International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 13 (2015) pp 33893-33897
 https://dx.doi.org/10.37622/IJAER/10.13.2015.33893-33897

Analysis of Semi Automated Extraction of Domain Specific Cognitive Portrayals

M. Pushpa

Research Scholar, Department of Computer Science,
Bharathiar University, Coimbatore, Tamilnadu, India

Dr. K. Nirmala

Associate Proffessor, Department of Computer Science,
Quiad-e-Millath Government Arts College for women,Chennai, Tamilnadu, India

Abstract
Cognitive portrayals are quantifiable text that represents certain concept. This paper has considered David Merrill's four cognitive structures namely Activation, Demonstration, Application and integration that can be quantified from domain specific documents. As these cognitive portrayals represents concepts, extractions from a document falls under 'Concept mining'. Although success rate of extraction is the highest through manual method, the paper attempts to compare the manually computed results with that of a semi-automated process uses Term frequency – inverse document frequency a text mining approach. While pure keywords are extracted automatically concept words need judgments by analyst and hence semi automation is proposed. Textual instructional document of a selected topic has been considered for the analysis.
A four parametric comparisons are made with the document and one semi automated extraction techniques conclusion are drawn from the study.

Keywords: concept mining, Text extraction, cognitive portrayals and performance comparison

I. INTRODUCTION

Keywords are used for textual extraction while concept keywords (explained later) are used for concept extractions from textual documents. Keyword extraction may be fully automated and is being adapted widely concept may be extracted through content analytical methods. Content analytical studies have been reported by many researchers. David Robertson [12] created a coding frame for a comparison of modes of party competition between British and American parties. It was developed further in 1979 by the Manifesto Research Group aiming at a comparative content-analytic approach on the policy positions of political parties. This classification scheme was also used to accomplish a comparative analysis between the 1989 and 1994 Brazilian party broadcasts and manifestos as reported by Carvalho F. (2000). It is also noted that every content analysis should depart from a hypothesis based study. Concept mining technique attempts for extraction of data from instructional materials like text books, question banks etc.,
SaleemaAmershi et al [4] have attempted extraction of data from instructional textual documents for conceiving concepts under exploratory learning environments. They continue to state that this approach, however, is often difficult and time consuming, especially for novel applications such as exploratory learning environments, for which there is still limited knowledge on what constitutes effective exploratory behaviour. The authors further substantiate that the few existing approaches to this problem have been very knowledge intensive, relying on time-consuming, detailed analysis of the target system, instructional domain and learning processes. Since these approaches are so domain/application specific, they are difficult to generalize to other domains and applications.
In view of the above an attempt is made by us to device a semi automated process to extract near concept words that are specified for instructional purpose from two types of documents namely text book and question papers.

II. CONCEPT WORDS

According to Masaru Ohba et al [2], concept word or concept keyword is a word that represents a key concept in understanding, and dirent (directory entry) is an instance of concept keywords. Unfortunately there is no clear definition of concept keywords, since they are highly based on subjective judgment. The authors have used three types: Ideal concept keywords, which have proven to improve program understanding by some objective measurements Human-selected concept keywords, which a developer or reviewer believes are ideal concept keywords. Machine-extracted concept keywords, which a method like TF/IDF (Term Frequency Inverse Document Frequency) produced as an approximation of ideal or human selected ones. We have proposed a semi automated extraction technique in which near concept keywords or doubtful sentences are analyzed by a developer or reviewer.

III. CONCEPT BASED COGNITIVE PORTRAYALS

David Merrill [1], in his 'First Principles of Instruction' divides any instructional event into four phases, which he calls 'Activation', 'Demonstration', 'Application' and 'Integration'. Central to this instructional model is a real-time problem-solving theme, called 'Problem'. Merrill suggests that fundamental principles of instructional design should be relied on and these apply regardless of any instructional design model used. Violating this would produce a decrement in learning and performance.

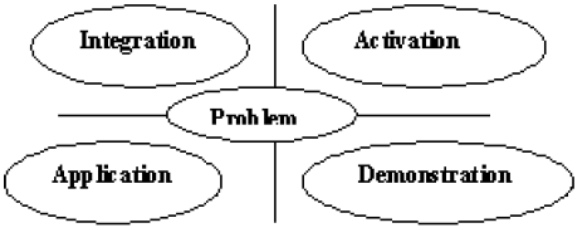

Fig.1 First Instructional principle

Each phase is is a cognitive portrayal for learning a concept. Those concept are briefly explained below:

A.Activation

This is the first or starting phase in the instructional process. New knowledge builds on the concept learner's existing knowledge. Concept learners recall or apply knowledge from relevant past experience as a foundation for new knowledge. This could be from previous courses or job experiences undergone by the learner. For instance, recall the old relevant information such as data types and memory addresses. During Merrill's Activation phase, prior knowledge (or experience) is recalled and emotions are triggered. Not only pre-knowledge should be activated during this phase, but mental models as well. If these mental models consist of misconceptions, the instructional process could modify them. This phase incorporates skills like 'Mental ability' of learning while conceiving knowledge.

B. Demonstration

New knowledge is demonstrated to the learner through this cognitive portrayal. Concept learners learn when the media (say instructor or textual materials or e-content media like voice or video) demonstrates what is to be learnt, rather than merely telling information about what is to be told. The learner observes or perceives while this portrayal is active. For example learning the principles involved in stacks or ques. The media used in the process is expected to play a relevant instructional role. Explain with examples, understand information with meanings, predict consequences, order, group, and infer causes are some samples for demonstration. Demonstration focuses the learner's attention on relevant information and promotes the development of appropriate mental models. It shows actions in a certain sequence, which can simplify complex tasks and facilitate learning.

C. Application

The concept learner applies his acquired new knowledge to a numerical problem. This is the practice phase, where learners are required to use their knowledge and skill to solve relevant problem. Writing programs using 'stacks' and 'queues' for a specific problem is a good example. The purpose of a practice phase in the instructional events is to provide an opportunity for learners to develop proficiency and become experts. During this phase, cognitive processes come into play; and there is a search for meaningful patterns and mental programmes occur in the learner's mind.

D. Integration

New knowledge is integrated into the concept learner's terminal behavior. This is the transfer phase where learners apply or transfer their newly found knowledge or skills into their workday practices. This is felt, if learners can a) demonstrate their new knowledge or skills, b) reflect-on, discuss their new knowledge and skills and c) create, invent and explore new ways to use their new knowledge and skills. Seeing patterns and organizing by recognition of hidden meanings, are some samples. Use old ideas to create new ones (relate knowledge from several areas). Assess values of ideas (make choices based on supported arguments). Most of the instructional events end with an assessment phase. During this phase learners have to prove themselves, that they have acquired the new knowledge and skills. David Merrill [1] calls this as the Integration phase, during which the learners get the opportunity to prove new capabilities and show newly acquired skills.

The four phases explained above, are actually cognitive portrayals that trigger the concept learners' inherent abilities namely: activating or manual ability, (demonstrating) perceiving ability, applying ability and integrating ability.

Action verbs may be used, which indicate the depth of understanding of any concept, expected from the concept learner. With simple definitions of the four phases (components) or abilities of Merrill's model, several action verbs could be taken from published literature apart from Merrill's own verbs. Simple definitions of these four components and the relevant action verbs, as taken from literature are presented.

IV. PROPOSED METHODOLOGY

Keyword search has a lot of problems. If is prone to being over-inclusive, i.e. finding some non-relevant documents, and under-inclusive, i.e. not finding some relevant documents. As a tremendous amount of time have spent on researching conceptual search in the emergence of concept search technologies are new and interesting and using these technologies you can find documents that keywords search can't find, therefore, concept search must be better than keyword search.

Concept mapping provides a framework for organizing conceptual information in the process of defining a word. The framework of the concept map contains category or target

word, properties of the concept word and examples of the concept or target word. Our method for generating concept word includes the

i) Identification of concepts based on FPI
ii) Identify the concept word with the help of the action verb
iii) Determining the relationship between the action verb and the concept
iv) Classification of keywords

With simple definitions of the four phases (components) or abilities of Merrill's model, several actions verbs could be taken from published literature apart from Merrill's own verbs. Some of the human generated concept words used in the proposed system for extraction purposes are detailed below:

1) Activation (Concept: "*Where do I start?*")

a) Does the instruction direct learners to recall, relate, remember, repeat or recognize the knowledge from relevant past experience that can be used as a foundation for the new knowledge (problem).
b) If learners have limited prior experience, does the instruction provide relevant experience that can be used as a foundation for the new knowledge?

Based on the above questions a set of concept key words for this phase, as taken from literature are presented below:
list, define, tell, name, locate, identify, distinguish, acquire, write, underline, relate, state, recall, select, repeat, recognize, reproduce, measure, memorize.

2) Demonstration (Concept: "Don't just tell me, show me!")

a) Does the instruction demonstrate (show example) of what is to be learnt rather than merely providing information about what is to be learnt?
b) Are the demonstrations (examples) consistent with the content being perceived?

Based on the above questions a set of concept key words for this phase, as taken from literature are presented below:
demonstrate, summarize, illustrate, interpret, contrast, predict, associate, distinguish, identify, show, label, collect, experiment, recite, classify, discuss, select, compare, translate, prepare, change, rephrase, differentiate, draw, explain, estimate, fill in, choose, operate, perform, organize.

3) Application (Concept: "Let me do it!")

a) Do learners have an opportunity to practice and apply their newly acquired knowledge or skill?
b) Are the application (practice) and assessment (tests) consistent with the stated or implied objectives?

Based on the above questions a set of concept key words for this phase, as taken from literature are presented below:
apply, calculate, illustrate, solve, make use of, predict, construct, assess, practice, restructure, classify.

4) Integration (Concept: "Watch me!")

a) Does the instruction provide techniques that encourage learners to integrate (transfer) the new knowledge or skill into their everyday professional life?
b) Does the instruction provide an opportunity for learners to create, invent, or explore new and personal ways to use their new knowledge or skill?

Based on the above questions a set of concept key words for this phase, as taken from literature are presented below:
analyze, resolve, justify, infer, combine, integrate, plan, create, design, generalize, assess, decide, rank, grade, test, recommend, select, explain, judge, contrast, survey, examine, differentiate, investigate, compose, invent, improve, imagine, hypothesize, prove, predict, evaluate, rate.

Some of the concept words may be found repeating in two or more phases. As these words are only indicative of a concept they cannot be accepted *per-se* in a particular concept category. Many terms, in similar lines, cannot be used *per-se* and needs interpretations. Therefore a semi automated approach is proposed by us.

Term Frequency based automated analysis:
1. Text book analysis
2. Question paper analysis

Manual method for benchmark:
1. Text book analysis
2. Question paper analysis

V. TERM FREQUENCY BASED SEMI AUTOMATED ANALYSIS

Term frequency classifier is a technique that extracts words from a document, and indicates how much important that word is to a document. The importance of a concept word to a document is determined by knowing the relation ship. Number of times the concept word exists in a document is proportional to its importance.
i = Number of times the concept word is appearing in 'n' worded document.
The term frequency **tf = i/n**. By definition, if there are a total N documents and the concept word is appearing 'k' times, the inverse document frequency is

$$\mathbf{Idf = \log(n/k) * tf}$$

In a number of cases, the word may be appearing in an extended form(Such as grammatically added with 's','ed','ing', ect.,). In a few cases the word might refer to different conceptual context. For example 'list three types of data structures' may fall under activation concept. But 'list' those data structures, which according to your judgment would be efficient" will fall under 'Integration' concept. Therefore a semi automated algorithm is suggested by us. In such doubtful cases as per repetition of keywords in a sentence or ambiguous categories, our algorithm will throw the sentence for developers or viewers judgment for

categorization. Our presentation includes comparison of results

i) Fully automation (Concept word without human intervention)
ii) Semi automation.(Concept words with human intervention)
iii) Manual content Analysis (Fully analyzed by a human)

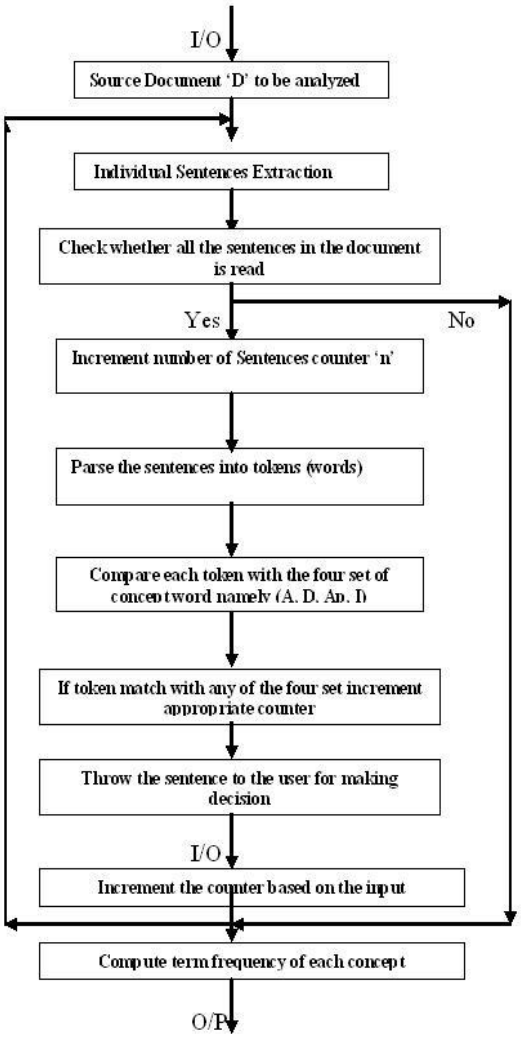

Fig.2 Proposed System Architecture Flow

Results are presented in Table I& Table II.

VI. MANUAL METHOD OF CONTENT ANALYSIS

The objective of manual content analytical exercise is to determine the context on the concepts in terms of the four cognitive portrayals that are represented in the chosen instructional materials. Merrill [1] has stated that any subject content material can be represented in any one of these four cognitive portrayals or combination of these cognitive portrayals. To determine the quantities of these four portrayals, the instructional material is divided into pieces of materials that are used for a unit time of reading, say 10 minutes duration each. Such pieces are analyzed for the presence of each cognitive portrayal. The analysis is done either using the action verb of each portrayal or verifying it with the definition of each portrayal. Accordingly the same chosen material that had been adapted for semi automated analysis is now adapted for the manual method also. As the manual method is executed as per definition of portrayal, the values obtained will be exact and the total value of the portrayal would yield to 100%. Therefore these values are taken as benchmark values. The benchmark values obtained from manual method is presented in Table II.

VII. RESULTS AND DISCUSSIONS

As the experiment deals with the two independent documents namely textual instructional document and question paper, the inverse document frequency would not be necessary for the intended comparative study.

Term frequency method extracts only textual keywords, where as the experiment aims for extracting concepts. In regard to this critical issue, the action verbs of each cognitive portrayal, a few extensions of each word with respect to grammar have been added for analysis. For example the word 'list' is added with a few more words like 'listed', 'listing', 'lists' etc.,

The result have been tabulated in Table I& Table II.

TABLE I. TERM FREQUENCY OF INSTRUCTIONALTEXTUAL INSTRUCTIONAL DOCUMENT

S.No.	Cognitive Portrayal	Term frequency extracted from the document in ratio of %		
		Fully Automated	***Semi Automated***	***Manual Method***
1	Activation	15%	17%	22%
2	Demonstration	30%	35%	40%
3	Application	15%	18%	22%
4	Integration	10%	12%	16%
5	Uncertain	30%	18%	

TABLE II. TERM FREQUENCY OF QUESTION PAPER DOCUMENT

S.No.	Cognitive Portrayal	Term frequency extracted from the document in ratio of %		
		Fully Automated	***Semi Automated***	***Manual Method***
1	Activation	19%	20%	21%
2	Demonstration	31%	33%	35%
3	Application	20%	20%	24%
4	Integration	15%	17%	20%
5	Uncertain	15%	10%	

VIII. CONCLUSION

Term frequency based algorithm for quantifying cognitive portrayals is working well although it can extract only concept keywords. As presumed earlier, the program can never near to manual methods of quantifying cognitive portrayals. The results on the quantification of cognitive portrayals of textual instruction materials by the two methods compare very poorly. However the results on the quantification of question papers compare well. Thus it is concluded term frequency based extraction technique can be applied to documents that have action verb based questions.

References

[1] S. David Merrill M, (2007), "Converting e *sub3*-learning to e *3rd power*-learning: an alternative instructional design method" in Carliner and P. Shank (Eds.), e-Learning: Lessons Learned, Challenges Ahead (Voices from Academe and Industry). Pfeiffer/Jossey-Bass.

[2] MasaruOhba, Katsuhiko Gondow (2005),Toward mining "concept keywords" from identifiers in large software projects. ACM SIGSOFT Software Engineering Notes 30(4): pp. 1-5

[3] AdityaParameswaran, AnandRajaraman, Hector GarciaMolina, (2010), "Towards The Web Of Concepts: Extracting Concepts from Large DataSets", Proceedings of the VLDB Endowment, Vol. 3, No. 1, VLDB Endowment 21508097/10/09...pp 566-577.

[4] SaleemaAmershi, and Cristina Conati (2009), "Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments", Journal of Educational Data Mining, Article 2, Vol 1, No 1

[5] Salton G 1983, McGill M. Introduction to modern Information Retrieval, McGrawHill

[6] ArunK.Pujari (2001), Data mining techniques, universities Press(India) Pvt. Ltd.

[8] S.Saraswath, "Design of textual presentation from online information using hybrid approaches", Vol 1 ICTACT Journal on soft computing, OCT 2010, Issue 02, issn 0976-6561, pp 105-112

[9] Amershi, S., Conati, C.(2006) Automatic Recognition of Learner Groups in Exploratory Learning Environments. Proceedings of ITS 2006, 8th International Conference on Intelligent Tutoring System.

[10] Merceron, A., Yacef,K.(2008) Interestingness Measures for Association Rules in Educational Data. Proceeding of the First International Conference on Educational Data mining.

[11] Gupta, V., &Lehal, G. S., (2010), "A Survey of Text Summarization Extractive Techniques. Journal of Emerging Technologies in Web Intelligence", 2(3): pp. 258-268.

[12] David Robertson (1976)-Wikipedia, (2008) http://en.wikipedia.org/wiki/Content_analysis

[13] Kowalski G. J, and Maybury M. T, (2012), "Information Storage and Retrieval Systems", Fifth Indian Edition, Springer, New Delhi, pp.14-15.

[14] ManishaPravin Mali, Mohammad Atique, (2012) "A review of Text Classification using Fuzzy logic", Proceeding of the International conference on Mathematics in Engineering and Business Management, Vol.2, pp.324-329.

[15] PouyaKhosravianDehkordi and FarshadKyoumarsi, Islamic Azad University, Shahrekord Branch and Iran, (2013), "Using Gene Expression Programming in Automatic Text Summarization", Middle-East Journal of Scientific Research 13 (8): 1070-1086.

[16] SuleymanCetintas and Luo Si, Yan Ping Xin, Dake Zhang and JOO Young Park, Ron Tzur, (2010), "A Joint Probablistic Classification Model of Relevant and Irrelevant Sentences in Mathematical Word Problems", Journal of Educational Data Mining, Article 3, Vol.2, No.1, pp.83-101.

[17] Merrill M.D., (2002),"First Principles of Instruction", Englewood Cliffs, NJ: Educational Technology Publications.

[18] Pushpa M, and Nirmala K, (2012) "Text Categorization Using Activation Based Term Set", International Journal of Computer Science Issues (IJCSI) pp.

[19] Carvalho F. 2000-Wikipedia, (2008) http://en.wikipedia.org/wiki/Content_analysis