The Mechanics of the Presentation Mining Framework

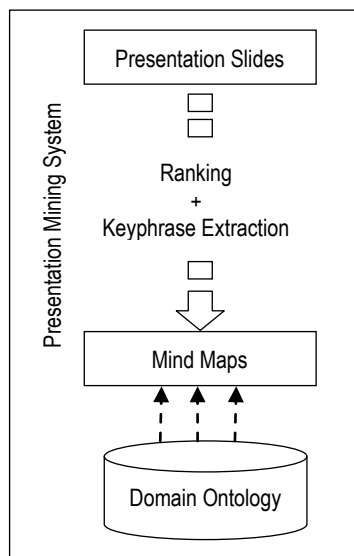Vinothini Kasinathan[a], Aida Mustapha[b], Masrah Azrifah Azmi Murad[a], Rahmita Wirza Rahmat[a], Evi Indriasari Mansor[a]

[a]Universiti Putra Malaysia, 43000 UPM Serdang, Selangor, Malaysia
[b]Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia

Graphical abstract

Abstract

This paper presents the mechanics of a presentation mining system that mines keywords and key phrases from a collection of PowerPoint slides and generates a mind map using the extracted words and phrases. The core of presentation mining lies in two stages; ranking the potential phrases and extracting the keywords and key phrases. The keywords and key phrases form a mind map, which is then evaluated against a domain ontology. The results of recall and precision are also compared between the existing key phrase extraction system called the KP-Miner and the proposed presentation mining system. The key phrase extraction algorithm by the proposed presentation mining system achieved higher recall and precision than KP-Miner, hence producing a more accurate visualization of the PowerPoint slides in the form of mind map.

Keywords: Presentation mining; text mining; keyphrase extraction

Abstrak

Artikel ini membentangkan mekanisme sistem perlombongan pembentangan yang melombong kata kunci dan frasa kunci daripada satu koleksi slaid pembentangan dan menjana peta minda menggunakan perkataan-perkataan dan frasa-frasa yang telah diekstrak. Asas kepada perlombongan pembentangan adalah berdasarkan kepada dua tahap; pengaturan frasa yang berpotensi dan pengekstrakan kata kunci serta frasa kunci. Kata-kata dan frasa-frasa kunci kemudiannya membentuk peta minda, yang kemudiannya dibandingkan dengan ontologi bidang. Keputusan susulan dan kepersisan juga dibandingkan dengan sistem pengekstrakan frasa kunci sedia ada yang dipanggil KP-Miner. Algoritma pengekstrakan frasa kunci dari system perlombongan pembentangan yang dicadangkan ini mencapai susulan dan kepersisan yang lebih tinggi daripada KP-Miner, seterusnya menghasilkan gambaran slaid pembentangan yang lebih jitu dalam bentuk peta minda.

Kata kunci: Perlombongan pembentangan; perlombongan teks; pengekstrakan frasa kunci

1.0 INTRODUCTION

In the field of education, two important aspects of teaching are (i) understanding the different types of learners in a classroom environment and (ii) identifying the suitable tools or technologies to enhance the learning process. According to the Visual Auditory Kinesthetic (VAK) model [1], there are three types of learners; auditory, visual, and kinesthetic. Amongst all types of learners, around 65.0% of the learners are visual learner [2] across a variety of domain such as

48.2% in occupational therapy [3], 88.7% in engineering and architecture [4], and 56.5% in medical [5].

Although visual learners are good in memorization, they are facing challenges in learning through oral lectures or reading [6] The ability of visual learner to visualize and understand the logic of subject matter is extremely important especially in higher level modules. Without the visual stimulations, visual learner tends to be multi-tasking in the class like texting, online browsing while listening to the lecture at the same time [7]. Due to this, visual learners have poor concentration in the class especially when it is oral lectures.

To enhance learning among the visual learners, classroom tools have shifted from board-based teaching to slide presentation such as the Microsoft PowerPoint. In PowerPoint slides, the modules are summarized according to chapters; picture, diagrams, and charts are also included in as the slide content. However, this tool is not without drawback. Several researches have raised issues regarding student understanding mainly due to the ineffectiveness of the presentation to control what the learner thinks [8] or how the learners reconstruct the materials based on their own understanding (Kinchin, 2008).

Weimer [8] also mentioned that around 80% of the learner tends to copy the slides' content rather than creating notes of their own. The concern of this issue is that whether the learner brainstorms and understands the study subject or just simply memorizes the points in the slides. Research has also shown that 25% of the students feel bored with slide-based presentation as the slides are usually crowded with too much of information and have no highlights for important terms [9].

As an alternative to slide-based presentation, researchers have proposed the concept of knowledge visualization in promoting effective learning. According to Zhang et al. [10], visualization accelerates learning by providing a different representation of the same material to the brain and it works better because human brain processes images way better than verbal or textual.

The essence of visualization lies in the selection of important keywords or key phrases available in a particular material as they provide learners an overview of a document. Key phrases are formed by a sequence of keywords that can be ranged from one to three words or more. To realize the concept of knowledge visualization, Kasinathan et al. [11] proposed a concept called presentation mapping that maps keywords from a slide presentation in a graphical form such as mind map or concept map to support visualization, hence the motivation of this research. The background research is summarized in Figure 1.
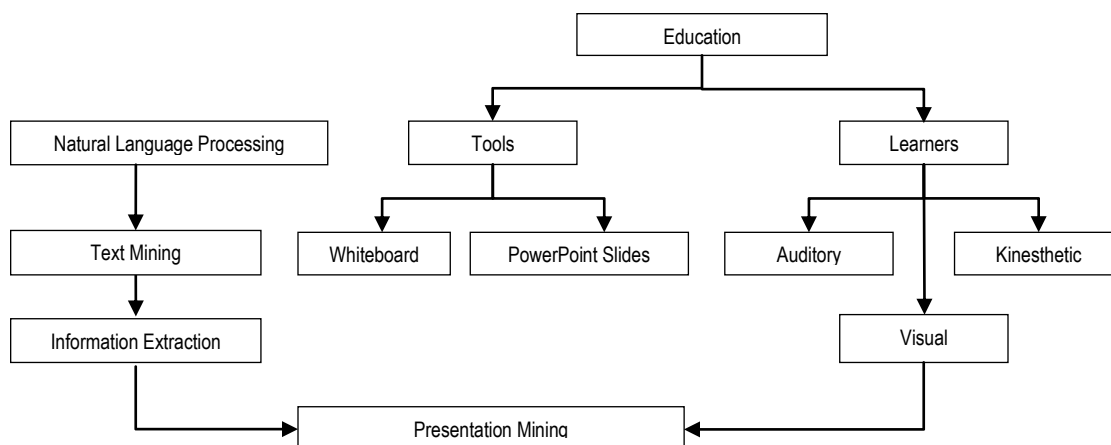

Figure 1 Motivation of research

Based on Figure 1, the objective of this paper is to develop a tool called the Presentation Mining system that extracts keywords and key phrases from the slide presentations using a set of Natural Language Processing (NLP) and text mining techniques and generate a mind map from those keys.

The remainder of this paper is organized as follows. Section 2 presents the mechanics behind the Presentation Mining system, Section 3 presents the evaluation framework, and finally Section 4 concludes with some direction of future works.

2.0 PRESENTATION MINING

Presentation mining is a system that mines keywords and key phrases from a collection of PowerPoint slides and generates a mind map using the extracted words and phrases (Figure 2). The system framework has three parts; pre-processing, mining, and visualization. The difference between presentation mining and text mining lies in their input and output. A conventional text mining framework takes plain text as input. The text is then processed using various natural language processing techniques to produce a variety of output such as

named entity tags [10,12-13] or part-of-speech (POS) tags [14-16].

In presentation mining, the system reads any number of PowerPoint slides and feeds the slides to a pre-processing stage. This stage performs several standard natural language processing techniques such as standardization, sentence segmentation, hyperlinks removal, tokenization, lemmatization, and symbol removal. The core of presentation mining lies in two stages; ranking the potential phrases and extracting the keywords and key phrases. Finally, the visualization module generates the keywords and key phrases extracted back into a Microsoft PowerPoint document as a singular mind map using the Smart Art Basic Radial. This decision is to ensure a cohesive implementation.
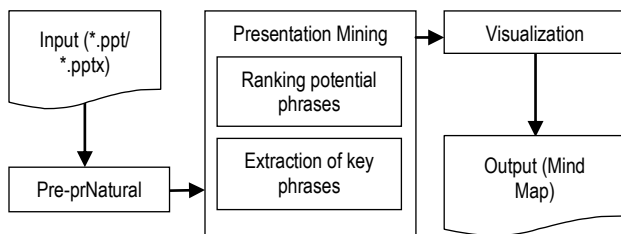

Figure 2 Presentation mining framework

2.1 Ranking: How to Identify Potential Phrases?

Post-processing the slides, each slide is now contains a number of candidate phrases along with their *n*-grams frequencies. *N*-gram weights are frequencies of candidate phrases. The minimum value for *n* is set to 1 and the maximum is 3. For example, if the candidate phrase is "Artificial Intelligent System", a 1-gram (unigram) will weigh "artificial", "intelligent", and "system"; a 2-gram (bigram) will weigh "Artificial Intelligent" and "Intelligent System"; and a 3-gram (trigram) will weigh "Artificial Intelligent System".

From the set of *n*-grams generated, the presentation mining system will return the *c*-gram. *c*-gram weights are the independency of an *n*-gram among its *n*-grams, whereby each *n*-gram is assigned with a *c*-value calculated using Equation 1.

$$c\text{-gram} = \begin{cases} \log_2|a| \cdot f(a) \\ \log_2|a|(f(a) - \dfrac{1}{P(T_\alpha)\sum_{b\in T_\alpha} f(b)}) \end{cases} \quad (1)$$

where *a* is the *n*-gram, $f(\cdot)$ is the frequency of occurrence in slides, T_α is the set of extracted candidate key phrases that contain a, and $P(T_\alpha)$ is the number of the candidate key phrases. Next, the weights for each candidate phrase are sorted in descending order. This will result as having a trigram on the top most and unigram at the bottom.

Table 1 shows the weights for all candidate phrases based on *n*-gram and *c*-gram. From this table, the phrases which do not have weight are the original

phrase, and phrases which are assigned weight are based on *n*-grams and *c*-gram of the original phrase. Bulleted phrases are identified as potential phrases among the *n*-grams by going through the process shown in Table 2.

Table 1 Preprocessing to identify the weight of the words

Phrase	Weight	C-gram
Linguistic		
• Linguistic	0.00619	0
Knowledge representation		
• Knowledge representation	0.01497	1
• Knowledge	0.01138	0
• Representation	0.01238	0
Norvig Artificial Intelligence		
• Norvig Artificial Intelligence	0.01541	1.58496
• Norvig Artificial	0.00973	0
• Artificial Intelligence	0.01276	2
• Norvig	0.00619	0
• Artificial	0.00708	0
• Intelligence	0.001138	0
Simple optimal agent design		
• Simple optimal agent	0.02343	1.58496
• Optimal agent design	0.02711	1.58496
• Simple optimal	0.00928	0
• Optimal agent	0.02034	0
• Agent design	0.02402	0
• Simple	0.00619	0
• Optimal	0.00619	0
• Agent	0.0283	0
• Design	0.01355	0

Table 2 Steps involves in Identify potential phrases

Identify Potential Phrase
Step 1:
• Identify whether the original phrase is a unigram.
• If original phrase is not a unigram, get bigrams and trigrams only. Otherwise, get the unigram.
Step 2:
• Identify the highest *c-gram* among the selected *n*-grams from Step 1.
• Get c-gram with the highest *c-gram* from *n*-grams selected from Step 1.
Step 3:
• If the number of *n*-grams with highest *c-gram* is more than 1, proceed to Step 4. Otherwise, *n*-gram with highest *c-gram* is selected as the potential phrase and skip Step 4.
Step 4:
• Get the highest *n*-gram weight among the selected *n*-grams from Step 1.
• Get *n*-gram with the highest weight as the potential phrase.

Finally, in Table 3, "Linguistic" is a unigram and it is identified as having the highest *c*-gram among its *n*-grams and is selected as potential phrase among the n-grams. Where else, "Knowledge representation" is a bigram, hence, the remaining unigrams are being

filtered out and only bigrams are selected. Because there is only 1 selected *n*-gram, hence it is identified as having the highest *c*-gram among its n-grams and is identified as potential phrase. Accordingly, "Norvig Artificial Intelligence" is a trigram, hence, only bigrams and trigrams are selected.

In Step 2, "Artificial Intelligence" is the only *n*-gram having the highest *c*-gram, hence it is then selected as potential phrase. Finally, "Simple optimal agent design" is a quadgram, hence, only bigrams and tri-grams are selected. In step 2, "Simple optimal agent" and "Optimal agent design" have the same highest *c*-gram. Hence, it proceeds to step 4 by choosing *n*-gram with the highest weight.

Table 3 Calculation of weights to identify keywords and key phrases

Phrase	*n*-gram	*c-gram*
Linguistic		
• **Linguistic**	0.00619	0
Knowledge representation		
• **Knowledge representation**	0.01497	1
Norvig Artificial Intelligence		
• **Norvig Artificial Intelligence**	0.01541	1.58496
• **Norvig Artificial**	0.00973	0
• **Artificial Intelligence**	0.01276	2
Simple optimal agent design		
• **Simple optimal agent**	0.02343	1.58496
• **Optimal agent design**	0.02711	1.58496
• **Simple optimal**	0.00928	0
• **Optimal agent**	0.02034	0
• **Agent design**	0.02402	0

2.2. Extraction: How to Select Key Phrases?

After the process of identifying potential phrases, the system will now prepare a list of united phrases. Potential phrases from each slide will be grouped together and go through a filtration process. Potential phrases which are substring of another potential phrase will be removed, as well as duplicate potential phrases.

Now the system have a clean collection of united potential phrases. The system will then loop through each united phrases to get the slide number where the phrase occurs in. After the process of pre-processing each slides in the presentation file, each slide has now a large number of candidate phrases which are noun phrases along with their *n*-grams and *c*-gram, refer Table 4 and Table 5.

Table 4 An example of identify using n-gram

Before Unite	
Slide 1	• Expert system industry
	• AI
	• Modern AI
Slide 2	• Neural network
	• Cognitive neuroscience
Slide 3	• Expert system industry
	• Cognitive science
	• Machine resource
	• Physical system
Slide 4	• Neural network
	• Expert system industry

Table 5 An example keywords and key phrase repeating in the slides number

After Unite		
Phrases		Slide number occurrence
1.	Expert system industry	1, 3, 4
2.	Modern AI	1
3.	Neural network	
4.	Cognitive neuroscience	2, 4
5.	Cognitive science	2
6.	Machine resource	3, 3, 3
7.	Physical system	

After uniting potential phrases and identifying the slide number for each phrase, the system will now choose the top 50 potential phrases and rank the slide numbers descending resulting in having the slide number with the highest number of occurrence in the top and the lowest occurrence in the bottom. For this example, there is only seven phrases as shown in Table 5, hence all of them will be included in the top 50.

Next is the weight calculation step in which the weight of each candidate phrase is calculated to enable ranking. Beside the common way of obtaining the weight of a term in a document using TF-IDF, a splitting value (*S*) is introduced in this work to provide higher weights for terms whose length in greater than 1 and for the terms that appear somewhere in the beginning of the document. If the length is just 1 then the value of *S* is set to 1. Equation 2 is then used to calculate the weight of keywords and key phrases;

$$weight = TF*IDF + S \qquad (2)$$

where TF is the occurrence of phrase in each slide content and title, and it is added up to form *tFreq*. TF is *tFreq* divided by the number of distinct words in the PowerPoint slide. Meanwhile, IDF is the occurrence of phrase in each slide content. It will increase the value of document frequency (*dFreq*) by 1 if occurrence in current slide content and title is not 0. DF is number of slides divided by the value of *dFreq*, therefore IDF is the logarithmic value of DF.

Finally, *S* is to split phrase words by a hyphen. The presentation mining system will loop through every split words, get the POS tag for each split word and add up the weights for each word weight together to form *S*. If the split word is a stop word or a POS tag, the weight will not be calculated. Table 6 shows the key phrases appear in the slides.

Table 6 Number of keywords/keyphrases appear in slides

| Slide No.	Frequency
3	4
1	2
2	2
4	2

From Table 6, slide number 3 have the highest frequency of 4. Hence, the system will now go through each of the united phrase and find out the top 5 phrases which occurs in slide 3. In result, "Expert system industry", "Cognitive science", "Machine resource", and "Physical system" will be selected as the key phrases of slide 3. This is followed by selecting key phrases for slide 1, 2, and 4. The system will take a maximum number of 6 slides only in order to present it in a mind map.

3.0 EVALUATION

To evaluate the mind map produced by the presentation mining system, the keyword and key phrase nodes in the map are colored as red and blue based on their match to the domain ontology built using Protégé (Kasinathan et al., 2015). This section describes in detail the domain ontology as well as the evaluation metrics.

3.1. Domain Ontology

Although the presentation mining system is designed to be domain-independent, tor the purpose of evaluation, the input slides are limited to a textbook called the Artificial Intelligence: A Modern Approach (AIMA) by Russell and Norvig [17]. It is a widely used textbook for AI courses in 1,300 universities throughout 110 countries. According to the Citeseer (http://citeseerx.ist.psu.edu/index). it is also the 22nd most cited source in Computer Science.
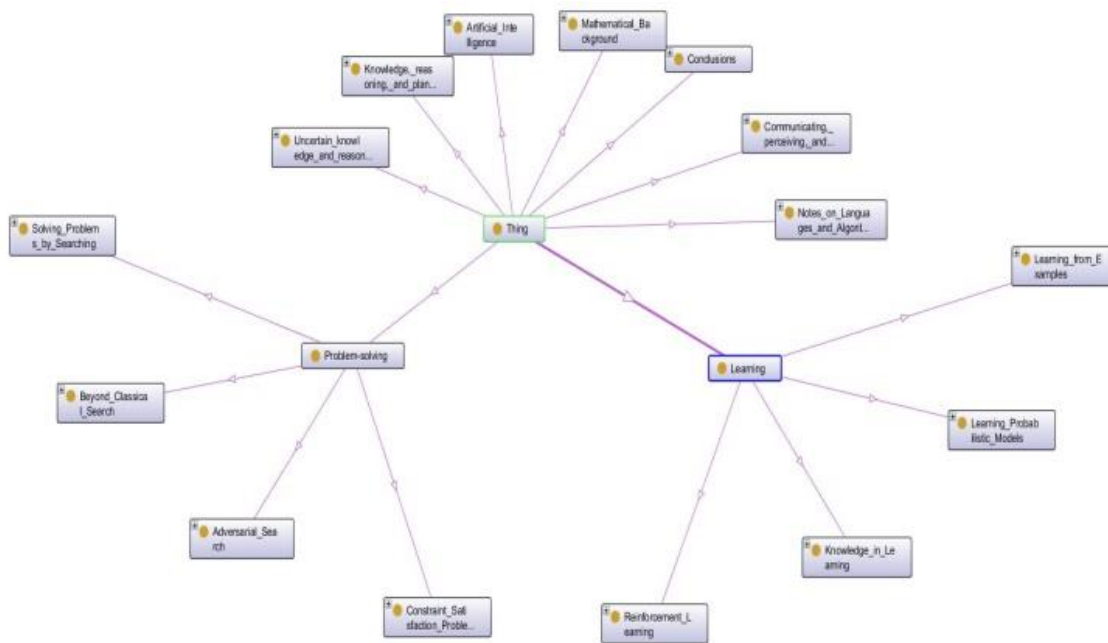
Figure 3 The ontology in Protégé based on the AIMA textbook by Russell and Norvig [17]
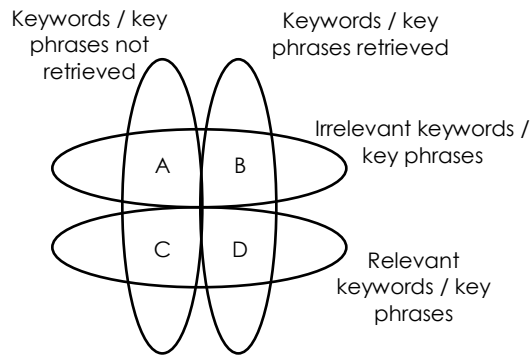
3.2. Evaluation Metrics

Measures are important when evaluating a text mining system. The same method is being used in the presentation mining system. Recall and precision are the basic measures used in evaluating search strategies and will be used throughout the evaluation. These measures assume that:

- there is a set of keywords and key phrases in the domain ontology that is relevant to the content of slides
- keyword and key phrase are assumed to be binary; either relevant or irrelevant (these

measure do not allow for degrees of relevancy)
- the actual retrieval set may not perfectly match the set of relevant

Figure 4 shows the definition of precision and recall for the presentation mining system with four different states; keywords and key phrases that were retrieved, that were not retrieved, the relevant and the irrelevant.

Keywords / key phrases not retrieved

Keywords / key phrases retrieved

Irrelevant keywords / key phrases

A	B
C	D

Relevant keywords / key phrases

Figure 4 Definition of precision and recall

The intersections of these states (A, B, C, D) represent the following:

- A is the number of irrelevant keyword and key phrase not retrieved (true negative)
- B is the number of irrelevant keyword and key phrase retrieved (false positive)
- C is the number of relevant keyword and key phrase not retrieved (false negative)
- D is the number of keyword and key phrase records retrieved (true positive)

The equations for states (A, B, C, D) are shown in Equation 3 and Equation 4. In the context of presentation mining, the recall addresses the question: "Given a correct key phrase, will the system extracts it?" and the precision addresses the question: "Given a correct key phrase extracted, how likely it is to be correct?" Note that in this paper keywords are assumed to part of key phrases and correctness refers to the existence of the phrase in a domain ontology, which is the A Modern Approach (AIMA) by Russell and Norvig [17] textbook in this experiment.

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

As noted earlier, keywords and key phrase are considered relevant or irrelevant based on the calculation of recall and precision. Obviously, keywords and key phrases are marginally relevant or somewhat irrelevant. Others may be very relevant and others completely irrelevant.

This problem is complicated by individual perception: what is relevant to one person may not be relevant to another. This is also important when considering "partial matches". Measuring recall is difficult because it is hard to determine how many relevant keywords exist in a slide.

3.3. Visualization

In the visualization stage of presentation mining (see Figure 2), the weighted keywords and key phrases extracted are then generated into a Microsoft PowerPoint document using the Smart Art Basic Radial as shown in Figure 5. The measurement of recall and precision are based on node color in the generated mind map, whereby blue nodes indicate keywords and key phrases which are in Protégé and the red nodes indicate that the keywords and key phrases extracted from slides via presentation mining. The representations of recall and precision based on Figure 5 are as follows:

- True Positive (Correct in blue) – Correct keywords or key phrases available in Protégé
- True Negative (Wrong in red) – Wrong keywords and key phrases extracted
- False Positive (Wrong in blue) – Wrong keywords or key phrases available in Protégé
- False Negative (Correct in red) – Correct keywords and key phrases extracted
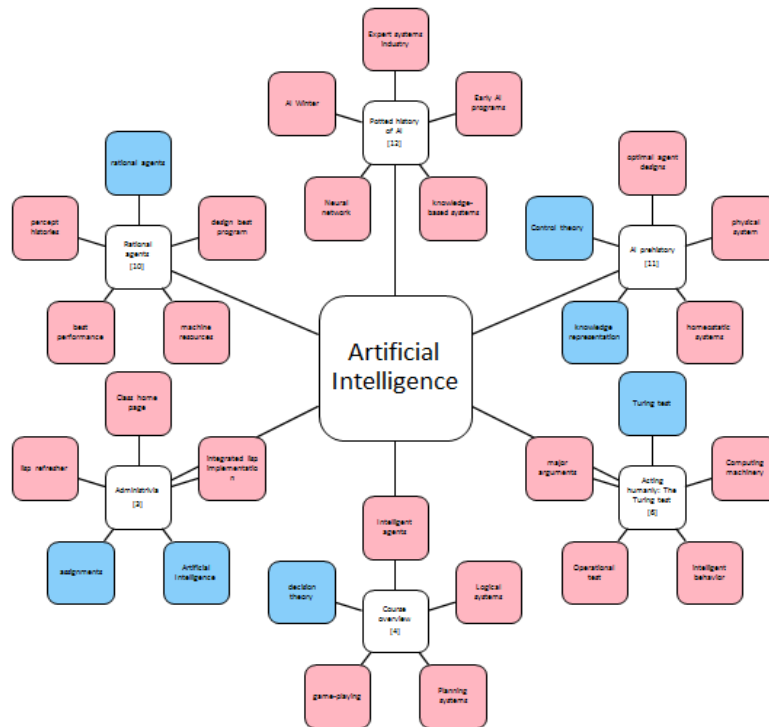
(3)

Figure 5 Output of the presentation mining system: a mind map

Note that validation against the ontology in Protégé is necessary to determine the different states that make up the calculations of recall and precision. However, red nodes are not necessarily wrong because the mind map only indicates that the words or the phrase is not available in the domain ontology.

3.4. Comparison with Existing Extraction System

The resulting keywords and key phrases extracted by the presentation mining system are also compared with those produced by another extraction system called the GenEx [18], KEA [19] and KP-Miner [20], KP-Miner [20] are some popular key phrase extraction systems proposed to extract words automatically rather than the time consuming manual key phrases assignment as well as the costly domain specialist.

While GenEx and KEA both treated the task of extracting key phrases as a supervised learning approach where key phrases are trained in order to build a model for identifying the probabilities of identified candidate key phrases, KP-Miner uses a non-learning approach. In KP-Miner, no training documents are required in order to identify key phrases within a given document. In an improved KP-Miner, an *n*-gram filtration technique is added to the extraction algorithm in order to enhance the accuracy of identified key phrases [21].

Table 7 shows the list of keywords and key phrases extracted using both extraction systems checked against a list of keys assigned by author for testing. The assigned keys are Turing Test, Neural Network,

Knowledge Representation, Intelligent Agents, and Knowledge-based systems**.**

Table 7 Extracted keywords and key phrases by KP-Miner and the proposed Presentation Mining system

KP-Miner (1 match)	Presentation Mining System (6 matches)
Linguistics knowledge representation	Expert system industry
Simple optimal agent designs	Early AI programs
Control theory homeostatic systems	Knowledge-based systems (match)
Mathematics formal representation	Neural network (match)
Psychology adaptation	AI Winter
Information-processing psychology replaced prevailing orthodoxy of behaviorism	Optimal agent designs
	Physical system
	Homeostatic systems
	Knowledge representation (match)
Level of abstraction	Control theory
Requires scientific theories of internal activities	Turing test (match)
Testing behavior of human subjects	Computing machinery
	Intelligent behavior
Cognitive revolution	Operational test
Machines behave intelligently	Major arguments
Suggested major components	Intelligent agents (match)
	Logical systems
Imitation game	Planning systems
Computing machinery	Game-playing
Intelligent agents (match)	Decision theory
Logical systems	Class home page
Planning systems	Integrated lisp implementation
	Artificial intelligence (match)
	Assignments

Philosophical issues	Lisp refresher
Game-playing	Rational agents
Class home page	Design best program
Norvig Artificial Intelligence	Machine resources
Integrated lisp	Best performance
implementation	Percept histories
Modern approach	
http	
Purpose of thinking	
Idea of mechanization	
Rules of derivation	
Intelligent behavior	
Greek schools developed	

Table 8 shows the recall and precision for both systems tested on all chapters from the A Modern Approach (AIMA) by Russell and Norvig [17] textbook. Note that both KP-Miner and the Presentation Mining system are unsupervised, keyword and key phrase extractions systems, which means there were no testing and training data required. Both systems take input of slides from 17 chapters.

Table 8 Extracted keywords and key phrases both extraction systems

Slides	Recall		Precision	
	KP-Miner	PM	KP-Miner	PM
Chapter 1	0	0.857	0	0.316
Chapter 2	0.571	1	0.571	0.333
Chapter 3	0.8	1	0.8	0.5
Chapter 4a	0.75	1	0.75	0.333
Chapter 4b	0.5	1	0.5	0.36
Chapter 5	1	0.75	1	0.27
Chapter 6	1	1	1	0.7
Chapter 7	0.5	0.857	0.5	0.5
Chapter 9a	0.75	1	0.75	0.4
Chapter 9b	1	0.8	1	0.267
Chapter 11	0.286	0.429	0.286	0.214
Chapter 13	0.5	1	0.5	0.153
Chapter 14	1	1	1	0.4
Chapter 15a	0.8	1	0.8	0.333
Chapter 15b	0.333	0.667	0.333	0.4
Chapter 16	0.4	0.571	0.4	0.333
Chapter 17a	0.5	1	0.5	0.429

*PM Presentation Mining System

From the table, the presentation mining system performed better than the KP-Miner in terms of higher precision. A precision value of 1 shows the best results and the presentation mining system achieved more optimal precision as compared to the KP-Miner. It also covers 10 chapters while the optimal precision for KP-Miner only covers eight chapters. In other readings, the precision for KP-Miner in Chapter 1 is 0 as compared to 0.857 for presentation mining system.

4.0 CONCLUSIONS AND FUTURE PLANS

This paper described the mechanics behind a presentation mining system from taking in a collection of slides, performing pre-processing, ranking and extracting key phrases, and finally generating a mind map based on the phrases. The input is limited to PowerPoint slides produced by Microsoft PowerPoint 2010 and newer version. The conversion of PowerPoint slides into plain text also excluded diagrams, charts, images and picture in the slides. The system is intended to be open-domain, however, for the sake of evaluation, testing is limited to the slides originated from an Artificial Intelligence book by Russel and Norvig [17]. The system was compared against an existing key phrase extraction system called the KP-Miner [20] to evaluate the correctness of the keywords or key phrases extracted.

In the future, the research plans to improve the presentation mining system to cater other types of content such as tables, charts, SmartArt graphics, and pictures. By reforming and presenting the same textual knowledge into a visual form, students could improve their understanding on the subject area while at the same time increase memory retention. Students will be able to access to deeper, more complex modes of knowing, understanding and valuing a discipline.

Acknowledgement

References

[1]   Kasinathan, V., Mustapha, A. 2015. Ontology Support for Web-based Presentation Mining. *In Proceedings of the Second International Conference on Advanced Data and Information Engineering, 25 - 26 April 2015, Bali, Indonesia.*
[2]   Turkington, C., & Harris, J. 2006. The Encyclopedia of Learning Disabilities.
[3]   Paulraj, S. J., Ali, A. R., & Vetrayan, J. 2013. Learning Style Preferences among Diploma Students of Middle-East Journal of Scientific Research. 14(5): 603-609.
[4]   Jubilo, A. B., & Faller, M. 2013. Index of Learning Styles of Engineering and Architecture Students for SY 2011-2012. Retrieved September, 2015.
[5]   Knublauch, H., Fergerson, R. W., Noy, N. F., & Musen, M. A. 2004. The Protégé OWL plugin: An Open Development Environment For Semantic Web Applications. *In The Semantic Web–ISWC 2004.* 229-243. Springer Berlin Heidelberg.
[6]   Knublauch, H., Fergerson, R. W., Noy, N. F., & Musen, M. A. 2004. The Protégé OWL plugin: An Open Development Environment For Semantic Web Applications. *In The Semantic Web–ISWC 2004.* 229-243. Springer Berlin Heidelberg.
[7]   Walker & Zur, A. W. 2015. On Digital Immigrants and Digital Natives: How the Digital Divide Affects Families, Educational Institutions, and the Workplace.
[8]   Weimer, M. 2012. Does PowerPoint Help or Hinder Learning?

[9]   Ding, X., & Liu, J. 2012. Advantages and Disadvantages of PowerPoint in Lectures to Science.

[10] Zhang, S., & Elhadad, N. 2013. Unsupervised Biomedical Named Entity Recognition: Experiments With Clinical And Biological Texts.

[11] Kasinathan, V., Mustapha, A., & Rani, M. F. 2013. Structured-based Algorithm for Presentation Mapping in Graphical Knowledge Display

[12] Finkel, J. R. 2007. March Named Entity Recognition and the Stanford NER Software.

[13] Zhou, G., & Su, J. 2002. Named Entity Recognition using an HMM-Based Chunk Tagger.

[14] Robin. 2009. Part-of-Speech Tagging

[15] Toutanova, K., & Manning, C. D. 2000. Enriching the Knowledge Sources used in a Maximum Entropy Part-of-Speech Tagger. *In Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. In conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13. 63-70. Association for Computational Linguistics.

[16] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. 2003. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. 173-180. Association for Computational Linguistics.

[17] Russel, S., & Norvig, P. Artificial Intelligence: A Modern Approach, 2003. EUA: *Prentice Hall.*

[18] Turney, P. D. 2000. Learning Algorithms for Keyphrase Extraction. *Information Retrieval*. 2(4): 303-336

[19] Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. 1999. KEA: Practical Automatic Keyphrase Extraction. *In Proceedings of the fourth ACM conference on Digital libraries*. 254-255. ACM.

[20] El-Beltagy, S. R. 2006. KP-Miner: A Simple System for Effective Keyphrase Extraction. In Innovations in Information Technology. 1-5. IEEE.