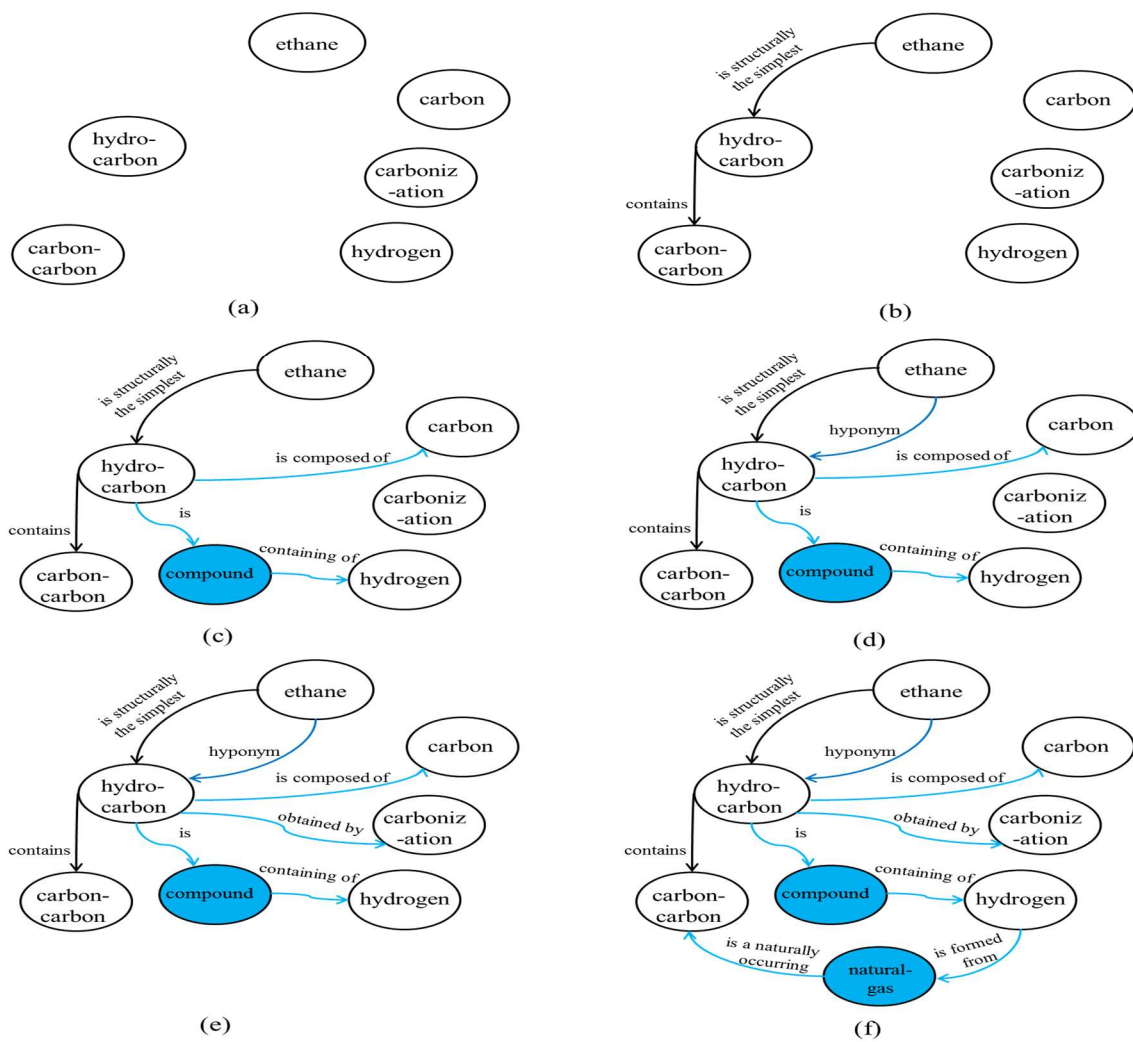
3. From  $RTX_i$ , the process uses *Terminal to Terminal Steiner Tree function* ( $TTST$ ) to extract the highest familiarity meaningful knowledge paths that connect the prose concepts  $C_L$ . The extracted path(s) is called Terminal to Terminal Steiner tree(s)  $TTST$  and it is represented in a graph  $G_{Ui}$ .
4. The process merges  $G_0$  and  $G_{Ui}$  graphs in a  $G_{temp}$  graph that represents the current assimilated knowledge among the learnable prose concepts  $C_L$ .
5. For each pair of concepts in  $G_{temp}$ , the model uses the *OE-Knowledge-Paths function*  $KP_{OE}$  to find an ontology engine path(s) connecting each pair of concepts. The found paths are represented in a graph  $G_{Wi}$ .
6. The process merges  $G_{temp}$  and  $G_{Wi}$  in the Illuminated Knowledge Graph  $IKG_i$  that represents the current assimilated knowledge among the prose concepts  $C_L$ .



**Figure 2.7. Comprehension Engine.**

The process performs steps 2-6 each time it reads a new reference text  $RTX_i$ , where  $G_0$  is replaced by  $IKG_i$  in step 4.

An example showing the impact of using reference texts and an ontology engine for adding knowledge associations among the prose concepts is a prose  $LTX$  about the chemical compound ‘Ethane’. Here,  $C_L$  has six learnable concepts *ethane*, *hydrocarbon*, *hydrogen*, *carbon*, *carbon-carbon* and *carbonization* and we need to find knowledge associations connect all the concepts in  $C_L$  using external consultations. Figure 2.8 explains the process of finding knowledge associations connecting  $C_L$  concepts using reference texts and an ontology engine.



**Figure 2.8.** The process of connecting the prose concepts  $C_L$  using reference texts and ontology engine. (a) a set of prose concepts  $C_L$ . (b) Knowledge path  $K$  from the prose LTX. (c) Knowledge path  $K$  using RTX1. (d) Knowledge path  $K$  using Ontology Engine OE. (e) Knowledge path  $K$  using RTX2. (f) Knowledge path  $K$  using RTX3.

### 2.3 Summary

The Chapter presented the details about the Knowledge Induction Process. The process is designed to substitute the lack of knowledge in a prose text by augmenting the highest familiarity knowledge associations among the prose concepts by using incremental external knowledge sources. It began by presenting the main definitions used by the

process. Next, it explained the two fundamental techniques used by the process. The concept representation generation technique explains the computational representation model that the process used to represent knowledge. The reference consultation technique utilizes a set of algorithms to augment knowledge from reference texts and an ontology engine. Lastly, it concluded the steps used by both techniques to achieve the Knowledge Induction Process task.



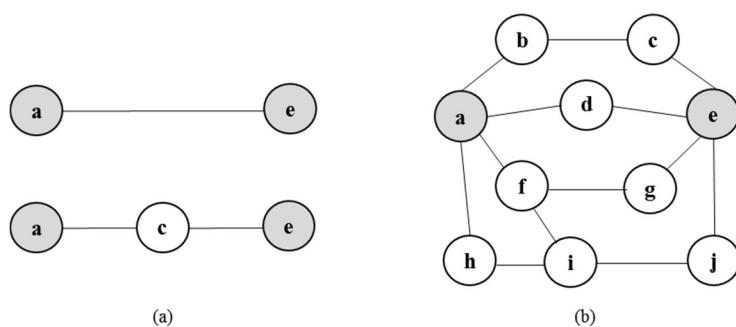
## CHAPTER 3

### **Knowledge Distillation Process of Prose Comprehension**

This chapter describes and defines in detail the steps in Knowledge Distillation Process.

The incremental use of external knowledge sources (reference texts and an ontology engine) for finding knowledge associations among prose concepts to enhance the prose comprehension results in adding external concepts to the prose concepts and knowledge associations among the whole set of concepts. This could result in ‘*too much knowledge*’. However, It is generally believed that human brains are not efficiently designed to assimilate ‘*too much knowledge*’ and only limited amounts of such knowledge can be acquired and retained (Johnson, 1980; Moseley, 2005; Patzer, 2012, 2006). For example, *cognitive economics* explain the fact that “People are flooded by information which must somehow be reduced and simplified to allow efficient processing and to avoid and otherwise overwhelming overload” (Mischel, 1979). Besides, ‘*too much knowledge*’ could involve familiar/unfamiliar, easy/complex, and/or repeated information which is likely to be misleading, and causes mess in comprehending the prose (Draper, Brown, Henderson, & McAteer, 1996). Concerning the grip of unmediated knowledge, Ball suggests “we have too much knowledge and not enough understanding” (Ball, 1998, p. 78). Thus, the Knowledge Distillation Process is proposed to grade the augmented knowledge, select the most familiar, easiest-to-understand knowledge associations that can be useful for

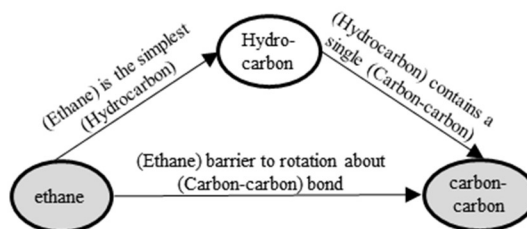
comprehending the relation between each pair of the prose concepts and present them to the reader as an enhanced text. Suppose a reader is given a prose and is asked to comprehend it. We assume the concepts are known and we focus on understanding the knowledge associations between each pair of concepts in the prose. In the simplest case, a sentence or a sequence set of sentences explains the knowledge associations between a pair of concepts. However, it becomes more difficult when there are multiple sets of sentences representing knowledge associations between the pair of concepts. Although the latter case could help in increasing the understanding and enhancing the comprehension of the association between the two concepts, it could confuse the reader rather than make the meaning clear. Consider the knowledge graph shown in Figure 3.1. Here, the simple case is shown in (a) where there is only one knowledge path representing the relation between concept  $a$  and concept  $e$ , while in (b), multiple knowledge paths exist between the pair:  $abce$ ,  $ade$ ,  $afge$ ,  $afije$ , and  $ahije$ . The question now is which knowledge path will serve as the best aid when comprehending the relation between  $a$  and  $e$ ?



**Figure 3.1. Examples of different types of paths between two concepts.**

It is possible that multiple Knowledge Paths  $K$  could exist between a pair of concepts containing both familiar and unfamiliar knowledge. While familiar knowledge is

represented in sentences consist of familiar words that are easy to understand, unfamiliar knowledge represented in sentences consist of unfamiliar words that can become obstacles to comprehension. To better illustrate this idea, the knowledge graph in Figure 3.2 shows two knowledge paths connecting *ethane* and *carbon-carbon* concepts; *[Ethane] barrier to rotation about the [carbon-carbon] bond* and *[Ethane] is the simplest [hydrocarbon] contains a single [carbon-carbon]*, which of them is considered the better option for comprehending the relationship between *ethane* and *carbon-carbon*. As is evident, the former knowledge path involves unfamiliar words that make the relation between the *ethane* and *carbon-carbon* hard to understand, while the latter knowledge path involves easy and familiar words that make the relation between the two concepts easier to understand. For that reason, finding a familiar knowledge path can help to better understand the relation between a pair of concepts among set of knowledge paths to enhance the prose comprehension.



**Figure 3.2. An Example of multiple knowledge paths between ethane and carbon-carbon.**

The simpler knowledge path is an example of what is known as an “*Alpha Knowledge Pathway*”  $K'$ . It is defined as the best knowledge subgraph that represents the relation between a single source concept  $c_i$  and a single destination concept  $c_j$  where each

word in the path is familiar and easy to understand. This pathway helps to best understand the relation between  $c_i$  and  $c_j$ , where  $c_i$  and  $c_j \in$  the list of the prose concepts  $C_L$ . In some cases, the “*Alpha Knowledge Pathway*”  $K'$  could be a single knowledge path comprehending the relation between two concepts.

All of the “*Alpha Knowledge Pathway*”  $Ks'$  between each pair of concepts in the list of prose concepts  $C_L$  are combined to form a *Skimmed Illuminated Knowledge Graph SIKG*, which is defined as an abstract graph from the *Illuminated Knowledge Graph IKG* that only contains the *Alpha Knowledge Pathway*  $K'$  between each pair of concepts in prose  $LTX$  and belongs to  $C_L$ .

Formally the second part of the problem is presented as the following:

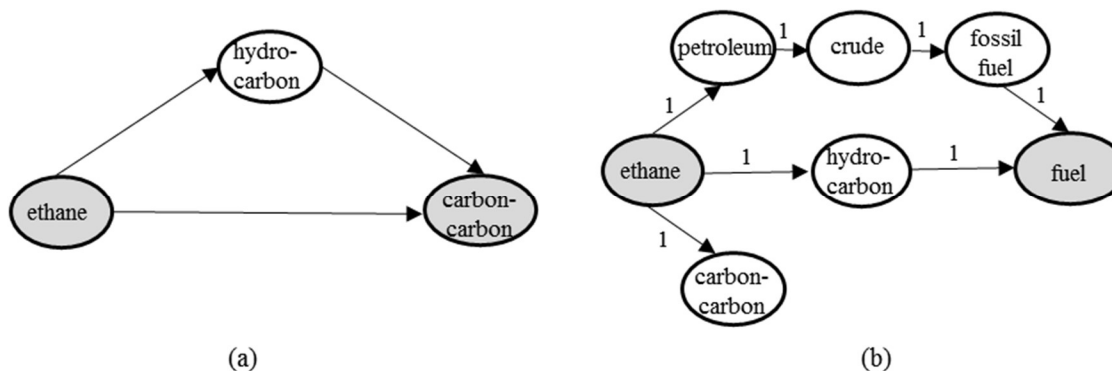
**Given:** a weighted directed Illuminated Knowledge Graph  $IKG$ , concept  $s$  and concept  $t$ ,  $s$  and  $t \in C_L$ . **Find:** (1) a connected *Alpha Label graph*  $g$  representing all the relations between  $s$  and  $t$  that maximize a “goodness” function. The connected *Alpha Label graph*  $g$  problem has two sub-problems: What is the appropriate “goodness” function that corresponds to good comprehension between the two concepts? and How to compute it? (2) an “*Alpha Knowledge Pathway*”  $K'$  from  $g$  that achieves the highest “goodness” (i.e. best knowledge subgraph to help in comprehending the relation between  $s$  and  $t$ ).

### 3.1 Goodness Paths(s) notion

It would be relevant to discuss the notion of the “goodness” path(s) and its relationship to the comprehension. The notion of the “goodness” path(s) extracted from large graphs has been frequently addressed (Rodrigues Jr et al., 2013; Ramakrishnan, Milnor, Perry, & Sheth, 2005; Faloutsos, McCurley, & Tomkins, 2004; Tong & Faloutsos,

2006; Koren, North, & Volinsky, 2006; Kasneci, Elbassuoni, & Weikum, 2009; Fang, Sarma, Yu, & Bohannon, 2011). Interestingly, the answer is non-trivial. One would think that the most obvious measures for choosing a good path between two concepts are the shortest-hubs path measurements. However, this is not always true. An example of this can be seen in the knowledge graph in Figure 3.3 (a). Here, it shows two knowledge paths between *ethane* and *carbon-carbon*. *[Ethane] barrier to rotation about the [carbon-carbon] bond* has length 1 and *[Ethane] is the simplest [hydrocarbon] contains a single [carbon-carbon]* has length 2. However, the former knowledge path is shorter, but its words do not look familiar. The latter knowledge path is longer but easier to understand. So, in this case, the “*Alpha Knowledge Pathway*” could not be captured by the traditional shortest-hubs path.

In attempt to refine this paradox one may suggest that we need to identify a form that shows the closeness of the edge sets in the path rather than the length. A graph theoretic analogy will determine a path that would provide maximum flow in terms of some assigned closeness value assigned to the edges through the path. Normally this will mean making an estimation of the narrowest edges’ capacity that makes the path. This approach can solve the problem with the previous case. However, it has its own limitations. An example of the drawbacks can be seen in Figure 3.3 (b), where in the paths *Ethane-petroleum-crude-fossil fuel* and *ethane-hydrocarbon- fuel*, both knowledge paths carry maximum flow of 1 even though one is shorter than the other. So, it is apparent that the measurement is also inadequate on its own in finding the “*Alpha Knowledge Pathway*” *K*’ between two concepts.



**Figure 3.3. Two simple graphs where both (a) shortest-hubs path and (b) network flow fail for discovering “Alpha Knowledge Pathway” between two concepts”.**

It is possible to design a goodness function that can combine the essence of both measurements. For that, we resort to an innovative electrical circuit theoretic concept recently introduced by Faloutsos (Faloutsos et al., 2004). While studying social closeness in a social network graph, Faloutsos and his associates found (Faloutsos et al., 2004) that both measurements fail to capture their preferred characteristics for selecting a good path in the social network. They investigated a new measure based on calculating the maximum delivered current flow for finding a good path in a social network. They applied their technique on a graph where the nodes represent famous people and the edges among the nodes represent the strength of acquaintance among them, where the strengths values representing the edges weights were delivered from the co-occurrences of the people’s names in web pages.

### 3.2 Equivalent Electrical Circuit (EEC)

Several relevant techniques exist that can be used for finding the *Alpha Knowledge Pathway*  $K'$  comprehending the relation between two concepts. Each of them is affected

by some measurements that can affect finding the right *Alpha Knowledge Pathway K'*. This section will discuss each technique, showing how each can affect finding the right *Alpha Knowledge Pathway K'*.

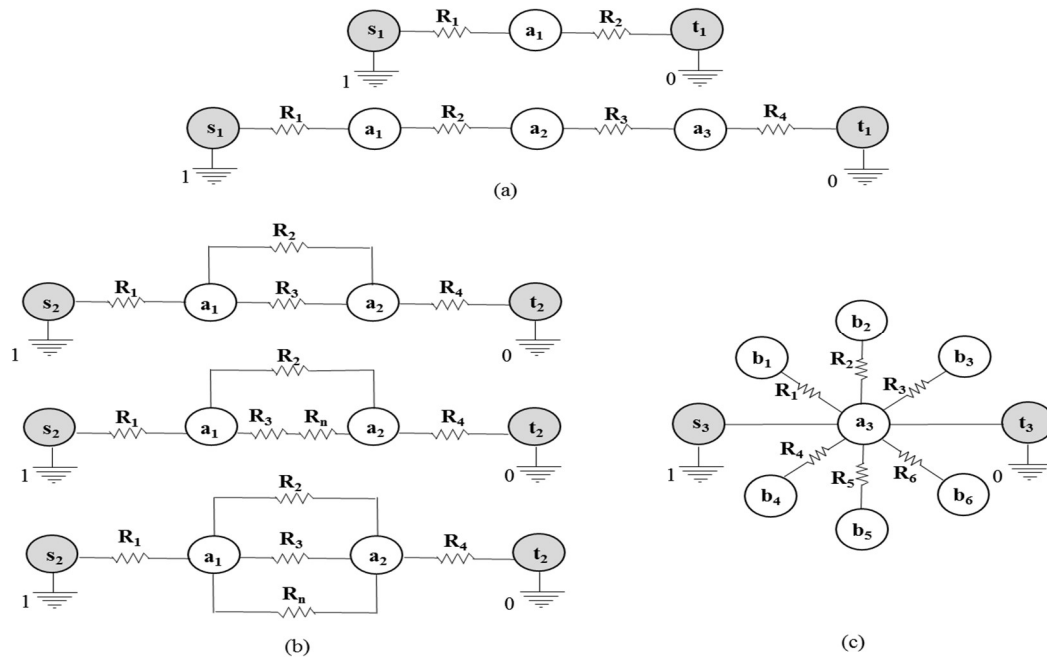
Effective Resistance (ER) models an *Alpha Label graph g* of all the knowledge paths between a source concept  $c_i$  and a destination concept  $c_j$  as an equivalent electric circuit (EEC). It treats the edges as resistors with resistance equal to the inverse of the edge weight, applying +1 voltage to  $c_i$  and 0 voltage to  $c_j$ , and solving a system of linear equations to estimate the voltages and the currents of the equivalent electric circuit. Faloutsos and his associates (Faloutsos et al., 2004) considered the same method<sup>1</sup> as a goodness function in their connecting paths study to calculate the maximum delivered current flow to find a good path between two query nodes.

One of the appealing properties of the EEC is that it distinguishes between long paths and short paths between the query nodes. An example is the electrical circuit in Figure 3.4 (a), connecting concept  $s_l$  with concept  $t_l$ . Here, it is shown that the length of the top path is shorter than the bottom path. A short path in EEC means it has fewer resistors carrying a lower resistance than a long path. For the purposes of this thesis, resistance refers to the inverse of the edge weight; low resistance means that the path representing the relation between  $s_l$  and  $t_l$  involves familiar and easy to understand words that help in comprehending the relation between  $s_l$  and  $t_l$ . The bottom path, however, has more

---

1. Faloutsos uses the term Effective Conductance EC.

resistors carrying higher resistance. Higher resistance refers to the high weight of the edges, meaning that the words of the path are unfamiliar and more difficult to understand. Thus, in Figure 3.4 (a), the top *EEC* ( $s_1, t_1$ ) is better than the bottom *EEC* ( $s_1, t_1$ ).



**Figure 3.4. Equivalent Electrical circuits EEC graphs.**

EEC also has a monotonicity property (Doyle & Snell, 1984) that states adding a new resistor between two concepts in a path. This can be done in series or in parallel. Consider the top equivalent electrical circuit graph shown in Figure 3.4 (b). Adding a new resistor in the series as shown in the middle graph of Figure 3.4 (b) means increasing the resistance, while adding a new resistor in parallel as shown in the bottom graph in Figure 3.4 (b) means decreasing the resistance. As was mentioned earlier, low resistance represents the relation between a pair of concepts that is easy to understand and helps in comprehending the relation between them. In other words, the existence of multiple edges



between two query nodes in a graph means that there are multiple sentences illuminating the relationship between the two concepts, thus enhancing comprehension of the relation between them. Referring to Figure 3.4 (b), this implies that the bottom *EEC* ( $s_2, t_2$ ) is better than the middle *EEC* ( $s_2, t_2$ ).

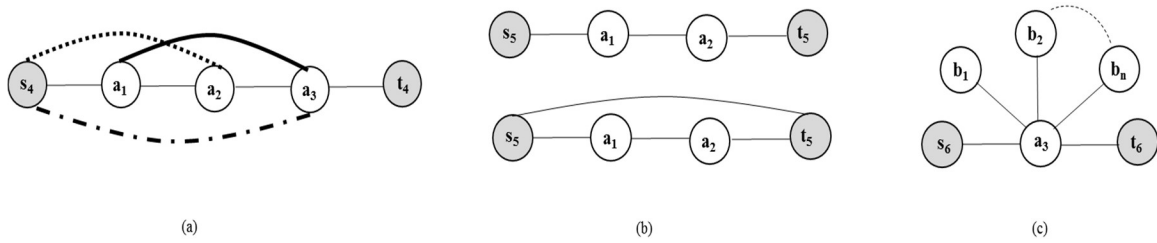
Another important property of using *EEC* appears in the case of a high degree concept that has edges to concepts of degree 1 (dead-end extension). As shown in Figure 3.4 (c), emanating from concept  $a_3$  to concept  $b_1, \dots, b_6$  have no impact on the equivalent electrical circuit graph representing the relation between concept  $s_3$  and concept  $t_3$ , and no impact in comprehending the relationship between them.

Another potential measure for discovering a good path between a pair of concepts which is *escape probability*, used in Random Walk (RW) (Koren et al., 2006). RW is a stochastic process on a graph. Given a graph, a source concept  $s$ , and a destination concept  $t$ ; RW selects a neighbor of concept  $s$  at random and goes there, then continuing the random walk from the newly chosen concept until it reaches  $t$ . In RW, the probability of transition from concept  $i$  to concept  $j$  is

$$P_{ij} = \frac{w_{ij}}{d_i}$$

where  $w$  is the weight of the edge between  $c_i$  and  $c_j$ , and  $d$  is the degree of  $c_i$ . Thus, for a given path  $P = v_1 - v_2 - \dots - v_r$ , the probability that a *RW* of  $P$  starts at  $v_1$  and reaches  $v_r$  can be expressed as:

$$\text{Prob}(P) = \prod_{i=1}^{r-1} \frac{w_{v_i, v_{i+1}}}{d_{v_i}}$$



**Figure 3.5. Pathological Cases in Random Walk Interpretation.**

The weight of the path  $P$  can be defined as:

$$\text{Wgt}(P) = d_{v_i} \cdot \text{Prob}(P)$$

$G = (V, E)$  is a weighted knowledge graph, where  $V$  is a set of vertices and  $E$  is a set of edges. In this context, the vertices denote to the concepts, each edge denoting to the sentential relation (sentence connecting two concepts) and each edge weight referring to the familiarity of the sentential relation between the two concepts.

Using  $RW$  for finding a good path between two concepts can be affected by the length of the path. The *escape probability* could be increased falsely if long paths must be followed, because when tracking a long path from a source concept  $s$  to a destination concept  $t$ , there is a chance in the random walk to backtrack and visit the same concepts many times as shown in Figure 3.5 (a). This can result in an overload of repeated information.

Moreover, the top path in figure 3.5 (b) adds a new edge in a path from concept  $s$  to concept  $t$  (*monotonicity*) in series. This contributes to an escape from  $s_5$  to  $t_5$ , falsely increasing the *escape property* (*similar to the long path case*) while adding a new edge from  $s_5$  to  $t_5$  in parallel as a path, this leads to increasing the escape property in a correct

manner as shown in the bottom path in Figure 3.5 (b), which results in finding new knowledge between  $s_5$  and  $t_5$ .

Furthermore, any walk through dead end extensions (*node of degree 1*) may not lead anywhere and will backtrack. This means that the random walk can make unlimited attempts to reach from concept  $s_6$  to concept  $t_6$  as shown in Figure 3.5 (c), also falsely increasing the *escape probability* to create an overload of no leading information.

For that, we are largely inspired by the ER measurement as it meets the desirable properties for measuring the “goodness function” in finding a connected path helps in comprehension the relation between two concepts. Where long or short paths, series or parallel paths, and/or dead-end extensions will not affect the flow of the current among the knowledge graph vertices thus the flow of the knowledge between the concepts in the knowledge graph.

### 3.3 Goodness Function Measurement

Before starting the Knowledge Distillation Process, a goodness function should be determined to select the *Alpha Knowledge Pathway  $K'$*  that best aids in comprehension of the relation between two concepts. Since most of the work for finding a good path between two concepts is based on the closeness of concepts (which has limitations) such as RW, the proposed goodness function is inspired by the ER measurement that most closely meets the properties needed for measuring the “goodness function”.

Here,  $G = (V, E)$  stands for an equivalent electrical circuit consisting of vertices  $V$  and a set of edges/resistors  $E$ . In this context, the vertices represent the concepts, each edge/resistor represents a sentential relation (sentence connecting two concepts) with

resistance equal to the inverse of the edge weight. The goodness function is defined by the highest delivered current flow carried between two vertices  $s$  and  $t$ , where the highest delivered current flow refers to the highest flow of knowledge familiarity. This represents the flow of comprehensibility between the two concepts.

Using Ohm's law and Kirchhoff's current law, a set of linear equations is created to estimate the voltages and the currents of the equivalent electrical circuit. Therefore, the delivered current flow for each path between a source concept and a destination concept is found. The path that has the highest delivered current flow is chosen to be the best path to represent the relation between the two given concepts.

### 3.4 Grading Process

Given an Illuminated Knowledge Graph  $IKG$  and a list of concepts  $C_L = \{c_i, \dots, c_n\}$  in prose  $LTX$ . The grading process performs the following steps for each pair of concepts  $c_i$  and  $c_j$  in  $C_L$ ,  $i \neq j$  to generate a Skimmed Illuminated Knowledge Graph  $SIKG$ ,  $SIKG \subset IKG$ , where  $SIKG$  joins all the "Alpha Knowledge Pathway"  $K'$  among each pair of concepts in  $C_L$ . The sequence of the Knowledge Distillation Process is shown in Figure 3.6.

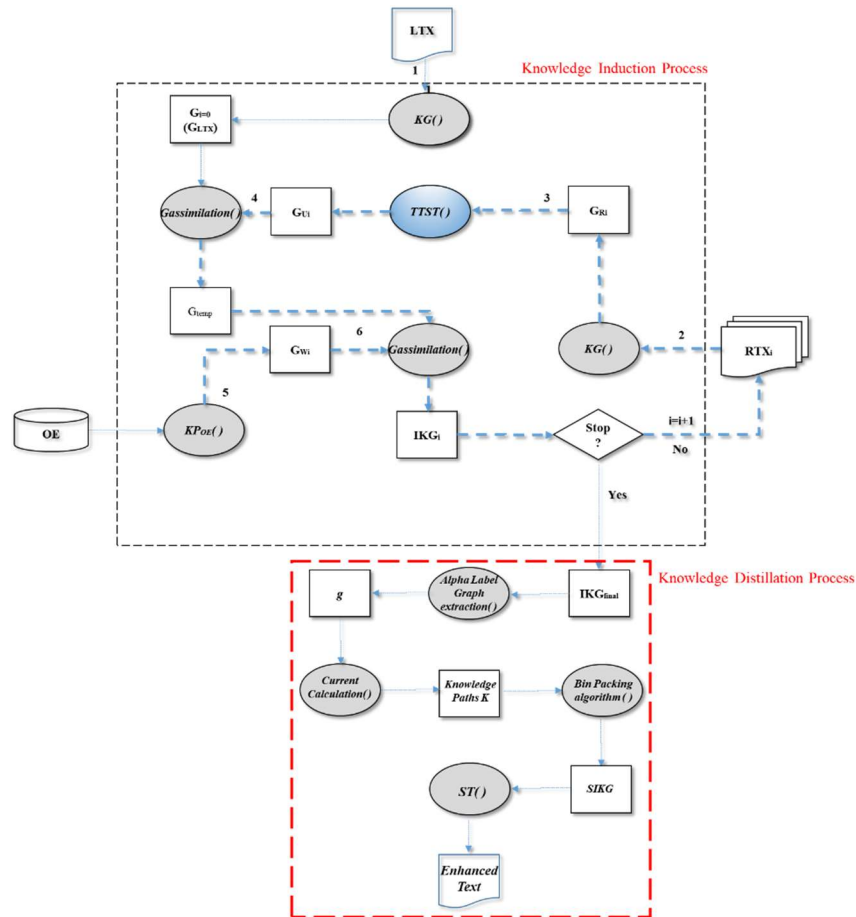


Figure 3.6. Comprehension Engine.

### 3.4.1 Alpha Label graph extraction

To find the *Alpha Knowledge Pathway K'* comprehending the relationships between two concepts, there should be an evaluation for each knowledge path connecting them. Thus, all the augmented knowledge paths between the two concepts is extracted in preparation for grading each of them. All of the extracted knowledge paths comprehending the relationships between the two concepts are represented in an *Alpha Label graph g*.

To extract the *Alpha Label graph g* for each pair of concepts in the list of prose concepts  $C_L$ , the Algorithm given in Figure 3.7 is performed. Here, the input of the

algorithm is the Illuminated Knowledge Graph *IKG* and a pair of concepts in  $C_L$ , where the output is a path or a set of paths representing all the knowledge paths  $K$  between the pair of concepts. The returned knowledge paths  $K$  are represented in the *Alpha Label graph*  $g$ .

The algorithm starts by initializing the path from *start* to *end* concepts with *start* as the first concept in the path. The *if* statement checks if the concept does not have an outgoing edge, and if so it returns nothing. The loop in line 7 searches for a connected concept with *start*; if the connected concept is not in the path, the algorithm calls itself with an update in the arguments where *start* will be updated to the new traversed concept. The path in line 1 is also updated to the new created path and so on until the *end* concept is reached. If a path between *start* and *end* concepts is discovered, the discovered *newpath* is appended to *paths* and so on. All the returned paths are represented in an *Alpha Label graph*  $g$ .

---

**Def find\_all\_paths():**

---

**Input:** IKG, start, end

**Output:** Alpha Label graph  $g$

```

1.     path = path + [start]
2.     if start == end:
3.         return [path]
4.     if not graph.has_key(start):
5.         return None
6.     paths=[]
7.     for node in graph[start]:
8.         if node not in path:
9.             newpath = find_all_paths(graph, node, end, path)
10.            if newpath:
11.                paths.append(newpath)
12.    return paths

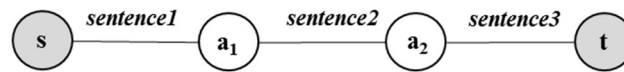
```

---

**Figure 3.7. Alpha Label graph  $g$  extraction Algorithm.**

### 3.4.2 Current Calculation

Each extracted knowledge path consists of a sentence or sequence of sentences comprehending the relation between two concepts. Each sentence presents a piece of knowledge. An example of a knowledge path  $K$  is  $s-a1-a2-t$ , as shown in Figure 3.8. Here, it consists of three sentences ( $sentence1$ ,  $sentence2$ , and  $sentence3$ ) representing three pieces of knowledge;  $sentence1$  comprehending the relation between  $s$  and  $a1$ ;  $sentence2$  comprehending the relation between  $a1$  and  $a2$ ; and  $sentence3$  comprehending the relation between  $a2$  and  $t$ . The amount of comprehension in each piece should be graded in order to grade the knowledge path. This can be achieved by calculating the current for each edge in the knowledge path.



**Figure 3.8.** An example of a knowledge path.

To calculate the current for each edge in *Alpha Label graph*  $g$ , given a weighted directed graph  $g = (C, E)$ , where  $C$  is a set of concepts and  $E$  is a set of edges representing the sentential relations between each pair of concepts. Here,  $g$  is interpreted as an equivalent electrical circuit of vertices and edges.  $R$  denotes the resistance of the edge  $e$  and each edge  $e$  represents a resistor with a resistance  $e \in E$ .  $\Omega$  denotes to the inverse of  $R$ . Suppose that the start concept  $s$  in  $g$  carries a voltage  $+1$  and the end concept  $t$  carries a voltage  $0$ .  $V_i$  denotes the voltage of vertex  $c_i$  and  $I_{i,j}$  denotes the current flow between each two concepts in  $g$ .

### *Ohm's Law*

To grade the comprehension amount of each piece of knowledge, Ohm's law can be applied to calculate the current between each two concepts in  $g$ :

$$\forall i, j: I_{i,j} = (V_i - V_j)/R_{i,j} \quad (3.1)$$

### *Kirchhoff's current law*

$$\forall i, j \neq s, t: \sum_i I(i, j) = 0 \quad (3.2)$$

The voltage at each concept of the circuit except the source and destination is calculated by combining Ohm's law and Kirchhoff's current law.

$$V_i = \sum_j V_j R_i / R_{i,j} \quad \forall i \neq s, t \quad (3.3)$$

Here,  $R_i = \sum R_{i,j}$  is the total resistance of the edges incident from concept  $c_i$ ,  $V_s = I$ , and  $V_t = 0$ . All of the current and the voltage equations are determined as a solution the linear equations.

To find the voltages of the concepts, the equations are expressed in a matrix form  $AX=B$ , where  $A$  is  $m$ -by- $m$  matrix,  $m$  represents the number of  $c_i$  concepts that are required to calculate their voltages, which are the number of all concepts in the graph except the source and destination concepts. Each row  $i$  in  $A$  represents a concept  $c_i$  and each cell  $a_{i,j}$  in  $i$  represents the weight of the edge incident from concept  $c_i$  to concept  $c_j$ ,  $i \neq j$ .  $a_{i,j}$  is the total weight of the edges incident from concept  $c_i$  in the case of  $i=j$ .  $B$  is a vector, where each  $b_i$  represents the weight of the edge between  $c_i$  and  $c_j$  multiplied by the voltage of  $c_j$ , where  $c_j$  is the source or the destination concept. Then the voltages of the  $c_i$  concepts are given by  $X=A^{-1}B$ , where  $A^{-1}$  is the inverse of  $A$ .



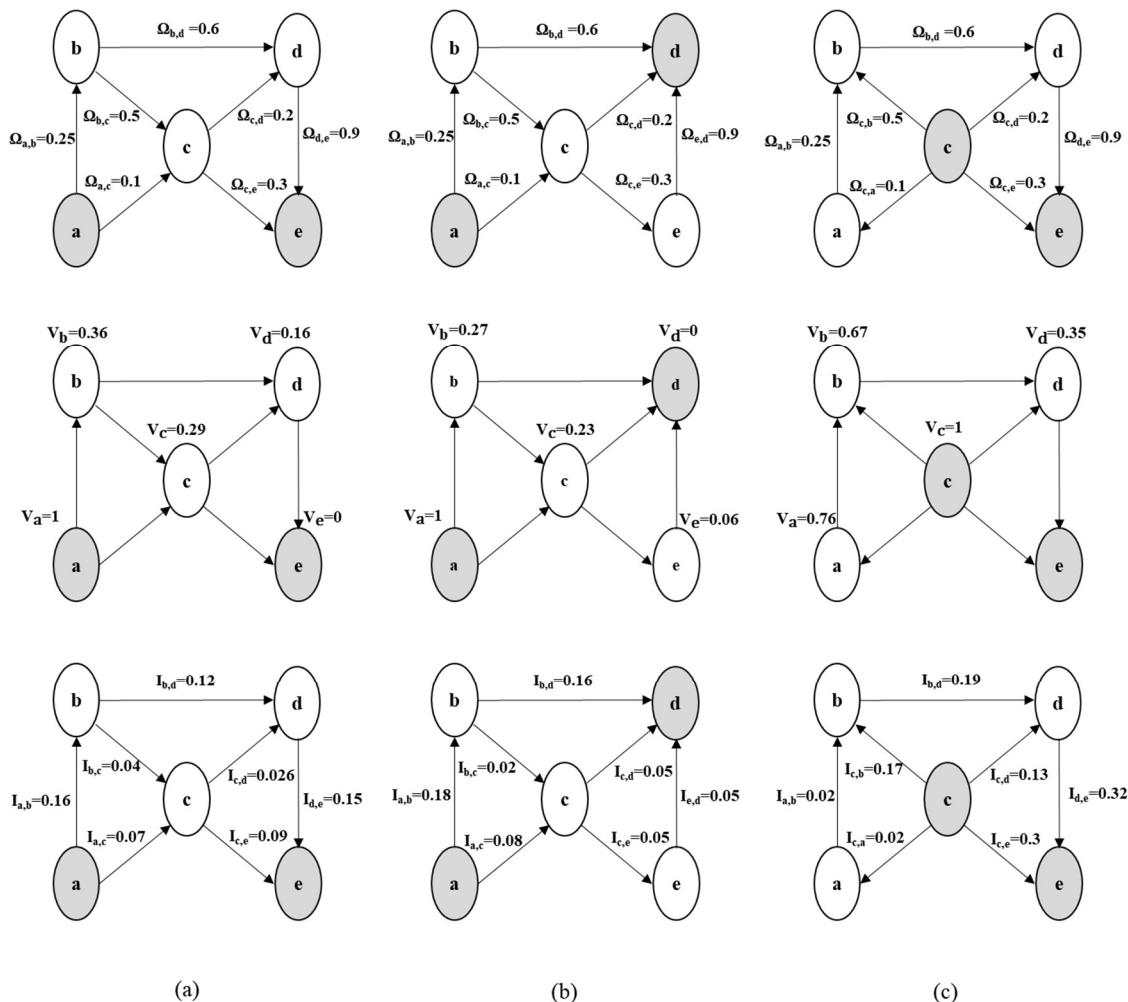


Figure 3.9. Example showing resistance, voltage, and current in three different Alpha Label graphs  $g$ .

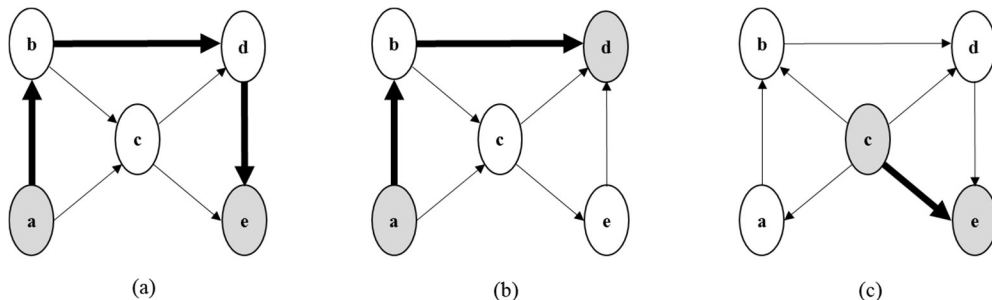


Figure 3.10. Example showing the "Alpha Knowledge Pathway"  $K'$  in the three Alpha Label graphs  $g$  in Figure 3.9.

Consider the Alpha Label graph  $g$  shown in Figure 3.9 (a), where  $V_a=1$ ,  $V_e=0$ . Equation 3.3 is applied to calculate the voltage of the three concepts  $V_b$ ,  $V_c$ , and  $V_d$  as the following:

$$V_b = \frac{(V_a \cdot R_b)}{R_{a,b}} + \frac{(V_c \cdot R_b)}{R_{b,c}} + \frac{(V_d \cdot R_b)}{R_{b,d}} \quad (3.4)$$

$$V_b = \frac{((1 \cdot 0.25) + (V_c \cdot 0.5) + (V_d \cdot 0.6))}{1.35} \quad (3.4.1)$$

$$1.35V_b = 0.25 + 0.5V_c + 0.6V_d \quad (3.4.2)$$

$$V_c = \frac{(V_a \cdot R_c)}{R_{a,c}} + \frac{(V_b \cdot R_c)}{R_{b,c}} + \frac{(V_d \cdot R_c)}{R_{c,d}} + \frac{(V_e \cdot R_c)}{R_{c,e}} \quad (3.5)$$

$$V_c = \frac{((1 \cdot 0.1) + (V_b \cdot 0.5) + (V_d \cdot 0.2) + (V_e \cdot 0.3))}{1.1} \quad (3.5.1)$$

$$1.1V_c = 0.1 + 0.5V_b + 0.2V_d + 0.3V_e \quad (3.5.2)$$

$$V_d = \frac{(V_b \cdot R_d)}{R_{b,d}} + \frac{(V_c \cdot R_d)}{R_{c,d}} + \frac{(V_e \cdot R_d)}{R_{e,d}} \quad (3.6)$$

$$V_d = \frac{((V_b \cdot 0.6) + (V_c \cdot 0.2) + (0 \cdot 0.9))}{1.7} \quad (3.6.1)$$

$$1.7V_d = 0.6V_b + 0.2V_c + 0 \quad (3.6.2)$$

After arranging the three linear equations, the resultant equations for  $b$ ,  $c$ , and  $d$  respectively are:

$$1.35V_b + -0.5V_c + -0.6V_d = 0.25 \quad (3.7)$$

$$0.5V_b + -1.1V_c + 0.2V_d = -0.1 \quad (3.8)$$

$$0.6V_b + 0.2V_c + -1.7V_d = 0 \quad (3.9)$$

To find the voltage of the concepts, the linear equations are expressed in a matrix form:

$$\begin{bmatrix} 1.35 & -0.5 & -0.6 \\ 0.5 & -1.1 & 0.2 \\ 0.6 & 0.2 & -1.7 \end{bmatrix} \begin{bmatrix} V_b \\ V_c \\ V_d \end{bmatrix} = \begin{bmatrix} 0.25 \\ -0.1 \\ 0 \end{bmatrix} \quad (3.10)$$

$$A = \begin{bmatrix} 1.35 & -0.5 & -0.6 \\ 0.5 & -1.1 & 0.2 \\ 0.6 & 0.2 & -1.7 \end{bmatrix}, \quad X = \begin{bmatrix} V_b \\ V_c \\ V_d \end{bmatrix}, \quad \text{and} \quad B = \begin{bmatrix} 0.25 \\ -0.1 \\ 0 \end{bmatrix} \quad (3.10.1)$$

$$\begin{bmatrix} V_b \\ V_c \\ V_d \end{bmatrix} = \begin{bmatrix} 1.35 & -0.5 & -0.6 \\ 0.5 & -1.1 & 0.2 \\ 0.6 & 0.2 & -1.7 \end{bmatrix}^{-1} \begin{bmatrix} 0.25 \\ -0.1 \\ 0 \end{bmatrix} \quad (3.10.2)$$

$$\begin{bmatrix} V_b \\ V_c \\ V_d \end{bmatrix} = \begin{bmatrix} 0.36 \\ 0.29 \\ 0.16 \end{bmatrix} \quad (3.10.3)$$

The voltage values for  $b$ ,  $c$ , and  $d$  are  $V_b \approx 0.36$ ,  $V_c \approx 0.29$ , and  $V_d \approx 0.16$ . By substituting in Equation 3.1, the resulting current for each edge in  $g$  is the following:  $I_{a,b} = 0.16$ ,  $I_{a,c} = 0.07$ ,  $I_{b,c} = 0.04$ ,  $I_{b,d} = 0.12$ ,  $I_{c,d} = 0.026$ ,  $I_{c,e} = 0.09$ , and  $I_{d,e} = 0.15$ . Figure 3.9 (b) and (c) present two more examples on calculating the current between two different source and destination concepts.

### 3.4.3 Skimmed Process

After grading each piece of knowledge in *the Alpha Label graph g*, the amount of comprehension achieved by each knowledge path should be graded in order to find the *Alpha Knowledge Pathway K'* comprehension of the relation between the two concepts. This can be achieved by calculating the delivered current flow for each knowledge path and selecting the one with the highest delivered current flow as the *Alpha Knowledge Pathway K'*.

Going back to Example 1, there are five knowledge paths between concept  $s$  and concept  $t$  ( $a-c-e$ ,  $a-b-d-e$ ,  $a-c-d-e$ ,  $a-b-c-e$ , and  $a-b-c-d-e$ ), where each of them is a

‘*sentence-chain path*’. To calculate the delivered current flow for the first knowledge path, 0.07 amber arrives from *a* to *c*, 0.09 amber moves from *c* to *e* (where the from *c* to *d* is 0.026 and the out from *c* to *e* is 0.09,  $0.09 + 0.026 = 0.12$ , noting that (the inverse of 0.12 is 8.33), the amber going from *c* to *e* is  $0.09 * 8.33 = 0.75$ ). Therefore, the total delivered current flow of the path equals  $0.07 * 0.75 = 0.05$ . Table 3.1 shows the delivered current flow for all the knowledge paths between the starting and ending concepts in Figure 3.9 (a, b, and c). Based on the consideration that the Alpha Knowledge Pathway *K*’ is a single knowledge path, along with the results shown in Table 3.1 (a), it is observed that *a-b-d-e* has the highest delivered current flow. Consequently, it is chosen as the “Alpha Knowledge Pathway” *K*’ between *a* and *e*.

**Table 3.1. The Delivered Current for All the Knowledge Paths in Figure 3.9**

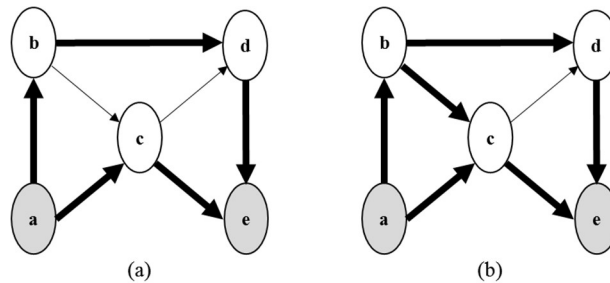
(a)			(b)			(c)		
K	Number of sentences	Delivered Current	K	Number of sentences	Delivered Current	K	Number of sentences	Delivered Current
1.	<i>a-c-e</i>	0.05	<i>a-b-c-d</i>	3	0.01	<i>c-a-b-d-e</i>	4	0.02
2.	<b><i>a-b-d-e</i></b>	<b>0.12</b>	<b><i>a-b-d</i></b>	<b>2</b>	<b>0.16</b>	<i>c-b-d-e</i>	3	0.17
3.	<i>a-c-d-e</i>	0.02	<i>a-b-c-e-d</i>	4	0.01	<i>c-d-e</i>	2	0.13
4.	<i>a-b-c-e</i>	0.04	<i>a-c-e-d</i>	3	0.04	<b><i>c-e</i></b>	<b>1</b>	<b>0.3</b>
5.	<i>a-b-c-d-e</i>	0.01	<i>a-c-d</i>	2	0.04	-	-	-

Suppose that selecting the Alpha Knowledge Pathway *K*’ is based on sentences budget. **Given** a set of knowledge paths  $K_1, K_2, K_3, \dots, K_n$ , of different delivered current flow, where  $n$  is the number of knowledge paths. **Find** Alpha Knowledge Pathway *K*’ based on budget  $u$ . In this case, *Bin packing algorithm – First Fit Decreasing* can solve this problem. Back to Example 1, when sorting the knowledge paths in descending order as

shown in Table 3.2, if the budget of the Alpha Knowledge Pathway  $K'$  is  $u=5$  sentences, then the Alpha Knowledge Pathway  $K'$  is the union of Knowledge Path 1 and 2 as shown in Figure 3.11 (a). If the budget of the Alpha Knowledge Pathway  $K'$  is  $u=6$  sentences, the Alpha Knowledge Pathway  $K'$  is the union of Knowledge Paths 1, 2, and 3 as shown in Figure 3.11 (b).

**Table 3.2. The Delivered Current for All the Knowledge Paths in Figure 3.8 sorting in descending order**

	K	Number of sentences	Delivered Current
1.	$a-b-d-e$	3	0.12
2.	$a-c-e$	2	0.05
3.	$a-b-c-e$	3	0.04
4.	$a-c-d-e$	3	0.02
5.	$a-b-c-d-e$	4	0.01



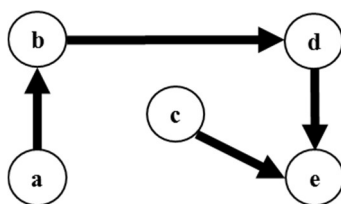
**Figure 3.11. Example showing the “Alpha Knowledge Pathway”  $K'$  in Example 1 based on sentences budget.**

#### 3.4.4 Skimmed Illuminated Knowledge Graph SIKG generator

After the *Alpha Knowledge Pathway*  $K'$  between each pair of concepts is selected, there is a need to collect all of the selected *Alpha Knowledge Pathway*  $K'$ s in order to

present their sentences to the reader as an *enhanced text*. This can be done by joining all of the *Alpha Knowledge Pathway K's* in a *Skimmed Illuminated Knowledge Graph SIKG*.

We initialize the Skimmed Illuminated Knowledge Graph *SIKG* to be empty. Then, the “*Alpha Knowledge Pathway*” *K* between each pair of concepts from the list of prose concepts  $C_L$  is added to it. For example, after adding all of the “*Alpha Knowledge Pathway*” *K's* in Figure 3.10, *SIKG* is shown in Figure 3.12. By using the *sentetest text generator ST()*, the *enhanced text* is derived from *SIKG* and presented to the reader to increase knowledge and enhance the comprehension. Presenting the sentences of the *enhanced text* “pre-order” according to the notion of traversing a tree.



**Figure 3.12. Skimmed Illuminated Knowledge Graph SIKG of Figure 3.10.**

### 3.5 Summary

The Chapter presented the details of the Knowledge Distillation Process. The process is designed to solve the problem that may be caused by the overabundance of knowledge resulting from the Knowledge Induction Process. The process is based on grading all of the augmented knowledge between each pair of concepts resulted from the Knowledge Induction Process, and selecting the most familiar, easiest-to-understand knowledge ‘*Alpha Knowledge Pathway*’ *K* that helps in comprehending the relation between the pair and presenting all the selected knowledge between each pair of concepts in an enhanced text. The chapter started by discussing the notion of the ‘goodness’ paths

from the measurements of a shortest-hubs path, a network flow, and the combination between them by using an equivalent electrical circuit. Next, it explained the grading process steps to grade each piece of knowledge between each pair of concepts, the skimmed process that selects the Alpha Knowledge Pathway between each pair. Then, it showed how all of the Alpha Knowledge Pathways are joined in a Skimmed Illuminated Knowledge graph to derive the enhanced text that is presented to the reader to enhance prose comprehension.

## CHAPTER 4

### **Computational Model Evaluation, Experiments and Results**

This chapter illustrates the details of the used computational evaluation model to assess the efficiency of the comprehension gained from the comprehension engine in both the Knowledge Induction Process (*the Illuminated Knowledge Graph IKG*) and the Knowledge Distillation Process (*the Skimmed Illuminated Knowledge Graph SIKG*). Next, both the design of the experiment and the content material are presented. Finally, the analysis of the results of each process is shown.

#### **4.1 Comprehension Model Evaluation**

To evaluate the proposed comprehension engine, there is a need for a computational evaluation model to measure the quantitative insight of the acquired knowledge, along with the learning process of prose comprehension obtained by the comprehension engine. Knowledge measurement is a difficult area still requiring significant exploration. “The fluid and intangible nature of knowledge makes its measurement an enormously complex and daunting task” (Mohamed A.F. Ragab & Amr Arisha, 2013) (Kankanhalli & Tan, 2008). Despite the amount of work done in the area of measurements, there is no clear answer about how to measure knowledge and the area is still very open. The acquired knowledge was assessed using: quantitative estimation, organization estimation, and comprehension efficiency.



#### 4.1.1 Information content

The amount of acquired knowledge influences comprehension, where the more acquired knowledge in the Illuminated Knowledge Graph  $IKG$  or the Skimmed Illuminated Knowledge Graph  $SIKG$  compares with the prose knowledge graph  $G_0$  expects to better enhance comprehension. This can be calculated by measuring the size of the obtained knowledge graph. The size of the graph is measured by the whole number of concepts  $C$  and the sentential relations  $E$  among them, where the concepts belong to three different sources prose  $LTX$ , reference text  $RTX$ , and Ontology Engine  $OE$ . In the Knowledge Induction Process, the knowledge graph transforms from  $(G_0, IKG_1, IKG_2, \dots, IKG_{final})$ , where each  $IKG_i$  represents an Illuminated Knowledge Graph after reading a new  $RTX_i$ , and  $IKG_{final}$  represents the Illuminated Knowledge Graph after reading all of the  $RTX_i$ . Therefore, the size of  $IKG$  is increased and knowledge is grown/overloaded respectively. For the Knowledge Distillation Process, however,  $SIKG$  is generated after joining the Alpha Knowledge Pathway of each pair of concepts belonging to the list of the prose concepts  $C_L$ . Similarly, the size of  $SIKG$  is increased and knowledge is grown/overload compared to the size of  $G_0$ . The knowledge growth rate  $\lambda$  can be calculated by Equation 4.1:

$$\lambda = \frac{|G|}{|G_0|} \quad (4.1)$$

For the Knowledge Induction Process,  $|G|$  denotes to the size of the Illuminated Knowledge Graph  $IKG$  after reading  $RTX_i$  and  $|G_0|$  denotes to the size of the prose knowledge graph. While in the Knowledge Distillation Process,  $|G|$  denotes to the size of

the Skimmed Illuminated Knowledge Graph *SIKG*. Together, the knowledge overload  $\gamma$  rate is also increased and it can be calculated by Equation 4.2:

$$\gamma = \frac{|G - G_0|}{|G_0|} \quad (4.2)$$

Calculating the amount of rare information that can be gained from the knowledge graph can be measured by the graph Entropy  $\delta$ , where high entropy is rare information in the knowledge graph and vice versa. According to (Shannon & Weaver, 1949), we calculate  $\delta$  using Equation 4.3:

$$\delta = - \sum_{i=0}^n p_i \log(p_i) \quad (4.3)$$

Where  $p_i$  is the probability of the outcome of concept  $c_i$  that is determined by Equation 4.4:

$$p_i = \frac{d_i}{2|E|} \quad (4.4)$$

$d_i$  represents the sentential relations of concept  $c_i$  and  $|E|$  is the number of the sentential relations  $E$  in the knowledge graph. To obtain a measure with a  $[0, 1]$  range,  $\delta$  is divided by  $\log(n)$ , where  $n$  represents the number of all the concepts in the knowledge graph  $|C|=n$ .

#### 4.1.2 Knowledge Organization

There is no doubt of the sentential relation  $E$ 's existence among concepts in the prose that affect comprehension, where the more the sentential relations  $E$  there are among the concepts, the more likely an understanding of the relations between them will occur. This falls under the notion of graph organization, which can be measured by calculating the cluster coefficient  $\beta$ . This offers a way to measure how strongly connected the concepts and their neighbors are in the knowledge graph (Drieger, 2013). According to (Watts & Strogatz, 1998), we suggest calculating  $\beta$  using Equation 4.5:

$$\beta = \sum_{i=0}^n \frac{2NIC_i}{d_i(d_i-1)} \quad (4.5)$$

Where  $NIC_i$  is the neighbors' interconnections coefficient of concept  $c_i$  which denotes to the number of the sentential relations between the first neighbors of concept  $c_i$  and  $d_i$  is the sentential relations of concept  $c_i$  which counts the first neighbors of concept  $c_i$ . The closer to 1 value indicates the higher clustered graph.

Knowledge graph organization can also be calculated by measuring the knowledge graph density  $\rho$ , which measures how the knowledge graph is to be completed and how well the concepts within it are integrated (Al Madi & Khan, 2015). A complete knowledge graph contains all possible sentential relations  $E$  and density equals 1. The graph density  $\rho$  is calculated by Equation 4.6 (Coleman & Moré, 1983).

$$\rho = \frac{|E|}{|C|( |C| - 1)} \quad (4.6)$$

### 4.1.3 Experiment

#### *Content Material*

An experiment was conducted on three proses  $LTX_i$ , and 8 concepts were selected as the list of the prose concepts  $C_L$  from each  $LTX_i$ . Wordnet (Miller, 1995) is a reliable Ontology Engine  $OE$  that has been used by many researchers in this field (Menaka & Radha, 2013; Chen, Chen, & Sun, 2010; Kamps, Marx, Mokken, De Rijke, & others, 2004). It is a huge lexical database developed by George Miller at the Cognitive Science Laboratory at Princeton University that is used as a dictionary of word senses and semantic relations between words (Menaka & Radha, 2013). This experiment utilized Wordnet

version 1.7. The reference text  $RTX$  used here is Wikipedia because it is considered one of the largest and most popular reference of articles on the internet (Conde, Larrañaga, Arruarte, Elorriaga, & Roth, 2016). For each  $RTX$ , there is a set of articles selected from Wikipedia about each concept in  $C_L$ . We applied the automated method used in a previous study (Babour, Nafa, & Khan, 2015) for the selection of the Wikipedia articles. For example, the 1<sup>st</sup> prose  $LTX1$  is about ‘Ethane chemical compound’ selected from *Encyclopedia Britannica* articles and the selected 8 concepts in its  $C_L$  are [‘Ethane’, ‘hydrocarbon’, ‘hydrogen’, ‘carbon’, ‘chemical’, ‘petroleum’, ‘carbonization’, ‘coal’]. Table 4.1 displays the selected  $LTX_i$ , as well as the  $C_L$  for each prose.

**Table 4.1. List of the proses used in the experiment**

	Prose Title	List of prose concepts $C_L$
<b>1<sup>st</sup> prose LTX1</b>	‘Ethane chemical compound’ (“ethane,” 2013)	[‘Ethane’, ‘hydrocarbon’, ‘hydrogen’, ‘carbon’, ‘chemical’, ‘petroleum’, ‘carbonization’, ‘coal’]
<b>2<sup>nd</sup> prose LTX2</b>	‘New Test for Zika OKed’ (Grens, 2016)	[‘Zika’, ‘infection’, ‘dengue’, ‘chikungunya’, ‘virus’, ‘aedes’, ‘mosquito’, ‘antibody’]
<b>3<sup>rd</sup> prose LTX3</b>	‘Anesthesia gases are warming the planet’ (DeMarco, 2015)	[‘Anesthetic’, ‘carbon’, ‘climate’, ‘oxide’, ‘desflurane’, ‘isoflurane’, ‘sevoflurane’, ‘halothane’]

The  $RTX_i$ s for LTX1 are [*Ethane, Hydrocarbon, Hydrogen, Carbon, Chemical substance, Petroleum, Carbonization, Coal*]. Table 4.2 shows details about each  $RTX_i$  for each  $LTX_i$ . For each  $LTX_i$ , the prose knowledge graph  $G_0$  represents the sentential relation among the concepts in its  $C_L$ . Then, in the Knowledge Induction Process, the system goes through all of the  $RTX$  and creates a set of Illuminated Knowledge Graphs  $IKG_i$ , where  $IKG_i$  represents the sentential relation among the concepts in  $C_L$  after reading a new  $RTX_i$ .

**Table 4.2. Break Down of the readable reference texts in each LTX**

	<b>1<sup>st</sup> prose LTX<sub>1</sub></b>	<b>2<sup>nd</sup> prose LTX<sub>2</sub></b>	<b>3<sup>rd</sup> prose LTX<sub>3</sub></b>
<b>1<sup>st</sup> reference text RTX1</b>	Ethane (“Ethane,” 2016)	Zika fever (“Zika fever,” 2016)	Anesthetic (“Anesthetic,” 2016)
<b>2<sup>nd</sup> reference text RTX2</b>	Hydrocarbon (“Hydrocarbon,” 2016)	Infection (“Infection,” 2016)	Carbon (“Carbon,” 2016)
<b>3<sup>rd</sup> reference text RTX3</b>	Hydrogen (“Hydrogen,” 2016)	Dengue fever (“Dengue fever,” 2016)	Climate (“Climate,” 2016)
<b>4<sup>th</sup> reference text RTX4</b>	Carbon (“Carbon,” 2016)	Chikungunya (“Chikungunya,” 2016)	Oxide (“Oxide,” 2016)
<b>5<sup>th</sup> reference text RTX5</b>	Chemical substance (“Chemical substance,” 2016)	Virus (“Virus,” 2016)	Desflurane (“Desflurane,” 2016)
<b>6<sup>th</sup> reference text RTX6</b>	Petroleum (“Petroleum,” 2016)	Aedes (“Aedes,” 2016)	Isoflurane (“Isoflurane,” 2016)
<b>7<sup>th</sup> reference text RTX7</b>	Carbonization (“Carbonization,” 2016)	Mosquito (“Mosquito,” 2016)	Sevoflurane (“Sevoflurane,” 2016)
<b>8<sup>th</sup> reference text RTX8</b>	Coal (“Coal,” 2016)	Antibody (“Antibody,” 2016)	Halothane (“Halothane,” 2016)

### *Results*

This section contains the analysis of information gained from both the Illuminated Knowledge Graph  $IKG_i$  of prose  $LTX_i$  in the Knowledge Induction Process and from the Skimmed Illuminated Knowledge Graph  $SIKG$  of prose  $LTX_i$  in the Knowledge Distillation Process.

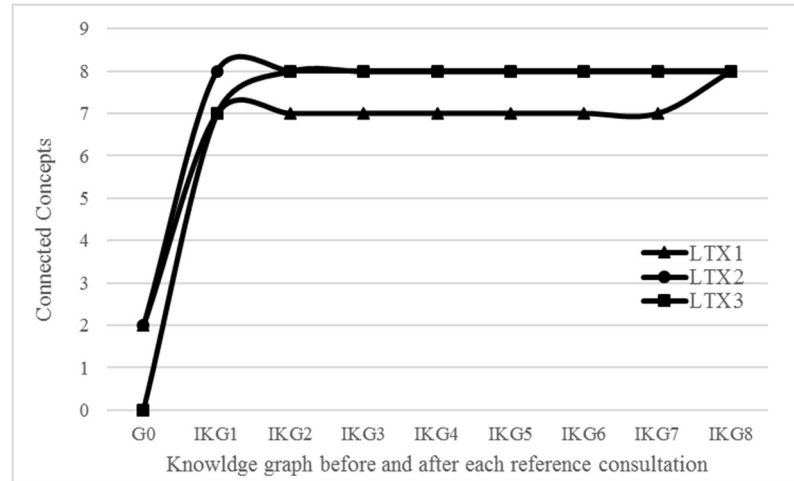
For the Knowledge Induction Process, Table 4.3 shows the average time needed to read the prose, the 8 references for each of the proses and finding the highest familiarity sentential relations connecting the prose concepts in each reference. As seen, the average

time needed to read each prose with its related references is in few minutes, which is considered a short amount of time.

**Table 4.3. The average time for reading the prose, the references and finding the highest familiarity sentential relations connecting the prose concepts in each reference in (h:m:s)**

	Time
<b>1<sup>st</sup> prose LTX<sub>1</sub></b>	(0:04:16)
<b>2<sup>nd</sup> prose LTX<sub>2</sub></b>	(0:05:18)
<b>3<sup>rd</sup> prose LTX<sub>3</sub></b>	(0:02:07)

Figure 4.1 shows the number of connected concepts found in the knowledge graph's list of prose concepts  $C_L$  before and after reading the reference texts  $RTX_i$ , the x-axis refers to the prose knowledge graph  $G_0$  and the Illuminated Knowledge Graph  $IKG_i$  after reading each reference text  $RTX_i$ , and the y-axis is the number of connected concepts per each knowledge graph. When the process reads a new  $RTX_i$ , the number of the connected concepts increased. It is observable that for  $LTX_1$ , the number of the connected concepts jumped from 2 to 7 after reading  $RTX_1$ , and got as high as 8 connected concepts after reading  $RTX_8$ . For  $LTX_2$ , two of the concepts are connected in the prose knowledge graph  $G_0$ , while the concepts become fully connected after reading  $RTX_1$ . Meanwhile in  $LTX_3$ , the connected concepts jumped from 0 to 7 after reading  $RTX_1$  and reached up to 8 connected concepts after reading  $RTX_3$ . This denotes the effectiveness of reading reference texts for connecting the prose concepts and illuminating sentential relations among them, thus enhancing prose comprehension.



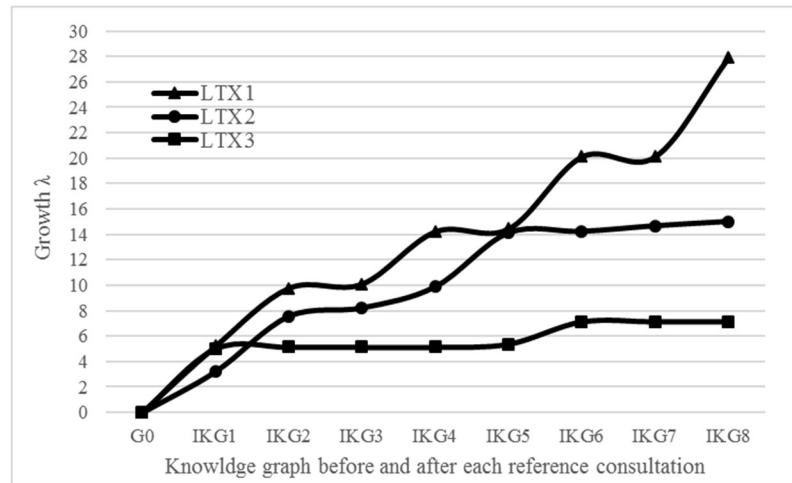
**Figure 4.1. Prose Concepts connectivity per the knowledge graphs.**

The breakdown of the total number of concepts  $C$  and the number of the sentential relations  $E$  in the prose knowledge graph  $G_0$  and the Illuminated Knowledge Graph  $IKG_{final}$  for each prose  $LTX_i$  are shown in Table 4.4, where the concepts are from the prose  $LTX$ , reference text  $RTX$ , and/or the Ontology Engine  $OE$ . As seen, there is a great variance in the number of concepts and the number of sentential relations between  $G_0$  and  $IKG_{final}$ . Increasing the size may be an indication to increasing the information. So, increasing the size in  $IKG_{final}$  is a good indicator to the plentiful information in the  $IKG_{final}$  which will further support in reinforcement of the prose comprehension.

**Table 4.4. Break Down of the total number of sentential relation and concepts in the three proses**

	1 <sup>st</sup> prose LTX <sub>1</sub>		2 <sup>nd</sup> prose LTX <sub>2</sub>		3 <sup>rd</sup> prose LTX <sub>3</sub>	
	G <sub>0</sub>	IKG <sub>final</sub>	G <sub>0</sub>	IKG <sub>final</sub>	G <sub>0</sub>	IKG <sub>final</sub>
<b>Sentential relation</b>	1	168	1	90	0	33
<b>Number of prose <math>LTX</math> concepts</b>	8	8	8	8	8	8
<b>Number of reference text <math>RTX</math> concepts</b>	0	12	0	16	0	5
<b>Number of ontology engine <math>OE</math> concepts</b>	0	63	0	21	0	11

Figure 4.2 displays the information growth  $\lambda$  per knowledge graph in each prose  $LTX_i$ , where the x-axis represents the prose knowledge graph  $G_0$  and the Illuminated Knowledge Graph  $IKG_i$  after reading each reference text  $RTX_i$ , and the y-axis represents the growth rate  $\lambda$ . In  $LTX1$ , the information is shown to have grown gradually after reading reference texts  $RTX1$  to  $RTX6$ , while no new information was added after reading  $RTX7$ . Interestingly, it started to increase again after reading  $RTX8$ . In  $LTX2$ , the information continued to grow gradually from reading  $RTX1$  to  $RTX8$ . In  $LTX3$ , information increase fluctuated; the information grew after reading  $RTX1$ , then no new information was added after reading  $RTX2$  to  $RTX4$ , it increased again after reading  $RTX5$  and  $RTX6$ , then no new information was added after reading  $RTX7$  and  $RTX8$ . This implies that reading references in some cases has a positive effect in increasing the knowledge while in other cases it has no effect.

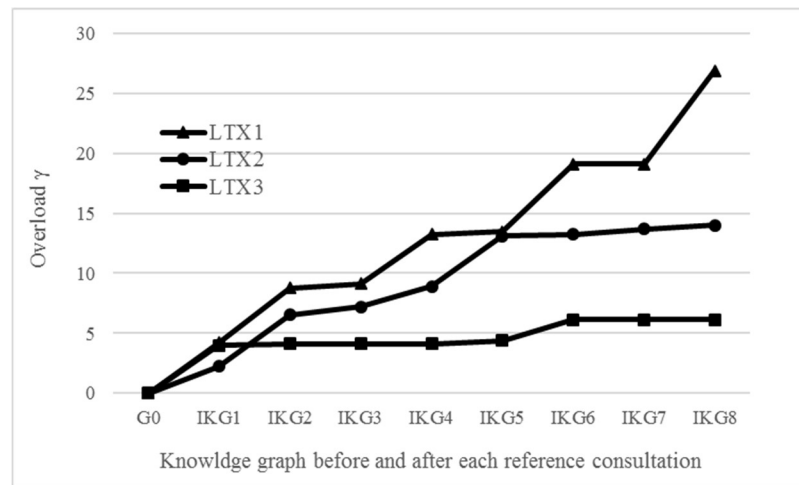


**Figure 4.2. Information growth rate  $\lambda$  per the knowledge graphs.**

The information overload rate  $\gamma$  in the knowledge graph is shown in Figure 4.3. Here, the x-axis represents the prose knowledge graph  $G_0$  and the Illuminated Knowledge



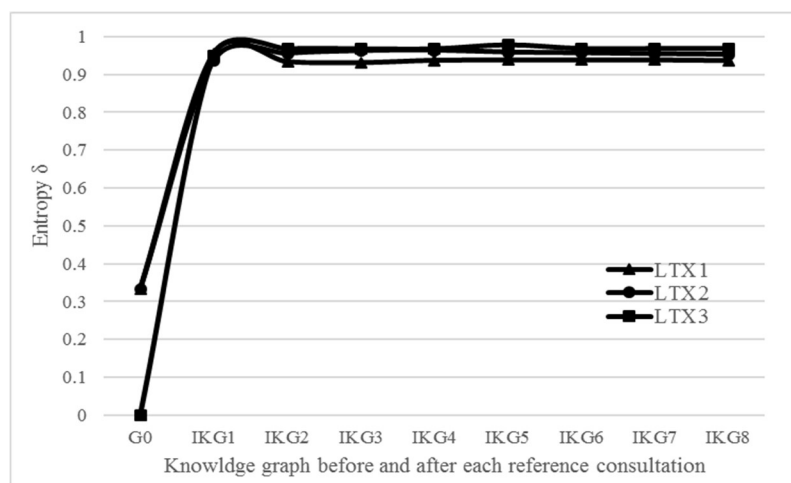
Graph  $IKG_i$  after reading each reference text  $RTX_i$ , and the y-axis represents the overload rate  $\gamma$ . Similarly, it is clear that information overload varies from being slightly to highly in the three proses  $LTX$  after reading new reference texts  $RTX_i$ .



**Figure 4.3. Information overload rate  $\gamma$  per the knowledge graphs.**

Figure 4.4. represents the entropy  $\delta$  per each knowledge graph, where the x-axis stands for both the prose knowledge graph  $G_0$  and the Illuminated Knowledge Graph  $IKG_i$  after reading each reference text  $RTX_i$ , and the y-axis is the entropy  $\delta$ . It was observed that  $\delta$  in the three proses  $LTX$  began with low values, then it jumped to high values after reading the 1<sup>st</sup> reference text  $RTX_1$ . This implies that  $RTX_1$  in the three proses  $LTX$  contributed to adding high amount of rare information. Then, in  $LTX_1$  the entropy value increased slightly after having read  $RTX_2$  to  $RTX_6$  and  $RTX_8$ . This indicates that texts  $RTX_2$  to  $RTX_6$  and  $RTX_8$  added a low amount of rare information, while no new information was added after reading  $RTX_7$ . In  $LTX_2$ , there was a little increase in entropy after reading texts  $RTX_2$  to  $RTX_8$ ; thus, a low amount of rare information was added. In  $LTX_3$ , there was also a slight increase in the entropy after reading  $RTX_2$  and  $RTX_5$ ; a low amount of rare information

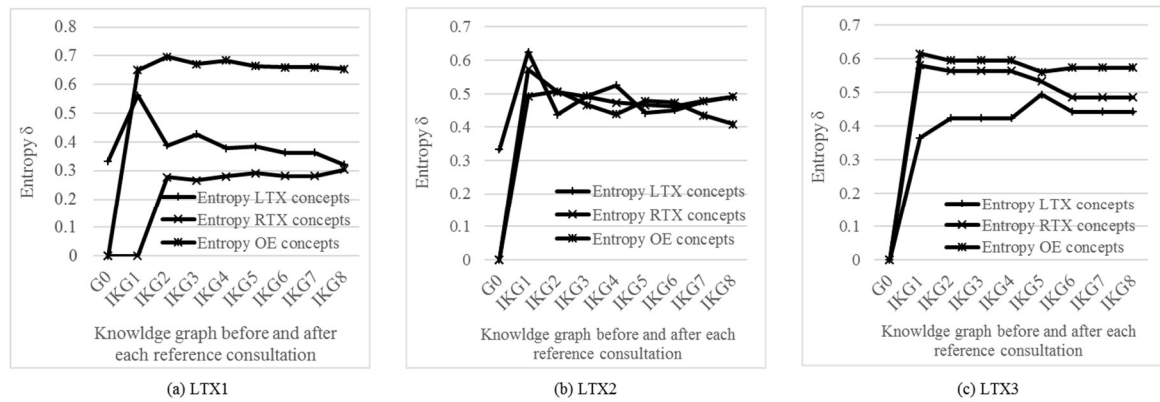
was added here, and no new information was added after having read texts *RTX3* to *RTX4*, and *RTX6* to *RTX8*. This verifies that some of the reference texts are highly effective in adding rare information, while other texts have little to no effect.



**Figure 4.4. Entropy  $\delta$  per the knowledge graphs.**

Additionally, Figure 4.5 represents the amount of rare information that was gained by the contribution of the prose *LTX*, reference text *RTX*, and the ontology engine *OE* concepts in the three *LTXs* separately. As shown in 4.5 (a), for *LTX1*, the entropy value of *OE* concepts jumped to a higher value than that of *LTX* and *RTX* concepts after reading texts *RTX1* to *RTX8*. This finding denotes that the amount of rare information added by the contribution of *OE* concepts was higher than the amount added by the contribution of *LTX* and *RTX* concepts. In 4.5 (b), for *LTX2*, the highest entropy value of *LTX*, *OE*, and *RTX* concepts varied after reading *RTX1* to *RTX8*; this implied that the highest amount of rare information gained by the contribution of *LTX*, *OE*, or *RTX* concepts was not concentrated on the contribution of any one of them. In 4.5 (c), for *LTX3*, the entropy value of *OE* concepts was the highest among *LTX* and *RTX* concepts after having read *RTX1* to *RTX8*.

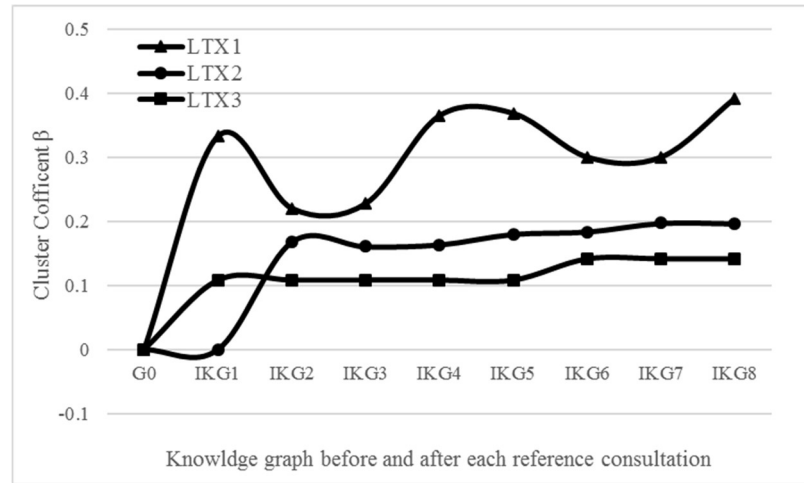
This refers to amount of rare information added by the contribution of *OE* concepts being higher than the amount added by the contribution of *LTX* and *RTX* concepts. This refers to the importance of *LTX*, *RTX*, and *OE* concepts in adding rare information and that there is no preference among them.



**Figure 4.5. Break down of the Entropy  $\delta$  per the knowledge graphs for the three LTX.**

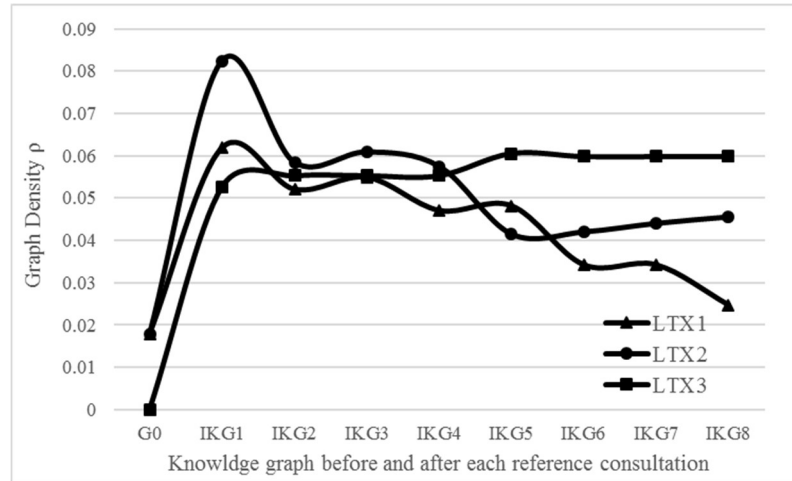
Moreover, the clustering coefficient  $\beta$  observed in each knowledge graph is shown in Figure 4.6, where the x-axis refers to the prose knowledge graph  $G_0$  and the Illuminated Knowledge Graph  $IKG_i$  after reading each reference text  $RTX_i$ , and the y-axis represents the clustering coefficient  $\beta$ . It is obvious that some of the graphs are highly clustered; this implies that many of the concepts within these graphs are highly related to each other. For *LTX1*, the graph tended to be highly clustered, especially after reading *RTX1*, *RTX4*, *RTX5* and *RTX8*. In *LTX2*, the cluster coefficient jumped to a higher value after reading *RTX2*, lowered slightly after reading *RTX3*, then it slightly increased after reading *RTX4* to *RTX8*. In *LTX3*, the cluster coefficient increased after reading *RTX1*, became steady after reading *RTX2* to *RTX5*, increased again after reading *RTX6*, then went back to being steady after reading *RTX7* and *RTX8*. This indicates that some of the reference texts increased

clustering of the concepts together, indicating an enhancement of the comprehension, while others decreased or did not affect clustering of the concepts.



**Figure 4.6. Cluster Coefficient  $\beta$  per the knowledge graphs.**

Additionally, Figure 4.7 gives us a view of the integration among the concepts contained within each knowledge graph, where the x-axis represents the prose knowledge graph  $G_0$  and the Illuminated Knowledge Graph  $IKG_i$  after reading each reference text  $RTX_i$ , and the y-axis represents the graph density  $p$ . According to the graph, in the three proses  $LTX$ , the  $p$  went between high and low values, indicating the variance among the reference texts for adding sentential relations among the concepts, making them closely.



**Figure 4.7. Density  $\rho$  per the knowledge graphs.**

For the Knowledge Distillation Process, Table 4.5 shows the time needed to read the 8 references for each of the proses, finding the highest familiarity in the sentential relations connecting the prose concepts in each reference, grading all of the knowledge paths between each pair of concepts in the list of prose concepts  $C_L$ , and selecting the Alpha Knowledge Pathway  $K'$  between each pair of concepts. In the experiment, Alpha Knowledge Pathway is considered to be a single knowledge path designed to help in comprehending the relation between pair of concepts. Therefore, the resulting Skimmed Illuminated Knowledge Graph  $SIKG$  works as the minimal Skimmed Illuminated Knowledge Graph.

**Table 4.5. The average time for reading the prose, the references, finding the highest familiarity sentential relations connecting the prose concepts in each reference and finding the Alpha Knowledge Pathway between each pair in (h:m:s)**

	Time
1 <sup>st</sup> prose LTX <sub>1</sub>	(0:05:19)
2 <sup>nd</sup> prose LTX <sub>2</sub>	(0:06:16)
3 <sup>rd</sup> prose LTX <sub>3</sub>	(0:02:13)

Table 4.6 shows the breakdown of the total number of concepts  $C$ , as well as the number of sentential relations  $E$  in the prose knowledge graph  $G_0$  and the Skimmed Illuminated Knowledge Graph  $SIKG$ . It is observed that the knowledge graph size, which is represented in the number of concepts and sentential relations  $E$  increased in the  $SIKG$ , in contrast to the  $G_0$ . It can be assumed that increasing the size indicates an increase of information. Therefore, the increasing the size of  $SIKG$  would imply that  $SIKG$  has more information than  $G_0$ .

**Table 4.6. Break Down of the total number of the sentential relations and concepts in the three proses**

	1 <sup>st</sup> prose LTX <sub>1</sub>		2 <sup>nd</sup> prose LTX <sub>2</sub>		3 <sup>rd</sup> prose LTX <sub>3</sub>	
	G <sub>0</sub>	SIKG	G <sub>0</sub>	SIKG	G <sub>0</sub>	SIKG
<b>Sentential relation</b>	1	28	1	22	0	23
<b>Number of prose LTX concepts</b>	8	8	8	8	8	8
<b>Number of reference text RTX concepts</b>	0	9	0	7	0	4
<b>Number of ontology engine OE concepts</b>	0	2	0	0	0	8

The quantitative metrics for  $G_0$  and  $SIKG$  are shown in Table 4.7. From Table 4.7, it is shown that in the three proses  $LTX$ , not all eight concepts were connected in the prose knowledge graph  $G_0$ , whereas they become fully connected in the Skimmed Illuminated Knowledge Graph  $SIKG$ . This indicates the efficiency of the Knowledge Distillation

Process for connecting the concepts in the list of the prose concepts  $C_L$ . At the same time, it is obvious that the information in  $SIKG$  grown and overload when compared with  $G_0$ . This verifies that  $SIKG$  contains more information than  $G_0$ . Entropy in the three  $LTXs$  is also raised, concluding that  $SIKG$  is richer with rare information than  $G_0$ . As it is shown in the graph, for  $LTX1$  and  $LTX2$ , the rare information gained by the contribution of the  $LTX$  concepts was the greater than the amount of information gained by the contribution of  $RTX$  and  $OE$  concepts. In  $LTX3$ , the rare information gained by the contribution of the  $OE$  concepts is the highest when compared with the gains of the  $LTX$  and  $RTX$  concepts. Furthermore, the cluster coefficient also increased for  $SIKG$  in  $LTX1$  and  $LTX2$ . One can infer that their concepts were better organized than those in  $G_0$ , while the cluster coefficient remained 0 in  $LTX3$ , leading to the idea that there are no sentential relations among the concepts' respective neighbors. Furthermore, the density in  $SIKG$  increased here; this suggests that the concepts in  $SIKG$  were closer than those in  $G_0$ .

**Table 4.7. Basic graph metrics analysis for the prose knowledge graph  $G_0$  and the Skimmed Illuminated Knowledge Graph  $SIKG$**

Metric	1 <sup>st</sup> prose $LTX_1$		2 <sup>nd</sup> prose $LTX_2$		3 <sup>rd</sup> prose $LTX_3$	
	$G_0$	$SIKG$	$G_0$	$SIKG$	$G_0$	$SIKG$
Connected concepts	2	8	2	8	0	8
Graph Growth	0	5.22	0	4.11	0	5.38
Graph Overload	0	4.22	0	3.11	0	4.36
Entropy	0.33	0.97	0.33	0.97	0	0.98
Entropy $LTX$ concepts		0.67		0.78		0.53
Entropy $RTX$ concepts		0.52		0.51		0.45
Entropy $OE$ concepts		0.45		0		0.58
Cluster Coefficient	0	0.06	0	0.03	0	0
Graph Density	0.018	0.08	0.02	0.1	0	0.06

#### 4.1.4 Comprehension Efficiency

To measure the efficiency of the comprehension, reachability among the prose concepts in the knowledge graphs needs to be calculated, where reachability measures the efficiency of traversal between each pair of prose concepts in the knowledge graph. This can be measured by calculating the diameter of the knowledge graph where the graph diameter is the longest of all the shortest Knowledge Paths between any two concepts in the knowledge graph (Minor & Urban, 2008). It is assumed that the diameter is the time needed to reach from one of the prose concepts to another. Table 4.8 shows the diameter of the knowledge graphs. As seen in the Knowledge Induction Process, it took a long time to reach from one prose concept to another when compared to the Knowledge Distillation Process.

**Table 4.8. Diameter of the Knowledge Graphs**

	<b>Knowledge Induction Process</b>	<b>Knowledge Distillation Process</b>
<b>1<sup>st</sup> prose LTX1</b>	15	5
<b>2<sup>nd</sup> prose LTX2</b>	10	4
<b>3<sup>rd</sup> prose LTX3</b>	19	5

The state of comprehension gained from the acquired knowledge needs to be assessed. This can be simulated from the concept recognition rates, which are affected by their relations with their neighbor concepts. For the purpose of this simulation, a novel measure was built to calculate the illumination value for each of the knowledge graph concepts, where the measure was based on the strength of the sentential relations among



the concepts and their neighbor concepts in the knowledge graph, along with prior knowledge of the concepts themselves.

Each concept in the knowledge graph underwent many phases  $\Theta$  for calculating its illuminated value  $h$  until it reached a stable value (also known here as its final illuminated value), where  $I$  refers to a fully illustrated concept. The phase  $\Theta_i$  is considered to be the reading of a set of sentences. At each phase  $\Theta_i$ , the concept represented by  $h$  referred to its current illumination value. The learning process at each phase was assessed by the value of  $|H|$  (the summation of  $h$  for each concept  $c_i$  in the list of prose concepts  $C_L$ ). The higher the value of  $|H|$ , the more the comprehension. To calculate  $h$  for each concept in the knowledge graph in each phase  $\Theta_i$ , Equation 4.7 (as seen below) was utilized. Solving 4.7 results  $|H|$  in each phase individually. Here,  $(A^T.H(\Theta).\alpha)$  represents the impact of the neighbors of each concept (the concepts that have sentential relations with the concept) on its illumination values;  $\alpha$  equals 0.5; and  $H(\Theta)$  represents the prior knowledge of the concept itself.

$$H(\Theta + 1) = (A^T \cdot H(\Theta) \cdot \alpha) + H(\Theta) \quad (4.7)$$

$$\begin{bmatrix} h_1 \\ h_2 \\ \dots \\ \dots \\ h_n \end{bmatrix} = \begin{bmatrix} a_{1,1} & \dots & \dots & \dots & \dots \\ a_{1,2} & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ a_{1,n} & \dots & \dots & \dots & a_{n,n} \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ h_2 \\ \dots \\ \dots \\ h_n \end{bmatrix} \cdot \alpha + \begin{bmatrix} h_1 \\ h_2 \\ \dots \\ \dots \\ h_n \end{bmatrix}$$

$A$  is an  $n$ -by- $n$  matrix,  $n$  representing the number of all of the concepts in the knowledge graph  $|C|=n$ . Each cell  $a_{i,j}$  in  $A$  represents the value of the sentential relation strength (weight) between two concepts. The value of  $a_{i,j}$  is calculated by Equation 4.8 (as seen below), where  $f_{i,j}$  represents the frequency of the relation type between concept  $c_i$  and

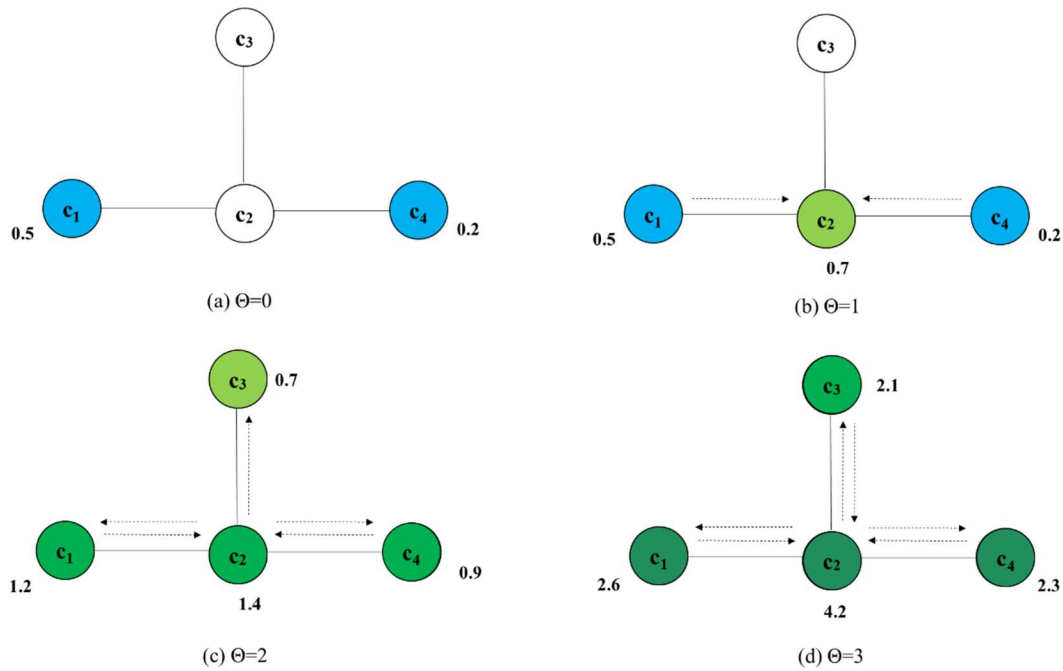
$c_j$  comes from the words frequency in the Gutenberg Project (Hart, 1971). The relationship between  $f_{i,j}$  and  $a_{i,j}$  is a direct relation, where the high frequency stands for high familiarity of the relation type.

$$a_{i,j} = -1/\log\left(\frac{f_{i,j}}{10^9}\right) \quad (4.8)$$

$H = \{h_1, h_2, \dots, h_n\}$  is a vector of the concept illumination values. Each value  $h_i$  in  $H$  represents an initial value for a concept. This initial value represents prior knowledge or the familiarity of the concept. It is calculated by Equation 4.9. Here,  $f_i$  is the frequency of the concept extracted from data from the Gutenberg Project (Hart, 1971). The relation between  $f_i$  and  $h_i$  is a direct relation, where a high frequency stands for a high familiarity of the concept.

$$h_i(0) = -1/\log\left(\frac{f_i}{10^9}\right) \quad (4.9)$$

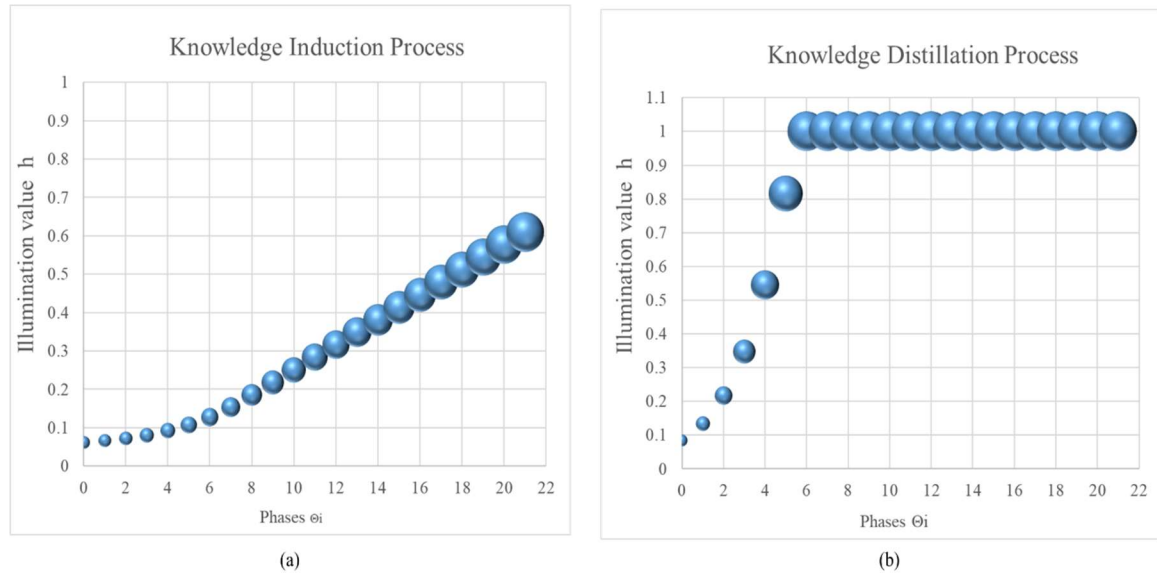
Figure 4.8 is an example that shows the variance of the concept illumination value  $h_i$  in the learning process over three phases  $\Theta_i$ . In this example, the sentential relation strength  $a_{i,j}$  between each pair of concepts  $c_i$  and  $c_j$  in the knowledge graph equals 1.  $h_i$  in Phase-0  $\Theta_0$  refers to the initial illumination value for each concept.



**Figure 4.8. Example of the variance of the illuminating value  $h_i$  in the learning process over three phases.**

One interesting growth characteristic that was tracked is the evaluation of the concepts values through the learning process over phases  $\theta$ . The value of each concept  $h_i$  at each phase  $\theta$  was calculated, where  $1$  refers to that the concept is fully illustrated. Figure 4.9. provides an example of the variance of  $h_i$  of the concept ‘carbonization’. In LTX1, it varied over 21 phases in (a) the Illuminated Knowledge Graph  $IKG_{final}$  and (b) the Skimmed Illuminated Knowledge Graph  $SIKG$ , where the size of the concept over the phases refers to its  $h_i$ . It can be seen that the increases in size over the phases indicates that the concept became more and more illustrated. However, it is obvious that  $h$  reaches 0.6 over 21 phases in the Knowledge Induction Process, which means that the concept was not fully illustrated over the 21 phases. Meanwhile in the Knowledge Distillation Process, the

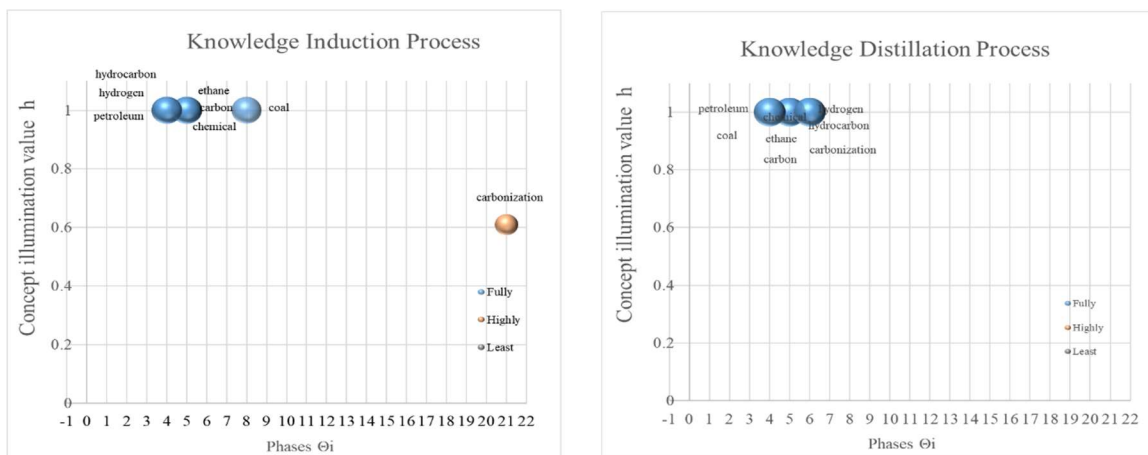
concept reaches 1 in phase 6, which means that the concept became fully illustrated in an earlier phase.



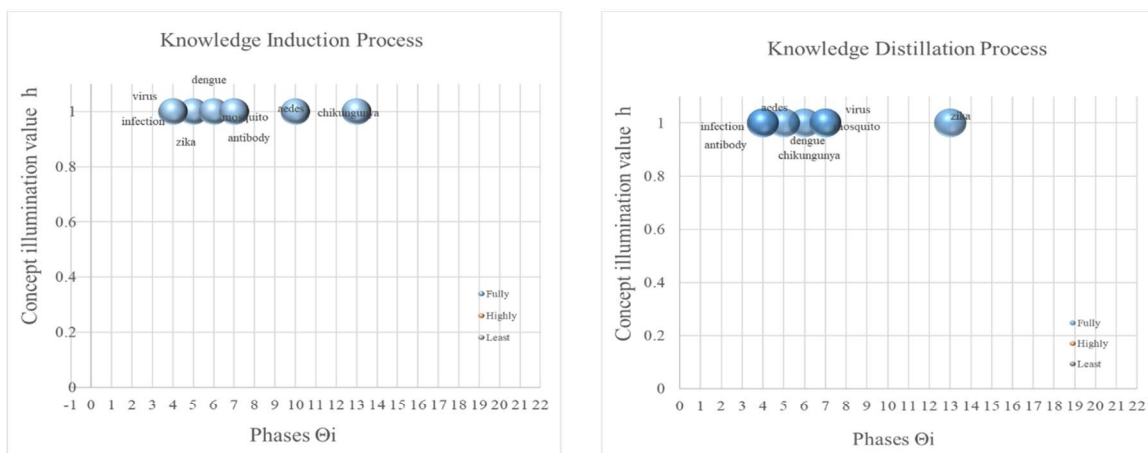
**Figure 4.9. Illustration values per phases for concept *carbonization* in (a) Knowledge Induction Process (Illuminated Knowledge Graph IKG) and (b) Knowledge Distillation Process (Skimmed Illuminated Knowledge Graph SIKG).**

It could be possible that some concepts were fully illustrated, while others were not. The results were then divided into three parts. The first part was for concepts that have values  $h_i(\Theta) = 1$  which are considered fully illustrated. The second part was for  $0.1 \leq h_i(\Theta) < 1$  when the concepts are highly illustrated and the third part for  $0 \leq h_i(\Theta) < 0.1$  when the concepts are the least illustrated. Figure 4.10 shows the illumination values  $h_i$  for each of the prose concepts  $C_L$  against the phases  $\Theta_i$ , where the x-axis represents  $\Theta_i$  and the y-axis represents the concept illumination value in which they were illustrated in the knowledge graph. The size of each concept indicates its illumination value  $h_i$ , and the color indicates to which part it belongs. For LTX1 in the Knowledge Induction Process, 7 of the 8 concepts were fully illustrated and one concept was considered highly illustrated, while in the

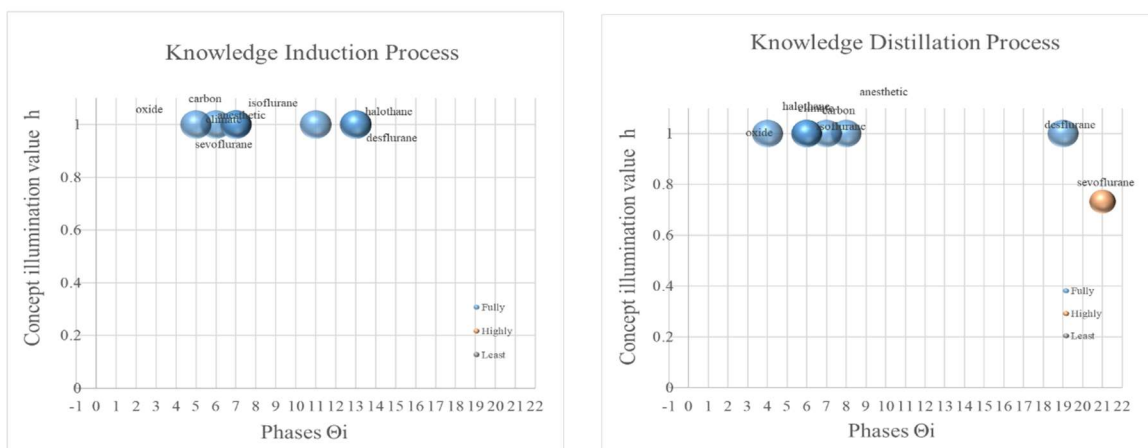
Knowledge Distillation Process, all 8 of the concepts were fully illustrated. For LTX2 in the Knowledge Induction Process and the Knowledge Distillation Process, all of the concepts became fully illustrated. For LTX3 in the Knowledge Induction Process, all of the concepts were fully illustrated, while in the Knowledge Distillation Process, 7 of the 8 were fully illustrated and one was highly illustrated. It can be seen that most of the concepts were fully illustrated either in the Knowledge Induction Process or the Knowledge Distillation Process; this means that they have strong connections with the concepts in the knowledge graph. In addition, it is observable that some of the fully illustrated concepts were recognized in earlier phases of the Knowledge Induction Process and later in the Knowledge Distillation Process and vice versa. For example, in LTX1, '*hydrocarbon*' was recognized in phase 4 of the Knowledge Induction Process and in phase 6 of the Knowledge Distillation Process, while '*coal*' is recognized in phase 8 of the Knowledge Induction Process and in phase 4 of the Knowledge Distillation Process. In addition, some of the concepts are fully illustrated in the Knowledge Induction Process but is highly illustrated in the Knowledge Distillation Process; and vice versa. For example, in LTX1, '*carbonization*' is highly illustrated in the Knowledge Induction Process but is fully illustrated in the Knowledge Distillation Process, while '*sevoflurane*' is fully illustrated in the Knowledge Induction Process but is highly illustrated in the Knowledge Distillation Process. All these differences are due to the number of sentential relations among the concepts of the knowledge graph.



(a) LTX1



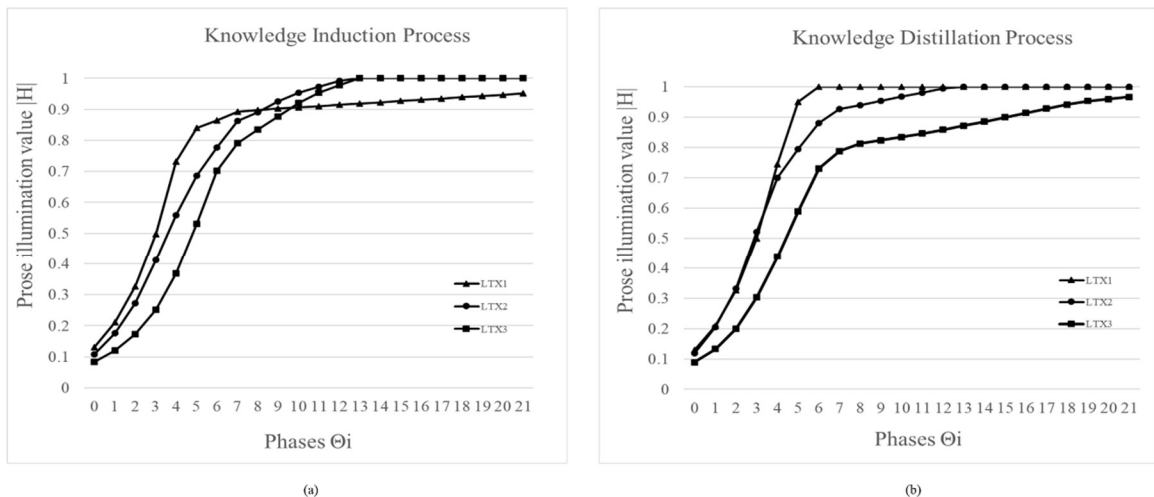
(b) LTX2



(c) LTX3

**Figure 4.10. Correlation among phases  $\Theta_i$  and the concept illumination value  $h_i$  for LTX1, LTX2, and LTX3 in the knowledge graphs.**

Figure 4.11 plots the variance in the concepts illumination value of the prose concepts  $C_L$ , where the x-axis represents the phases  $\Theta_i$  of the learning process and the y-axis represents the prose illumination values  $|H|$  in the knowledge graphs. 21 phases were examined. In the Knowledge Induction Process, it was noticed that in LTX1, the prose illumination value  $|H|$  of  $C_L$  nearly reached 1 over the 21 phases, which means that the prose concepts were highly illustrated in the Knowledge Induction Process, while the prose illumination value  $|H|$  of  $C_L$  reached 1 in phase 6 in the Knowledge Distillation Process. For LTX2, the prose illumination value  $|H|$  of  $C_L$  reached 1 in phase 13 for both the Knowledge Induction Process and the Knowledge Distillation Process. In LTX3, the prose illumination value  $|H|$  of  $C_L$  got up to 1 in phase 13 in the Knowledge Induction Process and nearly reached 1 over the 21 phases in the Knowledge Distillation Process. Again, the differences come back to the number of the sentential relation among the concepts of the knowledge graph.



**Figure 4.11. Illustration values per phases for LTX1, LTX2, and LTX3 (a) Knowledge Induction Process (Illuminated Knowledge Graph IKG) and (b) Knowledge Distillation Process (Skimmed Illuminated Knowledge Graph SIKG).**

#### 4.1.5 Human experiments Analysis

An experimental study was designed and implemented involving human readers to study and analyze the performance of the prose comprehension gained by the Knowledge Distillation Process.

The Kent State University Institutional Review Board (IRB) approved this study. The study was conducted on 36 male and female readers from Kent State University. The readers' backgrounds and education levels varied. They were between the age of 21 and 50 years. They were divided into three groups of 12 readers. Each group was given a test about a single prose among the three proeses presented in Table 4.1. Each test contained 36 questions: 8 questions were “*what is*” questions asking about the meaning of the concepts; 28 questions were “*what is the relation between concept  $c_i$  and concept  $c_j$* ” questions that asked about the relations between two concepts. Each reader answered the same set of questions 3 times in 3 attempts. In the first attempt (attempt1), the reader answered the questions based on his or her prior knowledge. In the second attempt, participants read a short prose about a specific topic. Then, they answered the same questions again based on prior knowledge and the information read in the prose. In the third attempt, participants read the *enhanced text* about all/some of the concepts in the prose and their relations. The *enhanced text* was derived from the Skimmed Illuminated Knowledge Graph *SIKG*. Afterwards, the reader answered the 36 questions, but this time based on prior knowledge, the prose read in the second attempt (attempt2), and the *enhanced text* in the third attempt (attempt3).



**Table 4.9. Rubric for incremental enhancement for knowledge comprehension**

1.	$Y_i =$	The question is answered correctly in the previous attempt. No new information is added in the current attempt.
2.	$Y_i +$	The question is answered correctly in the previous attempt. New correct information is added in the current attempt. The answer has further improved.
3.	$Y_i -$	The question is answered correctly in the previous attempt. New incorrect information is added in the current attempt. The answer is slightly worse but still correct.
4.	$N_i -/+$	The question is answered incorrectly in the previous attempt. New correct information is added in current attempt and the answer now is correct.
5.	$N_i =$	The question is answered incorrectly in the previous attempt. No new information is added in the current attempt.
6.	$N_i +$	The question is answered incorrectly in previous attempt. New correct information is added in current attempt. The answer is slightly better but still incorrect.
7.	$N_i -$	The question is answered incorrectly in the previous attempt. New incorrect information is added in the current attempt. The answer is still incorrect and has further degraded.
8.	$Y_i +/-$	The question is answered correctly in the previous attempt. New incorrect information is added in the current attempt and the answer now is incorrect.

Presenting the *enhanced text* can be done in many ways and by itself can be considered a problem worthy of being addressed (Khan & Hardas, 2013). An ‘pre-order’ traverse was applied in the Skimmed Illuminated Knowledge Graph *SIKG* to present its sentences as the *enhanced text* to the reader.

To assess the impact of the proposed Knowledge Distillation Process for increasing knowledge, we classify the recognized/not recognized concepts and the recognized/not recognized relations in transition-1 (from attempt1 to attempt2) and in transition-2 (from attempt2 to attempt 3) into eight categories each. All eight categories are presented in Table 4.3. For example, *what is carbonization?* is a question about the meaning of a concept from a prose. There were three attempts in two transitions to answer the question. For one of the answer cases: suppose that the reader in attempt1 answered the question with (*I don't know*)

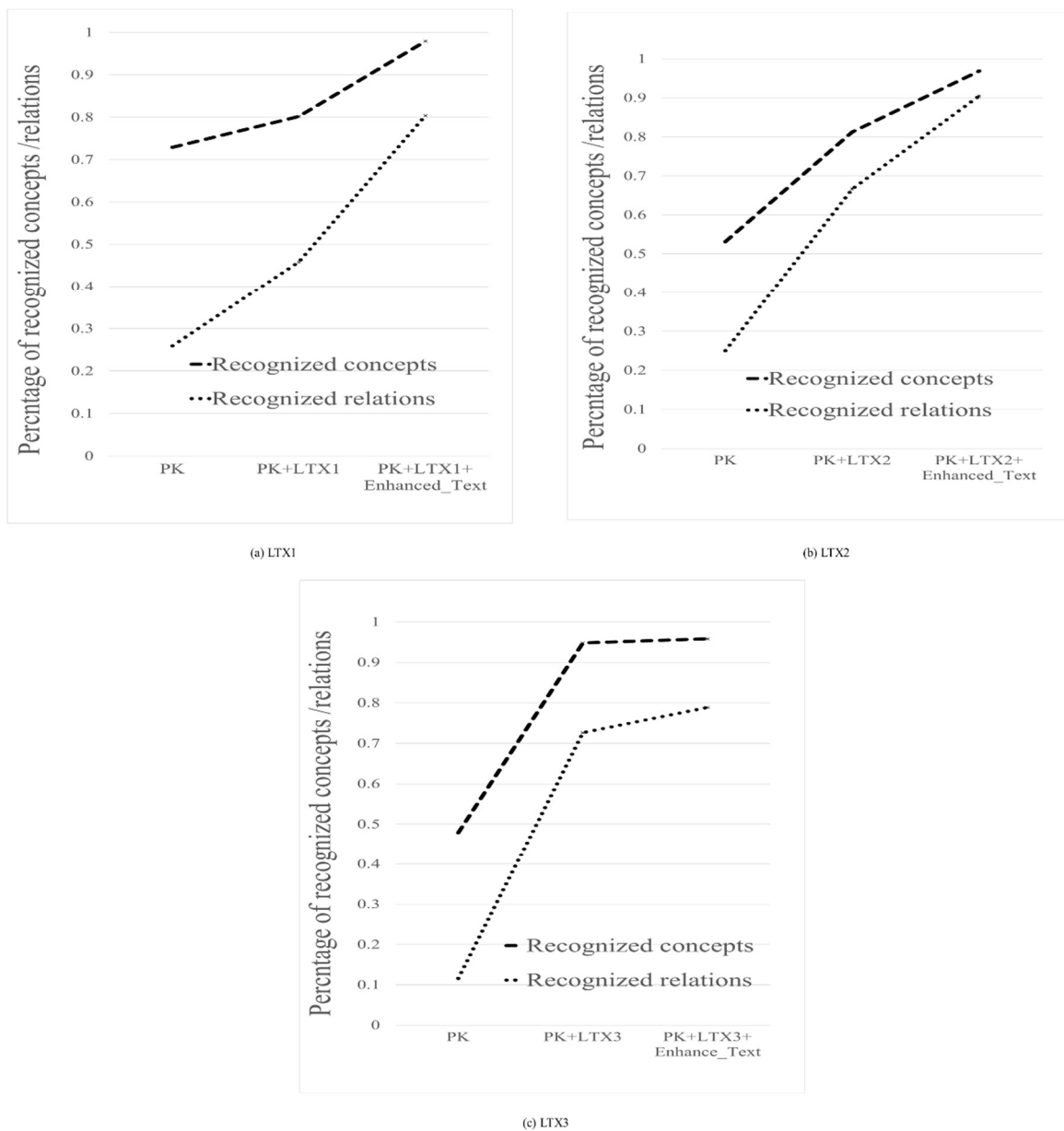
( $N$ ) which means he or she did not have any previous knowledge about the meaning of the concept, then after reading the prose in attempt2, he or she answered the question correctly with (*Carbonization is a chemical process used for processing of coal*) ( $N_{1-/+}$ ) which means that the knowledge in the first transition (from attempt1 to attempt2) had increased. Suppose the reader after reading the *enhanced text* modified his answer with a new correct response (*The conversion of dead vegetation into coal is called carbonization*) ( $Y_{2+}$ ); that meant the knowledge in the second transition (from attempt2 to attempt3) also increased. Another case: suppose the reader in attempt1 answered the question correctly with (*Carbonization is a chemical process used for processing of coal*) ( $Y$ ). This meant he or she had prior knowledge of the concept. Then, in attempt2 he or she answered the question with (*See Attempt1*) ( $Y_{1=}$ ); meaning that no new information was added in the first transition (from attempt1 to attempt2). Then, in attempt3, he or she modified the answer with incorrect information (*Carbonization is a process used for producing oxygen*) ( $Y_{1+/-}$ ); such a response indicates that the knowledge in the second transition (from attempt2 to attempt3) became distorted.

Figure 4.12 shows the average number of recognized concepts and the recognized relations in the three proses. It can be seen that in all cases, the values of the recognized concepts and the recognized relations gradually increased during the transition attempts in the three proses. In other words, more information was recognized in almost every transition. For the recognized concepts, the value increased by 0.1, 0.3, and 0.5 in the transition from attempt1 to attempt2 and by 0.2, 0.2, 0.01 in the transition from attempt2 to attempt3 respectively in the three proses. This indicates that the *enhanced text*

contributed in increasing knowledge about the concepts, thus overall prose comprehension. Similarly, the recognized relation values increased by 0.2, 0.4, and 0.6 in the transition from attempt1 to attempt2, and by 0.3, 0.2, and 0.06 in the transition from attempt2 to attempt3 respectively in the three proses; this verifies the effectiveness of the added information for increasing knowledge and comprehension.

In Figure 4.13, we can see the average of answers within the categories. It can be seen that the amount of new added knowledge ( $Y_{i+} + N_{i-/}$ ) for the recognized concepts increased by 0.2, 0.3, and 0.6 in the transition from attempt1 to attempt2 and increased by 0.6, 0.4, and 0.6 in the transition from attempt2 to attempt3 respectively in the three proses. This points directly to the impact of the Knowledge Distillation Process in adding new knowledge for recognizing the concepts. Likewise, for the recognized relations, they increased by 0.3, 0.4, and 0.6 in the transition from attempt1 to attempt2 and by 0.5, 0.4, and 0.3 in the transition from attempt2 to attempt3 respectively in the three proses; this refers to the effectiveness of the model in adding new knowledge for recognizing relations between concepts within the prose.

The path length between two concepts played an important role for recognizing the relation between them. The distribution of the knowledge paths length to the correct recognized relations answers is presented in Figure 4.14. It is partly clear that the shorter the knowledge path length, the more correct answers there were.



**Figure 4.12.** The average of the recognized concepts and the recognized relations in the three proses.

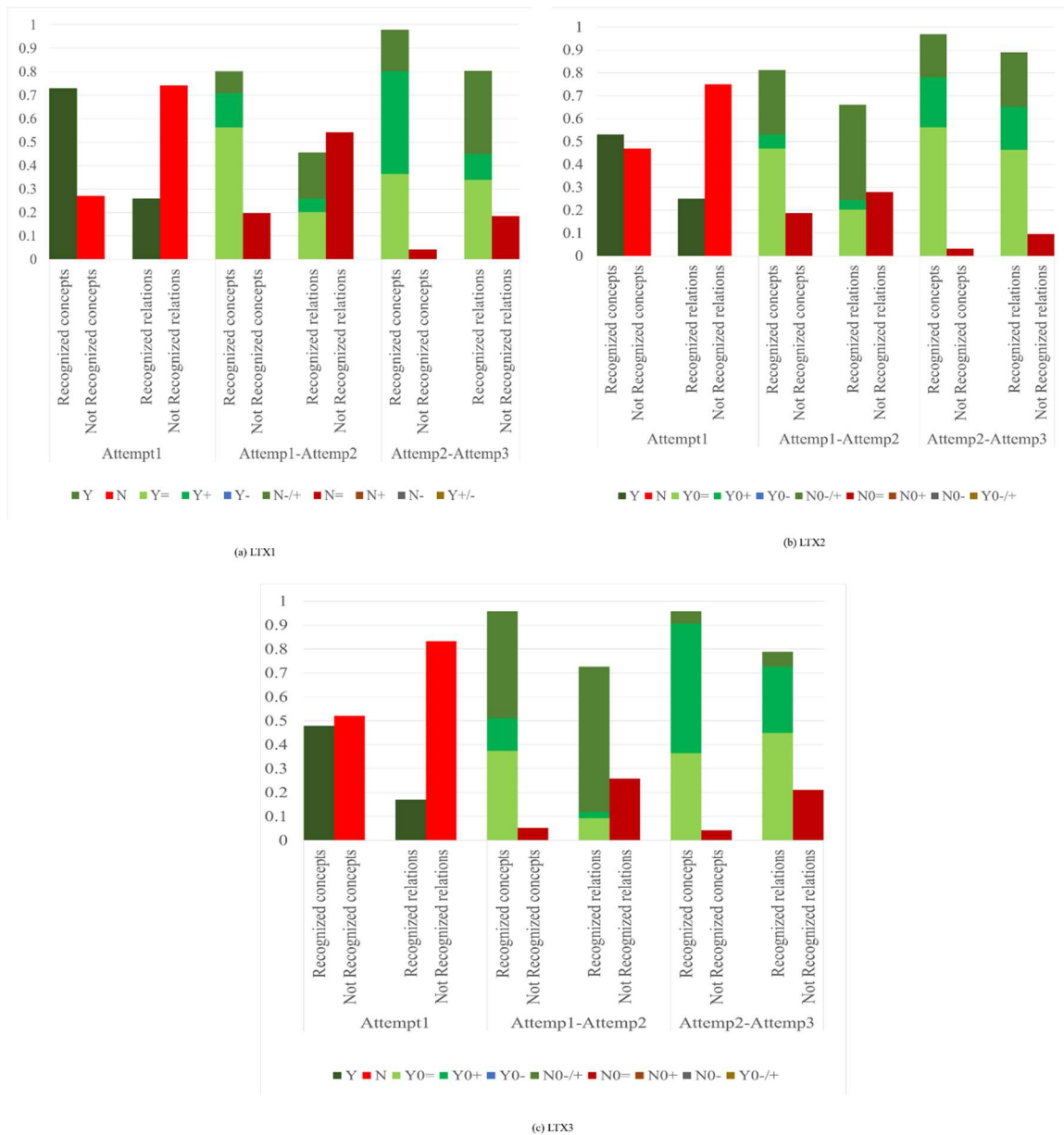
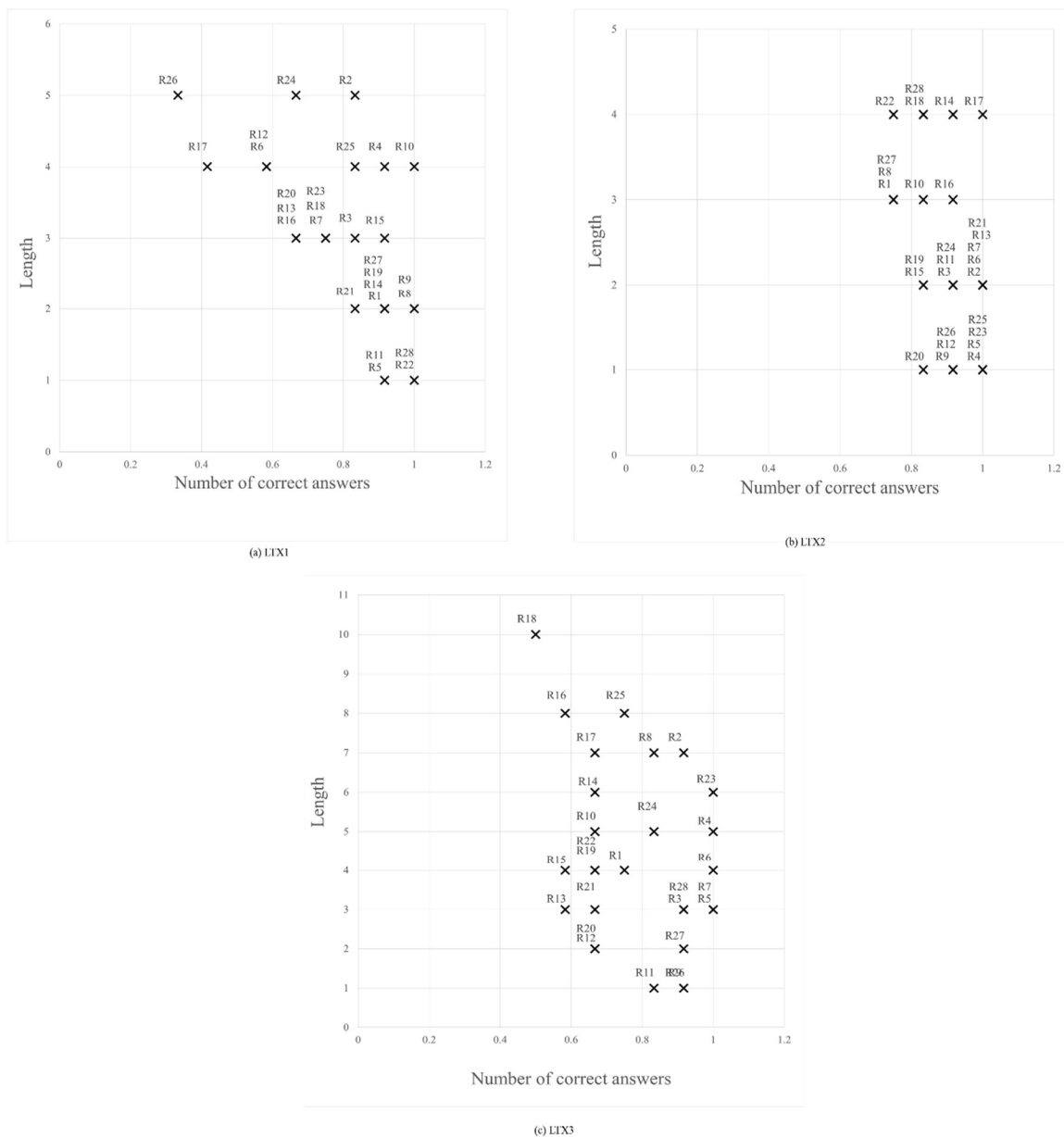


Figure 4.13. The average of the incremental enhancement of the recognized concepts and the recognized relations in the three proses.



**Figure 4.14. The distribution of the knowledge paths length to the correct recognized relations in the three proses.**

## 4.2 Summary

In this chapter, details were presented about the computational evaluation model that were used to evaluate the efficiency of the comprehension gained from the

comprehension engine. The model measured the quantitative estimation, organization estimation, and the comprehension efficiency of the acquired knowledge gained from the comprehension engine. Next, the content materials that were used in the experiment were presented. Then, the results gained from each process in the comprehension engine and its analysis were displayed. Finally, the design and the analysis of the experimental study were shown and explained; the study involved human readers to study the impact of acquired knowledge gained from the Knowledge Distillation Process on prose comprehension.

## CHAPTER 5

### Conclusions

In this chapter, the goals achieved and the future considerations of the presented work are discussed.

#### **5.1 Does the proposed comprehension engine help in improving the quality of comprehension?**

Chapter 4 showed how certain properties of the knowledge graph pointed to improving the quality of the prose comprehension.

1. *Information content:* it is seen in Figure 4.1 and Table 4.6 that the knowledge graphs in both processes begin with multiple disconnected concepts. Then with use of the comprehension engine, all of the concepts become connected, indicating that all the concepts are reachable to each other. This means that the sentential relations between the prose concepts are known due to the comprehension engine; the sentential relations between the prose concepts help in recognizing unknown concepts and in increasing the knowledge about the known ones. Thus, a connected knowledge graph helps in improving the quality of comprehension. In Table 4.4 and Table 4.6, it is also seen that the size of the knowledge graphs started with a low number of concepts and sentential relations. Then with use of the comprehension engine, this number increased. It was assumed that when the size of the knowledge graph was increased, actual knowledge was also increased. This points to the new knowledge that appeared, illuminating the reader to the prose concepts



and their sentential relations. Similarly, the study of “*knowledge growth*” seen Figure 4.2 and Figure 4.3, and the study of “*knowledge overload*” featured in Table 4.6, showed that the comprehension engine contributed to both growth and perhaps an overload of knowledge gained. Accordingly, increasing the size of the knowledge graph indicates to improving the quality of comprehension. In addition, from the findings displayed in Figure 4.4 and Figure 4.5, and the study of “*entropy*” shown in Table 4.6, it is obvious that the knowledge graphs in both processes start with low entropy values. Later, by the comprehension engine, the values increased. With the use of the comprehension engine, the amount of the new rare information increased, referring to an improvement in the quality of the comprehension.

2. *Knowledge organization*: as seen in Figure 4.6, Figure 4.7, and the study of “*cluster coefficient*” shown in Table 4.6, the comprehension engine increased the value of the cluster coefficient in the knowledge graphs. This points to the concepts becoming highly organized and connected, thus facilitating understanding the sentential relations among the prose concepts. Therefore, this indicates improvement of the quality of comprehension.

3. *Comprehension Efficiency*: the study of reachability in Table 4.8 shows how reachability among the prose concepts was reduced by the Knowledge Distillation Process; the number of sentences needed to understand the relation between two of the prose concepts decreased. Therefore, efficiency of the comprehension showed improvement. In addition, Figure 4.11 shows how the illumination values of the prose concepts in the knowledge graphs increased. This refers to the impact of the sentential relations in the

knowledge graphs for increasing knowledge of the prose concepts. For each knowledge graph, the illumination values of the prose concepts reaching stable values denotes how much the reader learned from the prose. Meanwhile, Figure 4.10 shows that the prose concepts were both fully and highly illustrated, indicating the improvement level in the quality of understanding provided by the comprehension engine. This indicates to that the comprehension engine improves the quality of the comprehension.

As seen from Figure 4.12 and Figure 4.13, enhanced text increased recognition of both the concepts and their relationships with each other, thus pointing to an improvement in the quality of the comprehension.

## **5.2 Does the proposed comprehension engine effect in saving time of learning?**

As shown in Table 4.3 and Table 4.5, the time necessary for the comprehension engine to read the reference texts, finding the highest familiarity sentential relations among the prose concepts, and then finding the Alpha Knowledge Pathway between each pair took only a few minutes. This is considered a short amount of time for learning when compared with the time needed for a human to read and decipher multiple reference texts (as was discussed in Chapter 1). This demonstrates the efficiency of the comprehension engine for saving time on reading.

## **5.3 Criticize and Challenges**

An interesting debate is whether or not all of the prose concepts are connected in the knowledge graph. It is probable that the knowledge graph may include disconnected concepts. In the comprehension engine, it is completely possible that an association relation

between two concepts may be recognized without a link that exists between them in the knowledge graph.

#### **5.4 Future work**

1. The association weight (familiarity value) and the concept illumination values were calculated based on the frequency of the relation type/concept in the Gutenberg Project. However, the ‘Gutenberg Project’ has some limitations. Its history was created by books. The community of the words was formed by book authors and experts, and word frequency here is based on their frequency in the books. Thus, we found the corpus missed a lot of common words. As our goal is to enhance comprehension for common readers, we suggest using another more common corpus such as a buildup corpus by Wikipedia, a newspaper, or tweets for a future work. For example, Wikipedia articles are written by anyone who cares about a topic, and its writers could be experts or semi-experts. So, its corpus words will be common and close to the reader. The more common the words, the more likely the understanding will be better, creating improvement of the quality of the comprehension.

2. In the experiment, we considered the Alpha Knowledge Pathway as a single knowledge path between each pair of prose concepts, where each edge in the path is represented by a single sentence comprehending the relation between each two concepts in the path. For future work, we will include the parallel relations between the concepts in the Alpha Knowledge Pathway in the experiment. This will add multiple sentences instead of a single sentence illuminating the relation between each two concepts in the knowledge

path. Therefore, this will strengthen understanding of the relation between the two concepts and increase the knowledge, thus leading to improving the quality of the comprehension.

3. The method used to find an Alpha Knowledge Pathway is based on grading all the knowledge paths between two of the prose concepts, then selecting the one which has the highest delivered current flow between them. It did not consider decreasing the number of external concepts as the method used for finding the highest familiarity knowledge path connecting the prose concepts that was described in Chapter 2. Learning new external concepts with their relations in order to learn the relations between the prose concepts still exists. Therefore, this will increase the burden on the reader for comprehending the prose. To decrease this burden, an elimination or even a reduction of the number of external concepts is required. So, after augmenting knowledge from the reference texts in the Knowledge Induction Process and forming the final Illuminated Knowledge Graph, we suggest applying the Terminal to Terminal Steiner Tree (TTST) algorithm discussed in Chapter 2 on the final Illuminated Knowledge Graph to find the Alpha Knowledge Pathway. The Alpha Knowledge Pathway would then connect all the prose concepts with few to no external concepts.

## **5.5 Summary**

The main contribution of this dissertation was to study the algorithms behind the prose comprehension. The problem of prose comprehension was explored and to help solve it, we recommended using an initial version of a set of algorithms that can be used to enhance the comprehension by creating a comprehension engine. Such an engine can enhance the prose comprehension by saving time and improving the quality of the

comprehension. The comprehension engine is able to read several reference texts and select the highest familiarity knowledge. It then presents this external knowledge to the readers in a short amount of time. In addition, it uses a graph called a knowledge graph as a computational representation model to represent the knowledge in the text. Some of the measurements on the knowledge graph are indicative of the quality of the learning which in turn enhances comprehension. Furthermore, the results of the human experiment on the output of the comprehension engine verifies the efficiency of the compression engine for improving knowledge.

## APPENDIX A

### Example of the Data used in the experiment

This appendix shows the processes used in the experiment.

LTX1- (Ethane)

Ethane, a colourless, odourless, gaseous hydrocarbon (compound of hydrogen and carbon), belonging to the paraffin series; its chemical formula is  $C_2H_6$ . Ethane is structurally the simplest hydrocarbon that contains a single carbon-carbon bond. The second most important constituent of natural gas, it also occurs dissolved in petroleum oils and as a by-product of oil refinery operations and of the carbonization of coal.

The industrial importance of ethane is based upon the ease with which it may be converted to ethylene ( $C_2H_4$ ) and hydrogen by pyrolysis, or cracking, when passed through hot tubes. Like propane and, to a lesser extent, butane, ethane is a major raw material for the huge ethylene petrochemical industry, which produces such important products as polyethylene plastic, ethylene glycol, and ethyl alcohol.

More than 90 percent of the ethane produced in the 1960s was burned as fuel without separation from natural gas. Ethane gas can be liquefied under pressure or at reduced temperatures and thus be separated from natural gas. Unlike propane, liquid ethane is not in common use as an industrial or domestic fuel.

## LTX2- (New Test for Zika OKed)

Current PCR-based Zika tests can't rule out infections with dengue or chikungunya viruses—infections that cause similar symptoms and are also transmitted by Aedes mosquitoes. Last week (March 18), the US Food and Drug Administration (FDA) granted the US Centers for Disease Control and Prevention (CDC) approval to start using a three-in-one assay that screens for all three viruses simultaneously.

“This [emergency use authorization] will potentially allow CDC to more rapidly perform testing to detect acute Zika virus infection,” the CDC said in a statement.

The test won't be available in hospitals or doctors offices, but will be used in a designated network of laboratories that assists in public health emergencies.

Last month, the FDA granted emergency use authorization (EUA) for these labs to use a diagnostic that can detect infections through antibodies in the patient's blood weeks after the virus has been cleared. However, the test cannot rule out the possibility that a positive result was caused by dengue, and research labs and biotech firms are working to develop a more-specific antibody assay.

“As there are no commercially available diagnostic tests cleared or approved by the FDA for the detection of Zika virus infection, it was determined that an EUA is crucial to ensure timely access to a diagnostic tool,” the CDC said in a February 26 press release.

## LTX3- (Anesthesia gases are warming the planet)

Anesthetics may make that tooth surgery bearable, but they are also contributing—at least somewhat—to climate change, a new study reveals. The gases act in much the same way as carbon dioxide (CO<sub>2</sub>), trapping energy from the sun in the atmosphere and warming the planet. Over the past decade, atmospheric concentrations of the commonly used anesthetics desflurane, isoflurane, and sevoflurane have risen globally to 0.30 parts per trillion (ppt), 0.097 ppt, and 0.13 ppt, respectively, scientists report in *Geophysical Research Letters*. Although those numbers may not seem like much—especially compared with CO<sub>2</sub>, which reached concentrations of 400 parts per million in 2014—the higher global warming potential of the anesthetics has some scientists worried. For example, every 1 kilogram of desflurane is equal to 2500 kilograms of CO<sub>2</sub>. They also tracked concentrations of another anesthetic, halothane—which many countries have phased out because it can damage the liver—and found its concentration had declined since 2000. Although nitrous oxide is also widely used as an anesthetic, the researchers purposefully did not include it in the study because, unlike the other gases, it is used in a variety of settings other than anesthetics, such as the food industry and in semiconductor manufacturing. Although no one is suggesting a return to the days of biting on a piece of leather or wood to distract from the pain of surgery, scientists argue that limiting or even eliminating the use of desflurane, the most potent of the three gases studied, would help. Also, the study's researchers point out, no mandate exists that requires used anesthetic be captured and disposed of, and as a result, almost all of it is released directly into the atmosphere.



**ATTEMPT1****[Ethane, hydrocarbon, hydrogen, carbon, chemical, petroleum, carbonization, coal]****Questions:****Please answer the following questions in few words:**

1. What is ethane?

 I don't know now. Answer -----

2. What is hydrocarbon?

 I don't know now. Answer -----

3. What is hydrogen?

 I don't know now. Answer -----

4. What is carbon?

 I don't know now. Answer -----

5. What is chemical?

 I don't know now. Answer -----

6. What is petroleum?

I don't know now.

Answer -----

7. What is carbonization?

I don't know now.

Answer -----

8. What is coal?

I don't know now.

Answer -----

9. What is the relation between ethane and hydrocarbon?

I don't know now.

Answer -----

10. What is the relation between ethane and hydrogen?

I don't know now.

Answer -----

11. What is the relation between ethane and carbon?

I don't know now.

Answer -----

12. What is the relation between ethane and chemical?

I don't know now.

Answer -----

13. What is the relation between ethane and petroleum?

I don't know now.

Answer -----

14. What is the relation between ethane and carbonization?

I don't know now.

Answer -----

15. What is the relation between ethane and coal?

I don't know now.

Answer -----

16. What is the relation between hydrocarbon and hydrogen?

I don't know now.

Answer -----

17. What is the relation between hydrocarbon and carbon?

I don't know now.

Answer -----

18. What is the relation between hydrocarbon and chemical?

I don't know now.

Answer -----

19. What is the relation between hydrocarbon and petroleum?

I don't know now.

Answer -----

20. What is the relation between hydrocarbon and carbonization?

I don't know now.

Answer -----

21. What is the relation between hydrocarbon and coal?

I don't know now.

Answer -----

22. What is the relation between hydrogen and carbon?

I don't know now.

Answer -----

23. What is the relation between hydrogen and chemical?

I don't know now.

Answer -----

24. What is the relation between hydrogen and petroleum?

I don't know now.

Answer -----

25. What is the relation between hydrogen and carbonization?

I don't know now.

Answer -----

26. What is the relation between hydrogen and coal?

I don't know now.

Answer -----

27. What is the relation between carbon and chemical?

I don't know now.

Answer -----

28. What is the relation between carbon and petroleum?

I don't know now.

Answer -----

29. What is the relation between carbon and carbonization?

I don't know now.

Answer -----

30. What is the relation between carbon and coal?

I don't know now.

Answer -----

31. What is the relation between chemical and petroleum?

I don't know now.

Answer -----

32. What is the relation between chemical and carbonization?

I don't know now.

Answer -----

33. What is the relation between chemical and coal?

I don't know now.

Answer -----

34. What is the relation between petroleum and carbonization?

I don't know now.

Answer -----

35. What is the relation between petroleum and coal?

I don't know now.

Answer -----

36. What is the relation between carbonization and coal?

I don't know now.

Answer -----

## ATTEMPT2

### [Ethane, hydrocarbon, hydrogen, carbon, chemical, petroleum, carbonization, coal]

#### Text1:

[*Ethane*], a colourless, odourless, gaseous [*hydrocarbon*] (compound of [*hydrogen*] and [*carbon*]), belonging to the paraffin series; its [*chemical*] formula is  $C_2H_6$ . [*Ethane*] is structurally the simplest [*hydrocarbon*] that contains a single carbon-carbon bond. The second most important constituent of natural [*gas*], it also occurs dissolved in [*petroleum*] oils and as a by-product of oil refinery operations and of the [*carbonization*] of [*coal*].

The industrial importance of [*ethane*] is based upon the ease with which it may be converted to ethylene ( $C_2H_4$ ) and [*hydrogen*] by pyrolysis, or cracking, when passed through hot tubes. Like propane and, to a lesser extent, butane, [*ethane*] is a major raw material for the huge ethylene petrochemical industry, which produces such important products as polyethylene plastic, ethylene glycol, and ethyl alcohol.

More than 90 percent of the [*ethane*] produced in the 1960s was burned as fuel without separation from natural [*gas*]. [*Ethane*] [*gas*] can be liquefied under pressure or at reduced temperatures and thus be separated from natural [*gas*]. Unlike propane, liquid [*ethane*] is not in common use as an industrial or domestic fuel.

#### Questions:

**Please answer the following questions in few words:**

1. What is ethane?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

2. What is hydrocarbon?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

3. What is hydrogen?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

4. What is carbon?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

5. What is chemical?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

6. What is petroleum?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

7. What is carbonization?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

8. What is coal?



- See Attempt 1.
- I don't know now
- Answer/ Modified Answer. -----

9. What is the relation between ethane and hydrocarbon?

- See Attempt 1.
- I don't know now
- Answer/ Modified Answer. -----

10. What is the relation between ethane and hydrogen?

- See Attempt 1.
- I don't know now
- Answer/ Modified Answer. -----

11. What is the relation between ethane and carbon?

- See Attempt 1.
- I don't know now
- Answer/ Modified Answer. -----

12. What is the relation between ethane and chemical?

- See Attempt 1.
- I don't know now
- Answer/ Modified Answer. -----

13. What is the relation between ethane and petroleum?

- See Attempt 1.
- I don't know now
- Answer/ Modified Answer. -----

14. What is the relation between ethane and carbonization?

- See Attempt 1.
- I don't know now
- Answer/ Modified Answer. -----

15. What is the relation between ethane and coal?

- See Attempt 1.
- I don't know now
- Answer/ Modified Answer. -----

16. What is the relation between hydrocarbon and hydrogen?

- See Attempt 1.
- I don't know now
- Answer/ Modified Answer. -----

17. What is the relation between hydrocarbon and carbon?

- See Attempt 1.
- I don't know now
- Answer/ Modified Answer. -----

18. What is the relation between hydrocarbon and chemical?

- See Attemp1.
- I don't know now
- Answer/ Modified Answer. -----

19. What is the relation between hydrocarbon and petroleum?

- See Attemp1.
- I don't know now
- Answer/ Modified Answer. -----

20. What is the relation between hydrocarbon and carbonization?

- See Attemp1.
- I don't know now
- Answer/ Modified Answer. -----

21. What is the relation between hydrocarbon and coal?

- See Attemp1.
- I don't know now
- Answer/ Modified Answer. -----

22. What is the relation between hydrogen and carbon?

- See Attemp1.
- I don't know now
- Answer/ Modified Answer. -----

23. What is the relation between hydrogen and chemical?

- See Attemp1.
- I don't know now
- Answer/ Modified Answer. -----

24. What is the relation between hydrogen and petroleum?

- See Attemp1.
- I don't know now
- Answer/ Modified Answer. -----

25. What is the relation between hydrogen and carbonization?

- See Attemp1.
- I don't know now
- Answer/ Modified Answer. -----

26. What is the relation between hydrogen and coal?

- See Attemp1.
- I don't know now
- Answer/ Modified Answer. -----

27. What is the relation between carbon and chemical?

- See Attemp1.
- I don't know now

Answer/ Modified Answer. -----

28. What is the relation between carbon and petroleum?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

29. What is the relation between carbon and carbonization?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

30. What is the relation between carbon and coal?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

31. What is the relation between chemical and petroleum?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

32. What is the relation between chemical and carbonization?

See Attempt 1.

I don't know now

Answer/ Modified Answer. -----

33. What is the relation between chemical and coal?

See Attemp1.

I don't know now

Answer/ Modified Answer. -----

34. What is the relation between petroleum and carbonization?

See Attemp1.

I don't know now

Answer/ Modified Answer. -----

35. What is the relation between petroleum and coal?

See Attemp1.

I don't know now

Answer/ Modified Answer. -----

36. What is the relation between carbonization and coal?

See Attemp1.

I don't know now

Answer/ Modified Answer. -----

### ATTEMPT3

#### [Ethane, hydrocarbon, hydrogen, carbon, chemical, petroleum, carbonization, coal]

##### Text2:

- [*Ethane*] can also be separated from [*petroleum*] gas, a mixture of gaseous [*hydrocarbon*] produced as a byproduct of [*petroleum*] refining.
- The [*chemical*] structure of [*petroleum*] is heterogeneous, composed of [*hydrocarbon*] chain of different length.
- Oil is a synonym of [*petroleum*].
- in the 19th century, the term [*petroleum*] was often used to refer to mineral oil produced by distillation from mined organic solid such as cannel [*coal*] (and later oil shale), and refined oil produced from them; in the united kingdom, storage (and later transport) of these oil were regulated by a series of [*petroleum*] act, from the [*petroleum*] act 1863 onwards.
- Because of this, [*petroleum*] may be taken to oil refinery and the [*hydrocarbon*] [*chemical*] separated by distillation and treated by other [*chemical*] process, to be used for a variety of purpose.
- [*Hydrogen*], as atomic h, is the most abundant [*chemical*] element in the universe, making up 75 % of normal matter by mass and more than 90% by number of atom.
- [*Hydrogen*] form a vast array of compound with [*carbon*] called the [*hydrocarbon*], and an even vaster array with heteroatoms that, because of their general association with living thing, are called organic compound.
- [*Petroleum*] is a kind of fossil fuel.
- Fossil fuel is a superordinate of [*coal*].
- As [*coal*] contains mainly [*carbon*], the conversion of dead vegetation into [*coal*] is called [*carbonization*].
- Fossil fuel is a superordinate of [*petroleum*].
- Generally, with catenation come the loss of the total amount of bonded [*hydrocarbon*] and an increase in the amount of energy required for bond cleavage due to strain exerted upon the molecule ; in molecule such as cyclohexane , this is referred to as ring strain , and occurs due to the ``destabilized " spatial electron configuration of the atom.
- Because of difference in molecular structure, the empirical formula remains different between [*hydrocarbon*]; in linear, or `` straight-run " alkane, alkene and alkyne, the amount of bonded [*hydrogen*] lessens in alkene and alkyne due to the `` self-bonding " or catenation of [*carbon*] preventing entire saturation of the [*hydrocarbon*] by the formation of double or triple bond.
- [*Hydrocarbon*] should be kept away from fluorine compound due to the high probability of forming toxic hydrofluoric acid.
- [*Hydrocarbon*] can also be burned with elemental fluorine, resulting in [*carbon*] tetrafluoride and [*hydrogen*] fluoride product.

- Venezuela also has large amount of oil in the orinoco oil sand, although the [hydrocarbon] trapped in them are more fluid than in canada and are usually called extra heavy oil.
- [Coal] is composed primarily of [carbon], along with variable quantity of other element, chiefly [hydrogen], sulfur, oxygen, and nitrogen.
- The simplest form of an organic molecule is the [hydrocarbon] a large family of organic molecule that are composed of [hydrogen] atom bonded to a chain of [carbon] atom.
- The buckyball are fairly large molecule formed completely of [carbon] bonded trigonally, forming spheroid (the best-known and simplest is the soccerball-shaped c60 buckminsterfullerene).
- Alternatively, the [hydrogen] obtained from gasification can be used for various purpose, such as powering a [hydrogen] economy, making ammonia, or upgrading fossil fuel.
- [Coal] gasification can be used to produce syngas, a mixture of [carbon] monoxide (co) and [hydrogen] (h2) gas.
- Primary [chemical] that are produced directly from the syngas include methanol, [hydrogen] and [carbon] monoxide, which are the [chemical] building block from which a whole spectrum of derivative [chemical] are manufactured , including olefin , acetic acid , formaldehyde , ammonia , urea and others.
- Compound based primarily on [carbon] and [hydrogen] atom are called organic compound, and all others are called inorganic compound.
- Compound is a kind of [chemical].
- [Chemical] is a superordinate of compound.
- Pre-combustion capture - this involves gasification of a feedstock ( such as coal ) to form synthesis gas, which may be shifted to produce a h2 and co2-rich gas mixture , from which the co2 can be efficiently captured and separated , transported , and ultimately sequestered , this technology is usually associated with integrated gasification combined cycle process configuration.
- Several different technological method are available for the purpose of [carbon] capture as demanded by the clean coal concept.

### Questions:

**Please answer the following questions in few words:**

1. What is ethane?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----



2. What is hydrocarbon?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

3. What is hydrogen?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

4. What is carbon?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

5. What is chemical?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

6. What is petroleum?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

7. What is carbonization?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

8. What is coal?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

9. What is the relation between ethane and hydrocarbon?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

10. What is the relation between ethane and hydrogen?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

11. What is the relation between ethane and carbon?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

12. What is the relation between ethane and chemical?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

13. What is the relation between ethane and petroleum?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

14. What is the relation between ethane and carbonization?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

15. What is the relation between ethane and coal?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

16. What is the relation between hydrocarbon and hydrogen?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

17. What is the relation between hydrocarbon and carbon?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

18. What is the relation between hydrocarbon and chemical?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

19. What is the relation between hydrocarbon and petroleum?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

20. What is the relation between hydrocarbon and carbonization?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

21. What is the relation between hydrocarbon and coal?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

22. What is the relation between hydrogen and carbon?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

23. What is the relation between hydrogen and chemical?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

24. What is the relation between hydrogen and petroleum?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

25. What is the relation between hydrogen and carbonization?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

26. What is the relation between hydrogen and coal?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

27. What is the relation between carbon and chemical?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

28. What is the relation between carbon and petroleum?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

29. What is the relation between carbon and carbonization?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

30. What is the relation between carbon and coal?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

31. What is the relation between chemical and petroleum?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

32. What is the relation between chemical and carbonization?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

33. What is the relation between chemical and coal?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

34. What is the relation between petroleum and carbonization?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

35. What is the relation between petroleum and coal?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----

36. What is the relation between carbonization and coal?

See Attemp1/ Attemp2.

I don't know

Answer/ Modified Answer. -----



## REFERENCES

- Aedes. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Aedes&oldid=797888558>
- Agrawal, A., & An, A. (2012). Unsupervised emotion detection from text using semantic and syntactic relations. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01* (pp. 346–353). IEEE Computer Society. Retrieved from <http://dl.acm.org/citation.cfm?id=2457613>
- Al Madi, N. S. (2014). *A Study of Learning Performance and Cognitive Activity During Multimodal Comprehension Using Segmentation-integration Model and EEG*. Kent State University. Retrieved from [http://rave.ohiolink.edu/etdc/view?acc\\_num=kent1416868268](http://rave.ohiolink.edu/etdc/view?acc_num=kent1416868268)
- Al Madi, N. S., & Khan, J. I. (2015). Is learning by reading a book better than watching a movie? a computational analysis of semantic concept network growth during text and multimedia comprehension. In *Neural Networks (IJCNN), 2015 International Joint Conference on* (pp. 1–8). IEEE. Retrieved from <http://ieeexplore.ieee.org/abstract/document/7280761/>
- Anesthetic. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Anesthetic&oldid=796835414>
- Antibody. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Antibody&oldid=797925412>

- Babour, A., Khan, J. I., & Nafa, F. (2016). Deepening Prose Comprehension by Incremental Free Text Conceptual Graph Mining and Knowledge. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2016 International Conference on* (pp. 208–215). IEEE. Retrieved from <http://ieeexplore.ieee.org/abstract/document/7864234/>
- Babour, A., Nafa, F., & Khan, J. I. (2015). Connecting the Dots in a Concept Space by Iterative Reading of Freetext References with Wordnet. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 441–444). IEEE. Retrieved from <http://ieeexplore.ieee.org/abstract/document/7396845/>
- Baker, L. (1979). Comprehension monitoring: Identifying and coping with text confusions. *Journal of Reading Behavior, 11*(4), 365–374.
- Ball, S. J. (1998). Educational Studies Policy Entrepreneurship and Social Theory. In *School Effectiveness for Whom?* Psychology Press.
- Bergenholtz, H., & Gouws, R. (2010). A new perspective on the access process. *Hermes, 44*, 103–127.
- Carbon. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Carbon&oldid=797606653>
- Carbonization. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Carbonization&oldid=795078223>
- Carver, R. P. (1984). Rauding theory predictions of amount comprehended under different purposes and speed reading conditions. *Reading Research Quarterly, 205–218*.

- Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, 36(2), 84–95.
- Chandran, D., Crockett, K., Mclean, D., & Bandar, Z. (2013). FAST: A fuzzy semantic sentence similarity measure. In *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on* (pp. 1–8). IEEE. Retrieved from <http://ieeexplore.ieee.org/abstract/document/6622344/>
- Chemical substance. (2016). In *Wikipedia*. Retrieved from [https://en.wikipedia.org/w/index.php?title=Chemical\\_substance&oldid=797496319](https://en.wikipedia.org/w/index.php?title=Chemical_substance&oldid=797496319)
- Chen, J.-M., Chen, M.-C., & Sun, Y. S. (2010). A novel approach for enhancing student reading comprehension and assisting teacher assessment of literacy. *Computers & Education*, 55(3), 1367–1382. <https://doi.org/10.1016/j.compedu.2010.06.011>
- Chikungunya. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Chikungunya&oldid=797661262>
- Climate. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Climate&oldid=797528936>
- Coal. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Coal&oldid=797262457>
- Coleman, T., & Moré, J. (1983). Estimation of Sparse Jacobian Matrices and Graph Coloring Blems. *SIAM Journal on Numerical Analysis*, 20(1), 187–209. <https://doi.org/10.1137/0720013>

- Conde, A., Larrañaga, M., Arruarte, A., Elorriaga, J. A., & Roth, D. (2016). litewi: A combined term extraction and entity linking method for eliciting educational ontologies from textbooks. *Journal of the Association for Information Science and Technology*, 67(2), 380–399.
- Davey, B. (1983). Think aloud: Modeling the cognitive processes of reading comprehension. *Journal of Reading*, 27(1), 44–47.
- DeMarco, E. (2015, April 7). Anesthesia gases are warming the planet. Retrieved July 24, 2017, from <http://www.sciencemag.org/news/2015/04/anesthesia-gases-are-warming-planet>
- Dengue fever. (2016). In *Wikipedia*. Retrieved from [https://en.wikipedia.org/w/index.php?title=Dengue\\_fever&oldid=798056254](https://en.wikipedia.org/w/index.php?title=Dengue_fever&oldid=798056254)
- Desflurane. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Desflurane&oldid=797545466>
- Di Vesta, F. J., Hayward, K. G., & Orlando, V. P. (1979). Developmental trends in monitoring text for comprehension. *Child Development*, 97–105.
- Doyle, P. G., & Snell, J. L. (1984). *Random walks and electric networks* (Vol. 22). Mathematical Association of America. Retrieved from <https://www.cse.buffalo.edu/~hungngo/classes/2005/Expanders/papers/general/randomWalk.pdf>
- Draper, S. W., Brown, M. I., Henderson, F. P., & McAteer, E. (1996). Integrative evaluation: an emerging role for classroom studies of CAL. *Computers & Education*, 26(1), 17–32. [https://doi.org/10.1016/0360-1315\(95\)00068-2](https://doi.org/10.1016/0360-1315(95)00068-2)

- Drieger, P. (2013). Semantic Network Analysis as a Method for Visual Text Analytics. *Procedia - Social and Behavioral Sciences*, 79, 4–17. <https://doi.org/10.1016/j.sbspro.2013.05.053>
- ethane. (2013). Retrieved July 24, 2017, from <https://www.britannica.com/science/ethane>
- Ethane. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Ethane&oldid=784178139>
- Faloutsos, C., McCurley, K. S., & Tomkins, A. (2004). Fast discovery of connection subgraphs. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 118–127). ACM. Retrieved from <http://dl.acm.org/citation.cfm?id=1014068>
- Fang, L., Sarma, A. D., Yu, C., & Bohannon, P. (2011). Rex: explaining relationships between entity pairs. *Proceedings of the VLDB Endowment*, 5(3), 241–252.
- Fathi, B. (2014). Experts and Specialised Lexicography: perspectives and needs. *Terminàlia*, 12–21.
- Gagné, E. D., & Memory, D. (1978). Instructional events and comprehension: Generalization across passages. *Journal of Reading Behavior*, 10(4), 321–335.
- Grens, K. (2016). New Test for Zika OKed. Retrieved July 24, 2017, from <http://www.the-scientist.com/?articles.view/articleNo/45638/title/New-Test-for-Zika-OKed/>
- Halothane. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Halothane&oldid=797545700>
- Hao, S. (2016). *Effects of faded scaffolding in computer-based instruction on learners' performance, cognitive load, and test anxiety*. The Florida State University.

Retrieved from  
<http://search.proquest.com/openview/fa48d8a7bb6bafcfccf8582ad6476999/1?pq-origsite=gscholar&cbl=18750&diss=y>

Hardas, M. S. (2012). *Segmentation and Integration in Text Comprehension: A Model of Concept Network Growth*. Kent State University. Retrieved from  
[http://rave.ohiolink.edu/etdc/view?acc\\_num=kent1334593269](http://rave.ohiolink.edu/etdc/view?acc_num=kent1334593269)

Hart, M. S. (1971). Project Gutenberg. Retrieved July 21, 2017, from  
<http://www.gutenberg.org/>

Huey, E. B. (1908). *The psychology and pedagogy of reading*. The Macmillan Company. Retrieved from  
[https://books.google.com/books?hl=en&lr=&id=bqFLMdfbNrsC&oi=fnd&pg=PA1&dq=The+psychology+and+pedagogy+of+reading.&ots=IqyI0535AM&sig=jeS\\_MYRg9oG-zrnKKvV1\\_eKyLmw](https://books.google.com/books?hl=en&lr=&id=bqFLMdfbNrsC&oi=fnd&pg=PA1&dq=The+psychology+and+pedagogy+of+reading.&ots=IqyI0535AM&sig=jeS_MYRg9oG-zrnKKvV1_eKyLmw)

Hydrocarbon. (2016). In *Wikipedia*. Retrieved from  
<https://en.wikipedia.org/w/index.php?title=Hydrocarbon&oldid=797145615>

Hydrogen. (2016). In *Wikipedia*. Retrieved from  
<https://en.wikipedia.org/w/index.php?title=Hydrogen&oldid=794824238>

Infection. (2016). In *Wikipedia*. Retrieved from  
<https://en.wikipedia.org/w/index.php?title=Infection&oldid=797795424>

Isoflurane. (2016). In *Wikipedia*. Retrieved from  
<https://en.wikipedia.org/w/index.php?title=Isoflurane&oldid=797545327>

- Johnson, M. (1980). *Toward Adolescence: The Middle School Years*. University of Chicago Press.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329.
- Kaluarachchi, A., Roychoudhury, D., Varde, A. S., & Weikum, G. (2011). SITAC: discovering semantically identical temporally altering concepts in text archives. In *Proceedings of the 14th International Conference on Extending Database Technology* (pp. 566–569). ACM. Retrieved from <http://dl.acm.org/citation.cfm?id=1951442>
- Kamps, J., Marx, M., Mokken, R. J., De Rijke, M., & others. (2004). Using WordNet to Measure Semantic Orientations of Adjectives. In *LREC* (Vol. 4, pp. 1115–1118). Citeseer. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.409.2489&rep=rep1&type=pdf>
- Kankanhalli, A., & Tan, B. C. Y. (2008). Knowledge Management Metrics: A Review and Directions for Future Research. *Http://Services.igi-Global.com/Resolvedoi/resolve.aspx?doi=10.4018/978-1-59904-933-5.ch282*, 3409–3420. <https://doi.org/10.4018/978-1-59904-933-5.ch282>
- Kasneci, G., Elbassuoni, S., & Weikum, G. (2009). Ming: mining informative entity relationship subgraphs. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1653–1656). ACM. Retrieved from <http://dl.acm.org/citation.cfm?id=1646196>

- Kaster, A., Siersdorfer, S., & Weikum, G. (2005). Combining text and linguistic document representations for authorship attribution. In *SIGIR workshop: stylistic analysis of text for information access*. Retrieved from [https://domino.mpi-inf.mpg.de/intranet/ag5/ag5publ.nsf/989344adb49e8b15c12565f50045087d/4800b599043bbf98c12570ad0039177b/\\$file/style05kaster.pdf](https://domino.mpi-inf.mpg.de/intranet/ag5/ag5publ.nsf/989344adb49e8b15c12565f50045087d/4800b599043bbf98c12570ad0039177b/$file/style05kaster.pdf)
- Khan, J. I., & Hardas, M. S. (2013). Does sequence of presentation matter in reading comprehension? a model based analysis of semantic concept network growth during reading. In *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on* (pp. 444–452). IEEE. Retrieved from <http://ieeexplore.ieee.org/abstract/document/6693560/>
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163.
- Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*(4), 294.
- Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. *Theoretical Models and Processes of Reading*, *5*, 1270–1328.
- Koren, Y., North, S. C., & Volinsky, C. (2006). Measuring and extracting proximity in networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 245–255). ACM. Retrieved from <http://dl.acm.org/citation.cfm?id=1150432>



- Kou, L., Markowsky, G., & Berman, L. (1981). A fast algorithm for Steiner trees. *Acta Informatica*, 15(2), 141–145.
- Levin, J. R. (1973). Inducing comprehension in poor readers: A test of a recent model. *Journal of Educational Psychology*, 65(1), 19.
- Maria, K., & MacGinitie, W. H. (1980). *Prior knowledge as a handicapping condition*. Research Institute for the Study of Learning Disabilities, Teachers College, Columbia University.
- Menaka, S., & Radha, N. (2013). Text classification using keyword extraction technique. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(12). Retrieved from <https://pdfs.semanticscholar.org/0efa/2c915196f1c9b6bd8854ada0ce381460c974.pdf>
- Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81.
- Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11), 39–41. <https://doi.org/10.1145/219717.219748>
- Minor, E. S., & Urban, D. L. (2008). A Graph-Theory Framework for Evaluating Landscape Connectivity and Conservation Planning. *Conservation Biology*, 22(2), 297–307. <https://doi.org/10.1111/j.1523-1739.2007.00871.x>
- Mischel, W. (1979). On the interface of cognition and personality: Beyond the person–situation debate. *American Psychologist*, 34(9), 740.

- Mohamed A.F. Ragab, & Amr Arisha. (2013). Knowledge management and measurement: a critical review. *Journal of Knowledge Management*, 17(6), 873–901.  
<https://doi.org/10.1108/JKM-12-2012-0381>
- Moravcsik, J. E., & Kintsch, W. (1993). Writing quality, reading skills, and domain knowledge as factors in text comprehension. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 47(2), 360.
- Moseley, D. (2005). *Frameworks for thinking: A handbook for teaching and learning*. Cambridge University Press. Retrieved from  
<https://books.google.com/books?hl=en&lr=&id=s1D2IXoNZjwC&oi=fnd&pg=PP1&dq=Frameworks+for+Thinking:+A+Handbook+for+Teaching+and+Learning&ots=9qYX5za-J-&sig=nuh74hM3s-zuEmyxAUWg3uO0NpA>
- Mosquito. (2016). In *Wikipedia*. Retrieved from  
<https://en.wikipedia.org/w/index.php?title=Mosquito&oldid=797321614>
- Owings, R. A., Petersen, G. A., Bransford, J. D., Morris, C. D., & Stein, B. S. (1980). Spontaneous monitoring and regulation of learning: A comparison of successful and less successful fifth graders. *Journal of Educational Psychology*, 72(2), 250.
- Oxide. (2016). In *Wikipedia*. Retrieved from  
<https://en.wikipedia.org/w/index.php?title=Oxide&oldid=789414027>
- Paris, S. G., & Myers, M. (1981). Comprehension monitoring, memory, and study strategies of good and poor readers. *Journal of Reading Behavior*, 13(1), 5–22.
- Patzer, G. L. (2006). *The power and paradox of physical attractiveness*. Universal-Publishers. Retrieved from

[https://books.google.com/books?hl=en&lr=&id=qQXE\\_dL1JNUC&oi=fnd&pg=PR11&dq=The+Power+and+Paradox+of+Physical+Attractiveness&ots=PjVfhGcw4p&sig=62ouk7\\_pzE\\_zQGjSmONzv3CQR4Y](https://books.google.com/books?hl=en&lr=&id=qQXE_dL1JNUC&oi=fnd&pg=PR11&dq=The+Power+and+Paradox+of+Physical+Attractiveness&ots=PjVfhGcw4p&sig=62ouk7_pzE_zQGjSmONzv3CQR4Y)

Patzer, G. L. (2012). *The physical attractiveness phenomena*. Springer Science & Business Media. Retrieved from

[https://books.google.com/books?hl=en&lr=&id=BNJeBAAAQBAJ&oi=fnd&pg=PA1&dq=The+Physical+Attractiveness+Phenomena&ots=erDYWjdidZ&sig=RFD\\_0ZaqAAfBNVojZQahcfz5EM0](https://books.google.com/books?hl=en&lr=&id=BNJeBAAAQBAJ&oi=fnd&pg=PA1&dq=The+Physical+Attractiveness+Phenomena&ots=erDYWjdidZ&sig=RFD_0ZaqAAfBNVojZQahcfz5EM0)

Petroleum. (2016). In *Wikipedia*. Retrieved from

<https://en.wikipedia.org/w/index.php?title=Petroleum&oldid=797608302>

Ramakrishnan, C., Milnor, W. H., Perry, M., & Sheth, A. P. (2005). Discovering informative connection subgraphs in multi-relational graphs. *ACM SIGKDD Explorations Newsletter*, 7(2), 56–63.

Rodrigues Jr, J. F., Tong, H., Pan, J.-Y., Traina, A. J., Traina Jr, C., & Faloutsos, C. (2013). Large graph analysis in the gmine system. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 106–118.

Sasser, D. P. (2004). *Identifying the Benefits of Knowledge Management in the Department of Defense: a Delphi Study*. Air University.

Sevoflurane. (2016). In *Wikipedia*. Retrieved from

<https://en.wikipedia.org/w/index.php?title=Sevoflurane&oldid=797688679>

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, Ill. Univ. Illinois Press, 1, 17.

- Spiro, R. J. (1980). *Constructive processes in prose comprehension and recall.* (RJ Spiro, BC Bruce and WF Brewer, (Eds.)) *Theoretical issues in reading comprehension.* Hillsdale, New Jersey: Lawrence Erlbaum Associates. Google Scholar.
- Takahashi, H., & Mastuyama, A. (1980). An approximate solution for the Steiner problem in graphs. *Math. Japonica*, 573–577.
- Tong, H., & Faloutsos, C. (2006). Center-piece subgraphs: problem definition and fast solutions. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 404–413). ACM. Retrieved from <http://dl.acm.org/citation.cfm?id=1150448>
- Virus. (2016). In *Wikipedia*. Retrieved from <https://en.wikipedia.org/w/index.php?title=Virus&oldid=797985669>
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440.
- Witte, P. L. (1978). An Investigation of the Imaging Behaviors of Good and Poor Fourth Grade Readers with Easy and Difficult Text. Technical Report No. 455. Retrieved from <https://eric.ed.gov/?id=ED159606>
- Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist*, 24(4), 345–376.
- Wittrock, M. C. (1992). Generative learning processes of the brain. *Educational Psychologist*, 27(4), 531–541.
- Zika fever. (2016). In *Wikipedia*. Retrieved from [https://en.wikipedia.org/w/index.php?title=Zika\\_fever&oldid=797239605](https://en.wikipedia.org/w/index.php?title=Zika_fever&oldid=797239605)

