International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

Content Grading System for Tamil based on Indexed Set Weights using PC-Kimmo

Thivaharan S [1]
[1]Asst. Prof (Sel. Grade), Dept. Of CSE,
PSG Institute of Technology and Applied Research,
Tamilnadu, India

Hariharan K [2], Christie Jerin Kumar R [3]
[2,3]Students, Dept. of CSE,
PSG Institute of Technology and Applied Research,
Tamilnadu, India

Abstract: **Tamil has rich vocabulary and has sentences with words that can be complexly inflected. These inflected words give high degree of meaning thereby helping to "making through" the seemingly equal Tamil Investigative contents. Content can be any of the following nature: textual, image, audio, video or combination of the aforesaid. Content Grading Systems (CGS) for any native language automates the process of manual valuation or grading the originality of the content. As in manual valuation which needs an answer key, CGS also refers answer key in the form of reference contents and makes the following decisions: 1). Rejecting the content completely, 2). Accepting the content completely, 3). Partially accepting the content. A CGS should aim at minimizing the hypothetical grading and should have reasons for the decision made. In this paper we propose basic frameworks of CGS for Tamil along with the Architecture, Supportive components and a minimized strategy to accomplish Grading. All contents irrespective of the language and nature consist of key terms. We propose in this paper, based on availability of key terms in the Cognitive Additive Lexicon (CAL), association of key terms, their position in the CAL Content Tree with additive information like probabilistic indexed weights, the grading Strategy of the content under investigation using sentence level path tracing.**

Keywords: ***Making through, Investigative Content, Reference Content, Tagging, Probabilistic Indexed set weights, sentence level path tracing.***

1. INTRODUCTION TO THE CONTENT GRADING SYSTEM:

CGS takes input as innovatively prepared content, which should be graded and based on the grading the following decision is made whether 1). The content violates the copyright ownership, 2). The content partially pirates or steals the information available in several copyrighted contents, 3). The content is original content, 4). The content has valid meaning and proposes a new idea. Content Grading Systems (CGS) not only grades the contents but also upgrades and patterns by itself without altering the existing base structure. Figure1 illustrates the basic black box model of CGS.

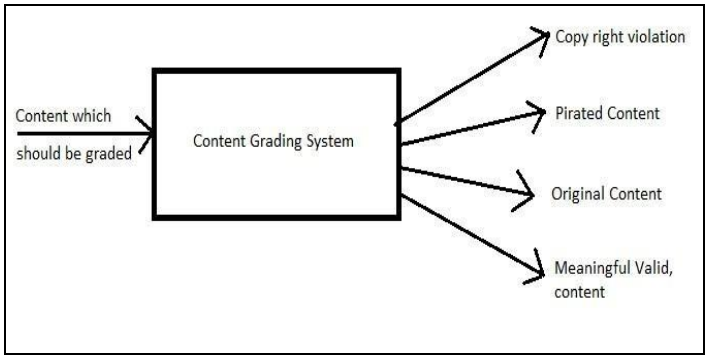

Figure 1. Black box model of Content Grading System

2. BASIC FRAMEWORK OF CONTENT GRADING SYSTEM

2.1. COMPOSITION OF CONTENT GRADING SYSTEM

CGS collectively and interactively uses several components in a defined order to effectively grade the content. Following is the list of components: 1). Content, 2). Content Tree, 3). Cognitive Additive Lexicon, 4). Morpheme extractor and semantic tagger. 5). Sentence level path tracer, 6). Content Tree Generator, 7). Decision maker, 8). Content Grader. Figure 2 illustrates the various components:

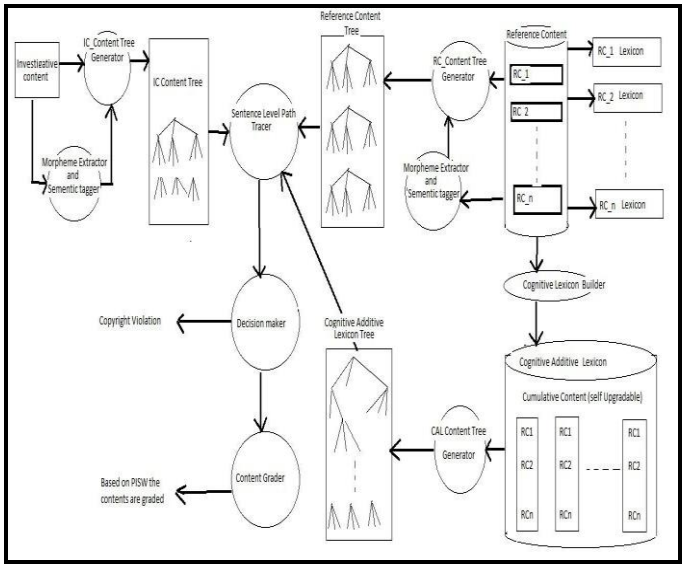

Figure 2. Components of Content Grading System

2.1.1. CONTENTS

Nature of contents can be any one of the following: 1). Text, 2). Images, 3). Audio, 4). Video, 5). Behavioural contents. Contents contribute ideas and have entities and their associated relationship. Content can be of two types:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

1). Investigative content (IC), 2). Reference Content (RC). Contents can be visualized as containing actors and their actions. Notational representation of contents is as follows:

Operand Operator Operand

For example, $IC_1 = E_1 \infty E_2$, denotes E_1 and E_2 as actors, ∞ as the action done by the objects / relationship between the entities. There is no binary or ternary relationship restriction.

2.1.2. CONTENT TREE

Content tree is the pictorial representation of notational representation of the content. Content tree is made up of nodes and lines connecting them. Nodes are the entities E_1, E_2, ..., E_n. Content trees should have at least one node and to the maximum "n" nodes and "m" levels, where "m" cane never be the greater than "n".

For example, the IC1 = E1 ∞ (E2 ∞ E4 ∞ (E5 ∞ E6) ∞ E6) is pictorially represented in Figure 3. Figure 4 illustrates the content tree as a simple Tamil content.

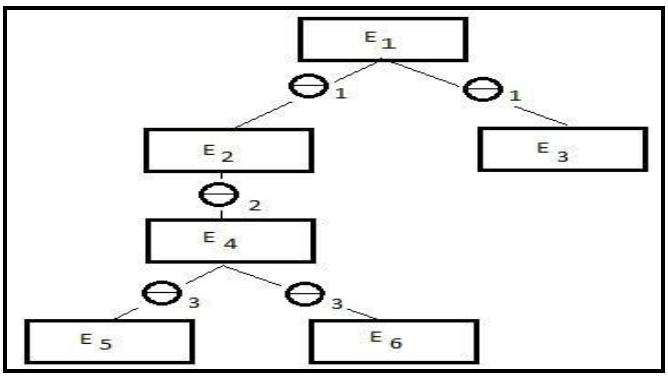

Figure 3. Pictorial representation of content tree

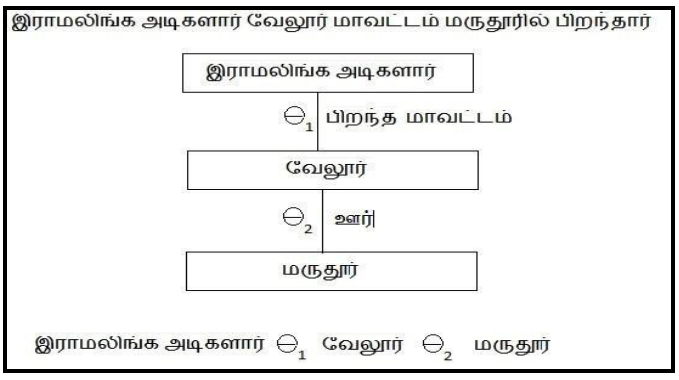

Figure 4. Sample content tree for a Tamil content

2.1.3. COGNITIVE ADDITIVE LEXICON (CAL)

Reference content (RC) is the content specific to different group of authors. The lexicon format of a particular Reference Content (RC_i), where i = 1 to n, is <category-1, category-2, ..., category-m>, where each category belongs to the tag of a native language under inspection.

CAL consists of

$$\begin{bmatrix} category-1,RC-1 & category-2,RC-1 & .. & .. & category-n,RC-1 \\ category-1,RC-2 & category-2,RC-2 & .. & .. & category-n,RC-2 \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ category-1,RC-n & category-2,RC-n & .. & .. & category-n,RC-n \end{bmatrix}$$

The entries in the lexicon have the following structure: {name of the entry, category, clitic / inflected form, additive info, possible follow-up entries, possible relations}. Cognitive additive lexicon enhances the details of the lexicon every time the content gets graded. So growth or population of entry in CAL is dynamic.

2.1.4. CAL – CONTENT TREE

CAL content tree is generated from the entries of CAL. Unlike RC-content tree where each RC-content has its own individualized content tree. Cal tree is a combination of the entire RC content tree. The root node of the RC content tree will never change. The root node is the central ideal entity, whereas in the CAL content tree, the root node can become a child node of some other node. CAL content tree combines the information of several RC content trees. This scenario is illustrated in figure 5.

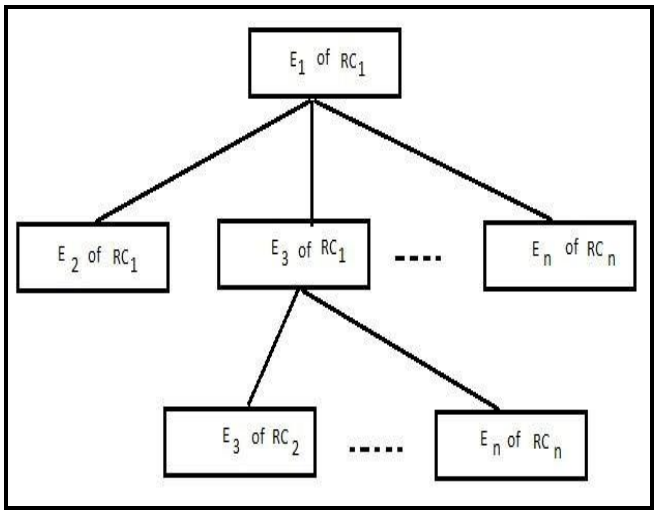

Figure 5. Cognitive Additive Lexicon – Content Tree

2.1.5. MORPHEME EXTRACTOR AND SEMANTIC TAGGER

Morpheme is the meaningful single most token of the word, which can be possibly inflected. In Tamil language all type of words irrespective of whether the word is a noun, verb, adverb, pronoun, preposition, they can be inflected. Using 2-level morphological analyser like PC-Kimmo, morphemes are extracted and checked against the lexicon. If the morpheme is available in the lexicon, it is tagged and added to the RC content tree or to the CAL content tree. Otherwise (i.e. the morpheme does not available in the lexicon) the corresponding tagging should

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

be identified. This is accomplished using the semantic tagger which uses Context Free Grammar (CFG). Semantic tagger identifies the actors and actions performed by them.

2.2. INTERACTION OF VARIOUS CGS COMPONENTS:

"Morpheme extractor & semantic tagger" works along with "Content tree generator" to generate the equivalent content tree. "Cognitive lexicon builder" adds the unavailable new words encountered in the reference lexicon (i.e. from RC_1 lexicon up to RC_n lexicon). "Sentence level path tracer" compares the content trees of IC and RC and also with CAL to check the originality of the content. "Content grader" takes the outcome of "Decision maker" and based on the relative probability associated with each node (i.e. Indexed set weights – ISW) and with the traced node in CAL content tree, grades and ranks the content.

3. PROCESS OF CONTENT GRADING SYSTEM

Case1: Deciding over the copyright violation;

Content tree of the Investigative Content (IC) and content tree of various "Reference Content (RC)" are compared. Comparison is normally done by tracing the IC content tree nodes with the RC content tree nodes. After node-by-node trace in all the of the RC content tree, it is decided that copyright is violated if any of the RC content tree matches exactly with the IC content tree. The degree of exactness is set by content grading system. After identifying the originality, the authors of the respective RC content tree is alerted / intimated about the violation.

Case 2: Partially pirated content

Sometimes a research idea is stolen not from a single copyrighted owner but collectively and in scattered manner from multiple RC contents. Using the RC content tree alone it is not possible to identify the case of contents stealing. Using a CAL content tree it is possible to find out this type of content stealing. In such cases exact match between RC content tree and IC content tree is not possible, but CAL content tree will reveal the content stealing.

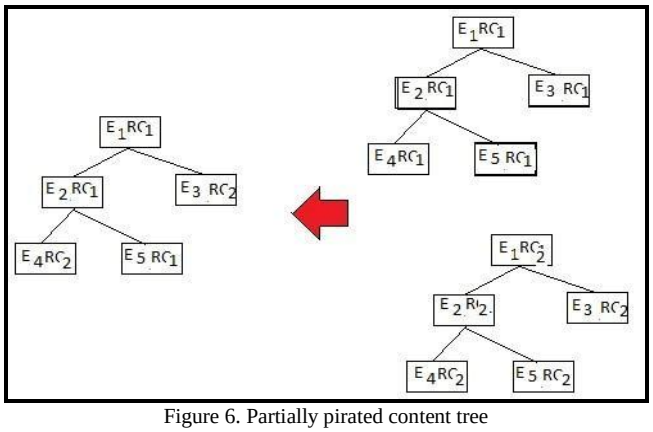

Figure 6. Partially pirated content tree

Case 3: Original content

Content of IC is checked against with every RC content tree, if no exact match is found, it denotes that the copyright ownership of the Reference Content is not violated. But IC has to be checked with CAL content tree to reveal the "Partially Pirated" status of the "IC". CAL content tree when checked against IC content tree will reveal the relevance of the content. Original content is then passed on to content grading.

Case 4: Meaningfulness of the content

An original content is graded with high rank among other original content, if it conveys relevant, innovative ideas which by the way extends the available knowledge (i.e. making the available knowledge a matured one). Content is said to be containing relevant, if there exist a path for the equally likely content in CAL content tree.

Content is said to containing innovative ideas if it adds more child nodes to the existing nodes but not orphan nodes. If the "IC" adds new nodes in the deepest levels of the existing CAL content tree, then it is said to innovate the existing knowledge and should be higher precedence over the equally likely "IC contents".

If the "IC" adds new nodes in between in cluttered manner in the existing CAL content tree, then it also proposes new idea, but the idea is not concrete and is not concentrating over some centralized topic.

3.1. INDEXED SET WEIGHTS (ISW)

With every node in the CAL content tree, a relative probability is associated. The relative probability is factor telling how far the newer idea is extended. Relative probability varies dynamically as new nodes are added to the existing Cal content tree. Indexed set weights are the relative probabilities.

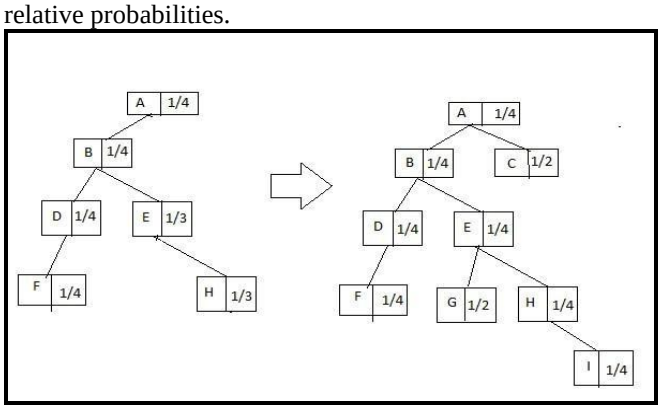

Figure 7. Example for the tagging of the Indexed Set Weights

In figure 7, it denotes a scenario of associated ISW with the nodes. It depicts how the ISW changes dynamically when new nodes are added to the existing content tree. Initially the tree contains 6 nodes, in which ABDF forms the path1 and ABEH forms the path2. Path1 has four nodes so all the nodes are attached a weight of (1/4). Path2 has three nodes originating from the "B", so they are attached a weight of (1/3) with them except the node B which is still (1/4) – this is indicated that B has a choice to make in the next node. And also denotes that D is attached first to B than E which

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

is attached later. Notice in figure 7 after the insertion of G to E and node I to H the ISW is dynamically updated without violating underlying restriction.
For the equally contents which needs to be assigned ranks, the path in the CAL content tree is traced, after complete trace in the CAL content tree, the combined indexed set weights is calculated. The one which gives least weight conveys high entropy of information and it is ranked higher, compared to other contents.

4. COGNITIVE GROWTH INDICATOR (CGI)

The content of CAL grows dynamically as and when new contents are graded. CAL content tree is used by the sentence level tracer to check for originality and meaningfulness. Initially the CAL content tree is small, because the CAL is in the process of learning curve state, by way of adding new nodes. But at some particular time, when available information in the CAL is adequate, very few nodes are updated and added. At this stage it indicates that the cognitive growth is becoming much matured. Maturity of content indicates that CAL content tree will serve and aid in better grading of the IC content.

4.1. EXPERMINETAL RESULTS

A cognitive growth line chart is prepared to determine the maturity of the CAL content tree. X-axis is taken as "number of entries in the CAL", Y-axis is taken as the "Number of paths traced in the CAL content tree". In figure8, it is evident that the curve behaves like a learning curve of a child slowly acquiring the cognitive skills as it grows and attains the adulthood (maturity), but the learning is a continuous process. At this point "A" – the cognitive skills stagnates. At point "B" – cognitive skills acquired till that point is used. At point "C" – high cognitive knowledge thrashing takes place. These results are obtained by taking chosen CAL contents of related ideology. X-axis has samples of approx. 1500 words.

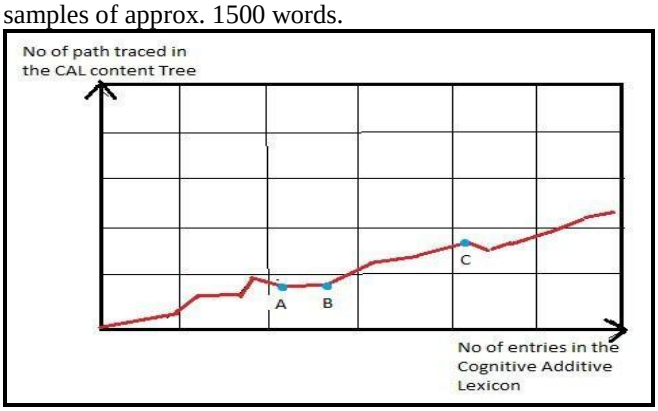

Figure 8. Cognitive Growth Indicator (PC-Kimmo)

5. FUTURE SCALE UP AND CHALLENGES

5.1. QUESTION ANSWER EVALUATION SYSTEM (QAES)

The same CGS can act as "QAES" system if applied the reverse way. If CGS rejects the content under investigation, then "QAES" accepts the content. If CGS partially accepts the IC content, then "QAES" also partially grades the content. If CGS accepts the content, then "QAES" rejects the content.
The contents of quantitative aptitude question are called as QAC. The QAC is transformed in to QAC content tree. Leaf nodes normally contribute the value. Based on the previously stored pattern, the content gets evaluated. The particular QAC is added to the CAL. Based on this the existing pattern is updated.

5.2. CONTENT SPECIFIC SEARCH

As the cognitive additive lexicon contains the indexed set weights, it is possible to give cross content information. Automated innovative ideas can also be generated.

5.3. CHALLENGES IN CONTENT GRADING SYSTEM

The cognitive part of system lies in CAL. Initially the lexicon is in the learning curve and CAL does this by way of rigorously populating its entries. While populating it also takes in to considerations, the relationship between entries of different categories. When the CAL content tree attains the maturity level (i.e. considerable value of cognitive growth indicator), the sentence level tracer must also be algorithmically modifies such that it does not take long time. As the CAL content tree becomes larger, the information available is also situated in a scattered manner. So neuro-fuzzy systems are to be used to extract innovative ideas and to make CGS system a efficient one.

6. CONCLUSION

The ideology behind this proposed research idea for content grading system (CGS) is taken from the influence human brains cognitive growth from inception to child to adult to elderly. As the human brain evolves to new idea and accustoms to the changing scenarios, the same is applied to the structure and behaviour of CAL. As the human brain takes decision collectively based on the available past events, the CAL content tree is used to take collective decisions. But sometimes the human brain itself is under immense confusion of what to decide mainly the content thrashing, the same happens to the content grading system. In order to make the CGS work effectively, it has to first populate the CAL with all relevant information. This ideology can be extended further in to the development of behavioural grading (i.e. Lie detector, image content grading, Audio & video content grading, motion capture detection)

ACKNOWLEDGEMENT

We are very much thankful to the family members, colleagues and management for patience and constant support in all walks of ours and helping with suggestions wherever possible.

REFERENCES:

[1] Dr. Andreas Mauthe; Dr. Peter Thomas (2004). Professional Content Management Systems: Handling Digital Media Assets. John Wiley & Sons. ISBN 9780470855423.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 03, March-2019

[2] PC-Kimmo Reference Manual - A two-level processor for morphological analysis version 2.1.0, http://www.ai.mit.edu/courses/6.863/doc/pckimmo.html
[3] "A reference Architecture for semantic content management" – Fabian Christ, Benjamin Nagel, s-lab, software quality lab, university of Paderborn, warburgerstr-100.
[4] "Two level Morphological Analyzer for english" – lauri kartunnen, Kent Wittenburg, http://www.stanford.edu/~laurik/publications/archive/kimmo/kimmo-english.pdf
[5] Document management system – john Kullen et al, Abstract of US Patient document No: 5893908.
[6] Document Management Systems – Tomomi maruyama et al, US Patient Application Publication, Publication number: US 2003/0046351 A1, Publication date: 11 – Feb – 2002
[7] "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles" - Carol Friedman ,Pauline Kra, Hong Yu, Michael Krauthammer and Andrey Rzhetsky - Proceedings for the Ninth international Conference on Intelligent Systems for Molecular Biology