PERSPECTIVE

52(245), May 1, 2016

ISSN 2278–5469 EISSN 2278–5450

Discovery

Big data methods for computational Tamil linguistics with emotion oriented semantic pattern mining

Chandrasekaran S[1], Chithra S[2]

1.Dean / Academic & Research, Sri Ranganathar Institute of Engg. & Technology, Coimbatore., India; Email: chandrasekaran_s@msn.com

2.Associate Professor / Tamil Department, Bharathiyar University, Coimbatore., India; Email: chithubu@gmail.com

Article History

Received: 06 March 2016

Accepted: 12 April 2016

Published: 1 May 2016

Citation

Chandrasekaran S, Chithra S. Big Data Methods for Computational Tamil Linguistics with Emotion Oriented Semantic Pattern Mining. *Discovery*, 2016, 52(245), 979-984

ABSTRACT

The objective of the paper is to propose a poetic big data analytic method towards the computational linguistics on Tamil Bharathiyar poems. The existing methods are performing good analysis from the text and data input from the e-books and web resources but lagging in collection and construction of higher level dictionaries of patterns within the poetic context. The big data method will be helpful to retrieve the view of the poet with respect to social and cultural information from the literary texts and emotions. It is scalable to any Tamil poems and E- transactions based on the age of that phrase or poem and the location of the player or poet respectively. The multiple entities within a single paragraph and multiple paraphrases for a single entity mixed with emotions are the major issues in the proposed research work. Byzantine Hypertext Attribute Relationship Augmented Type Intelligence (BHARATI) algorithm is proposed to form specific semantic patterns of the poet for that specific poem and also for future Internet Tamil on Devices (IToD) technology.

Keywords: Big Data, Computational Linguistics, Tamil Poems, multiple paraphrases, Emotions, semantic patterns, Internet Tamil on Device.

1. INTRODUCTION

The application of big data analysis methods for computational linguistics has been exercised and produced historical and social cultural predictions. In the linguistic application context, many literatures have been considered and new remarkable information are being unearthed from the poetic space. Distributed file systems and data streaming technologies are applied in the linguistic problems of corpora mining and machine translation involving advanced specific semantic patterns [1]. The pattern oriented methods are of two categories. One is the direct application of the pattern through identification of the player and the concept and stored in the knowledge base approach. The challenge in the knowledge base approach is the named entity disambiguation (NED) and the word sense disambiguation (WSD). The other works focused on the relationship between the noun, class and category items as in the YAGO. The revised YAGO project introduces the relational patterns. It focused on the pronouns, the context with which the nouns are coined and an info box to hold all other related texts. The earlier work on big data methods for computational linguistics focused on building large dictionaries of name and paraphrases for entities. The work proposed a syntactic ontological lexical (SOL) patterns and pattern synset of equivalent patterns [2]. These techniques solve the issues due to frequent sequence mining and graph algorithms and co-occurrence statistics. For an effective analytics on the big data, a new type of pattern, called GTS (Geographic-Temporal Semantic) Pattern, to represent users' frequent movement behaviors by considering all the three kinds of user intentions mentioned previously, i.e., geographic-triggered, temporal-triggered and semantic-triggered intentions was presented [3]. The big data derived or filtered from poems written in Tamil language poses a serious challenge in data collection and preprocessing. Big Data characteristics are given in HACE theorem which represents big data sources as heterogeneous, autonomous, complex and evolving. The phases include the Tamil incoming poems or poetic texts, acquisition and transliteration, extraction using patterns, aggregation with emotional features [5].The various big data sources are diverse in the context of schema, data storage model and its representation The computational linguistics has come to make use of a diverse portfolio of data processing methods, relevant to corpus processing, psycholinguistic modeling and formal grammar construction, which have ready applications to more traditional domains of language analysis [6].

The organization of the paper is as follows: Section II explores the need for big data methods for tamil poetic text anysis in terms of the feasibilities and policies of the government. Section III introduces the methodology of the poetic text or corpora acquisition from the transliteration and filtering. Section IV proposes the need for emotion flow and semantic patterns identified and collected methodology for the big data analytics performance enhancement. Section V proposes BHARATHI algorithm to retrieve and mine the semantic patterns detected within the selected hypertexts of Bharathiyar Tamil poem Kuyilpaattu of Tamil Poet Bharathiyar. Section VI concludes the limitations of the submitted work and future works on Tamil on Devices and Tamil computing in Big Data Methods.

2. BIG DATA FOR TAMIL COMPUTING AND TAMIL TEXT ANALYTICS

In the state of Tamilnadu, Tamil is the language of the government by the people and for the people. Even though the Tamil language is not given that much importance in the school and college level teaching and learning, the language is most predominantly used in most of the rural and urban areas. The challenge in using Tamil for Big Data is of many folds. One is the proper understanding, usage and pro-efficiency in Tamil language with all its grammar and the next one is the acceptance policy for using Tamil texts for our analysis with proper contexts and spoken environments. The work by R.S Vignesh Raj and his team explored the challenges of big data methodologies from the Tamil language perspective by considering the policy of the Tamilnadu government, literacy level of the Tamil people to get benefitted from the big data methods [4].

Big data is not language specific and hence the big data in Tamil poems can also be analyzed as per the standards and algorithms deployed for massive datasets. The correct interpretation of data, database design, meta data and its structure with semantics are the important challenges in the application of big data analysis methods to Tamil poetic texts. The big data analysis is totally dynamic heterogeneous and interrelated. Hence cross checking of conflicting cases, hidden relationships and poetic untrusted exaggerated texts can determine the flow of emotions. The massive or big data analytics of textual data needs specific methods for processing in terms of mining. The mining will be so complex if the dimensionality of the textual data is very high. In the case of poetic text, the issue of emotion and aesthetics of the poet will add to the complexity of the analytic methods. The Tamil text analysis needs parsing, searching for retrieval and mining stages specifically for poetic texts with poet or player's emotions.

3. POETIC TEXT CORPORA ACQUISITION

Original poem in native language is in Tamil and the Word-set is processed in many phases. Document filtering, Unicode Identification, Poetic Normalization, Poem based Classification, Entity Attribute Relationship Identifier, Hypertext Augmenter,

Transliteration and Emotion Flow Analyzer. The input pure poetic Tamil text has to be fed to the appropriate transliteration and to be parsed. The equivalent English is also shown in the boxes below.

காலை யிளம்பரிதி வீசுங்கதிர்களிலே நீலக்கடலோர்
நெருப்பெதிரே சேர்மணிபோல்
மோகன மாஞ்சோதி பொருந்தி முறைதவறா வேகத்திரைகளினால்
வேதப்பொருள்பாடிவந்து தவழும் வளஞ்சார் கரையுடைய செந்தமிழ்த்
தென்புதுவை யென்னுந் திருநகரின் மேற்கே, சிறுதொலைவில் மேவுமொரு
மாஞ்சோலை;

Righteousness will match the mokana mancoti yilampariti morning vicunkatirkalile nilakkatalor neruppetire cermanipe vekattiraikalin creepy but come vetapporulpati valancar shore west of the centamilt tenputuvai yennun tirunakar, in small mevumoru Manjolai;

கண்டதொரு காட்சி கனவு நனவு என்றறியேன்
எண்ணுதலும் செய்யேன் இருபது பேய் கொண்டவன்போல்
கண்ணும் முகமும் களியேறிக் காமனார் அம்பு நுனிகள்
அகத்தேயமிழ்ந்திருக்க

Kantatoru View dream consciousness to know
the counting of the eye, face and twenty ghost kontavanpe kaliyerik akatteyamilntirukka kamanar arrow tips

Many commercial development kits like Linguistic Development kit or Multi Lingual Parser or Corpora Linguistic Warehouse can be used to in all these phases. According to the online translator at http://en.eprevodilac.com/prevodilac-tamil-engleski, the kuyilpaattu poem with line number 76 to 79.

The entities that are more crucial in any poetic text analysis are the attributes within the poetic text. Like the name of the poet, scene over which the poetic lines are developed, the nature of statements whether they are self-talk or a conversation with other players of the poems in addition to the the poet experience within the poem or in his life period as given:

<Poet Name, Poetic Scene, Poet Self Address, Poetic Conversation, Poet Experience>

<கவிஞர்பெயர், கவிக்காட்சி, தமக்குள்பேச்சு, உரையாடல், கவிஞர் அனுபவம்>.

Within the poem, the primary player of the poem, the player's features and the players action and roles and the quantitative indication of the achievements by the player or by any other object and special information to add more aesthetics and rhyme to the poem itself are represented as shown below:

<Poetic Player, Player Features, Poetic Deed, Poetic Object Count, Special Information>

<கவி நாயகன், தன்மைகள், நாயகன் செயல், கவிப்பொருள் அளவு, பாடல்செய்திகள்>

For example, the above Bharathi poem named Kuyilpattu is taken as the linguistic corpora for the submitted research work. The poetic text are clustered based on the sequence and classified based on the poet's creative character and paraphrases.

Table 1 Parsed Poetic Hypertexts in source language

< கண்ட, காட்சி> ஒரு < கனவு, நனவு> என்று
<அறியேன்>
< எண்ணுதலும்> < செய்யேன்>
இருபது < பேய், கொண்டவன்> போல்
< கண், முகம்> < களிப்பு> ஏறி
< காமனார், அம்பு, நுனி> < அகத்தே>
<அமிழ்ந்திருக்க>
< கொம்பு, குயிலுருவம் > கோடி, பலகோடி
ஆய்
< உலகம், தோற்றம்> எல்லாம் உற <
சென்றே>
< மனை, போந்து> <சித்தம், தனது> அன்றி
< நாள் > ஒன்று < போவதற்கு>
< நான் > அனைத்தும் <பாடுபடு> <
படுதல்>
<தாளம், தறி> <படுமோ,
படுமோ>
< யார்> <படுவார்>

Relationship Augmented Type: The byzantine hypertext are classified based on the metaphors, analogies, grammar ,words, sequence of words, rhyme and the main focus or the concept with which they are created. The relationships that are derived from the above text clusters, it is possible to arrive at machine readable classes of texts and their importance as,

<Poetic Style, Parable, Metaphor, Text, Sequence, Theme, Rhyme, Rhythm, Aesthetic>.

Table 2 Tamil augmented things and their Types

1. காட்சி வருணனை : வருணனை: பாத்திர வருணனை
2. உவமை
3. உருவகம்
4. சொல், தொடர், வினா – அடுக்குகள்
5. தத்துவப் பின்புலம்
6. குயிலிற்குத் தரப்பெறும் அடைகள்
7. கவிஞனுக்குரிய விளிகள்
8. கவிதையழகு, தொடை, ஓசைநயம்

4. EMOTION PATTERNS AND EMOTIONAL PATTERN FLOW PROGRAMMING

An emotion is a complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response (Hockenbury & Hockenbury, 2007). The ability to recognize, understand and manage self emotions and or emotions of the others may be called emotional intelligence. In NLP corpora of actual human originated text, spoken or written, are the basis for identifying structures by applying statistical methods. This requires a fundamentally different view on text in contrast to what qualitative oriented researchers are used to [8]. The emotion can be computational represented as a chain of implications as a set of context, cause and conversation shown below: Subjective Experience → Physiological Response → Behavioral Expression. The emotions in the poetic text can be detected in a machine readable form through basic text analysis.

Rhythmic Negated Text (RNT) : The poetic text follows a similar unicode or rhythm with a negation in explaining its theme. The examples are as follows:

1. "காதலை வேண்டிக் கரைகின்றேன் இல்லையெனில்
   சாதலை வேண்டித் தவிக்கின்றேன்"
2. "காதலோ காதலினிக் காதல் கிடைத்திலதேல்
   சாதலோ சாதலெனச் சாற்றுமொரு பல்லவி"

Anoymous Repeated Text (ART): In the poetic creation, the poet will attach different tags to the same player or object based on the emotion flow of the poem. It is based on the playing concept and his creativity. The semantic pattern is decided based on the poetic instance or specific scene. The following tags are attached to the poetic primary player. Kuyil" in the Bharathiyar's song.

குயிலிற்கான அடைகள்
1.பெடைக்குயில் 2. ஒற்றைக்குயில்
3. மாயக்குயில் 4.கானக்குயில. 5. சின்னக்குயில்
6. பிள்ளைக்குயில் 7. நீலக்குயில் 8. சிறுகுயில்.

The poetic entity may be the poet herself or himself or the subject with which he or she is referring many physical or virtual objects created by himself. The features of those objects and the sequence with which they are referred will be useful in declaring her or his ideas with emotions. An entity attribute emotion relationship diagram will be useful to find out the basic inherent emotion of the poet in that poem.

Hypertext Augmented Type (HAT): In this category there are two methods based on the nature of the emotion expressed in the poem by the poet in that scene. They are SQT and ALT.

Successive Questioning Text (SQT): The poetic hypertexts or corpus can have successive questioning style to reveal the "Sadness" emotion or mood of the poet.

நாளொன்று போவற்கு நான்பட்ட பாடனைத்தும்
தாளம் படுமோ? தறிபடுமோ? யார்படுவார்?

Augmented Linked Text (ALT) : The texts are interlinked in order to bring aesthetics and rhyme in the poem. These can be parsed with ALT criteria to expose the metaphor used by the poet.

1. வீணை போன்ற உள்ளம் .....
உள்ள வீணை .....
2. வீடு போன்ற உள்ளம் .....
உள்ள வீடு .....

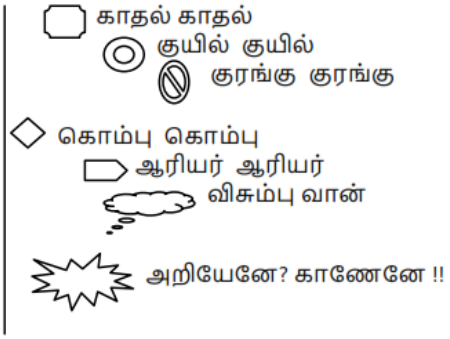
5. BYZANTINE HYPERTEXT ATTRIBUTE RELATIONSHIP AUGMENTED TYPE HIERARCHICAL INTELLIGENCE (BHARATHI) ALGORITHM FOR BIG DATA ANALYTICS

The algorithm is basically a byzantine nature of selecting the hypertexts and their attributes with their inter relationships. The word Byzantine is a term that represents a system or subsystem that has many intricate links that it would be very difficult to separate them as loosely coupled texts. There are so many hypertexts in the Bharathiyar Tamil poem Kuyilpaattu. They have a branching structure which resembles a tree. Hierarchical models of intelligence include the three stratum theory and the Cattell–Horn–Carroll theory can be applied on the augmented types of attributes and relationships of hypertexts selected in a byzantine manner.

1. Select the poem with its source markup format
  1. a. Identify the line numbers or lexical tokens
  1. b. Pickup any line as the starting point after second lines of the Start line
2. Collect and Transliterate the Poem into selected markup
  2. a. Poetic markup language or HTML or XML
  2. b. Identify the entities set noun, verbs, attributes
  2. c. Remove the delimiters like commas, semicolon
3. Parsing in terms of texts of unstructured type
  3. a. unstructured text without syntactic errors

3 .b. Identify the texts and clustering the texts
4. Searching the nouns, special information in the Poetic text
4. a. Retrieval of relationship among patterns
4. b. Semantic Pattern extraction
5. Mining with emotions and number of such emotions
5. a. Emotion related pattern mining
5. b. Decision with the counts of patterns in the entire poems.

The application of this algorithmic suite in multiple phases, it is possible to detect the semantic pattern based on the poet emotion within the poem. The keywords supplied to the information retrieval for further analysis may be based on the nouns or adjectives and the similarities between the scenarios. The output can be in terms of the number of counts of a specific term in the poetic text and the number of times the negated terms takes place in the emotion flow.



6. CONCLUSION

The Big Data in Future Digital Tamil Poetry (DTP) and Tamil on Devices (ToD) needs further research so that the Tamil texts with poetic and non-poetic style can be used in computing. The analysis of big data involves distinct stages of handling the incoming texts. The business or cultural case is to define and declare the various methods in designing and deploying digital poems so that the art-world can entertain itself and the society. The pedagogy of art and literature, computing science and information technology will introduce new means of developing poems with and without images or animations. The hypermedia will utilize the big poetic data analysis methods.

REFERENCE

1. *Data Science and Big Data Analytics*, Education Services, Wiley Publications 2014., pp.256- 259.
2. Gerhard Weikum ., et.al.,," Big Data Methods Computational Linguistics"., Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2012., pp.1- 10.
3. JJ Ching Ying., "Mining geographic-temporal semantic patterns in trajectories for location prediction", ACM Transactions on Intelligent Systems and Technology (TIST), Volume 5 Issue 1, December 2013.
4. R.S. Vignesh Raj., et.al., " Big Data in Tamil: Opportunities, Benefits and Challenges" ICTACT Special Issue on Soft Computing Models for Big Data, July 2015, Volume 05, Issue 04., pp. 1016-1020.
5. Frank Ohlhorst, "Big Data Analytics: Turning Big data into Big Money".,John Wiley & Sons, Inc.,2013, USA..,pp.113-122.
6. Emily M. Bender Jeff Good., Department of Linguistics., University of Washington University at Buffalo., "A Grand Challenge for Linguistics: Scaling Up and Integrating Models"..,
7. Gregor Wiedemann., "Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences"., FQS., Volume 14, No. 2, Art. 23 May 2013., pp.1- 24.