

# Pattern Identification on a Session-Based Application

Sari H<sup>1</sup>, Naseer C<sup>2</sup>

Student, Computer Science and Engineering, GEC Palakkad, Kerala, India<sup>1</sup>

Associate Professor, Computer Science and Engineering, GEC Palakkad, Kerala, India<sup>2</sup>

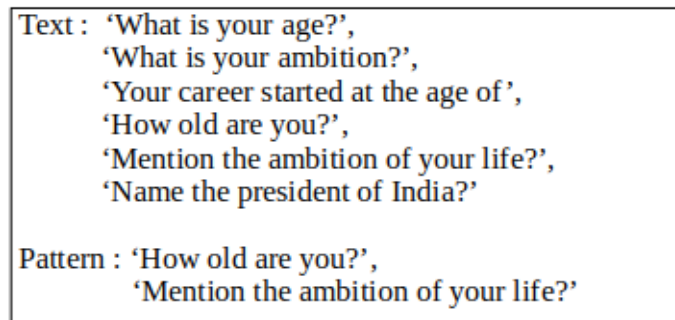
**Abstract:** Pattern identification in texts refers to the identification of repeating texts from set of sentences. Patterns are automatic discovery of regularities present in data through the use of computer algorithms. There is limited research carried out for such identification of patterns. The input to the system is first gathered and is then cleaned to remove the noisy elements present in the data. After cleaning the data, the similarity of the elements present in data is identified. The similar elements are grouped into segments and these segments are then analyzed to check whether repeating elements are present in the data. From this data, the necessary repeating insights are extracted which are the resulting pattern. The detection of patterns of any real world entity or substances of text or any other source is a difficult task for humans as well as for machines. It may be a time-consuming task if the detection of such patterns are done by the human. Also, human supervision is unable to deal with large quantities of data as there will be 'n' number of patterns. Therefore, automatic identification of such repeating texts has become an urgent need. For identifying patterns, context of text accompanying repeating sentences is very useful. In this work, pattern identification of text in semantics level is addressed by using ontology. After identifying similar sentences, the Sequence-to-sequence model is developed to identify patterns present from set of sentences given as input to the system.

**Keywords:** Pattern, Seq2seq, Ontology, Domain-Specific Words.

## I. INTRODUCTION

Pattern identification is an emerging trend in the field of big data. Pattern recognition is defined as the process in which automated recognition of patterns and regularities in data is found out on a regular basis. It has immense applications in the field of statistical data analysis, signal processing, image analysis, information retrieval, bioinformatics and so on. The semantics of text is also taken into consideration while recognising patterns among sentences. The semantic of text is considered by using ontology. The pattern of text are identified by passing it to Encoder-Decoder architecture [1].

Patterns are identified from set of sequences for the use in recommendation systems. The sequence of a pattern followed by individual helps to recommend it to a new user in providing recommendations. The three main models [2] for pattern recognition are statistical model, syntactic / structural model and template matching. The statistical model identifies whether the specific piece belongs (for example, whether it is a cake or not) to which entity and makes use of supervised machine learning technique. Syntactic model defines a more complex relationship between elements (for example, parts of speech) and makes use of semi-supervised machine learning technique. Whereas Template Matching model match the object's features with the predefined template and identify the object by proxy. One of the common use case of such a model is plagiarism checking.



Text : 'What is your age?',  
'What is your ambition?',  
'Your career started at the age of',  
'How old are you?',  
'Mention the ambition of your life?',  
'Name the president of India?'

Pattern : 'How old are you?',  
'Mention the ambition of your life?'

Fig. 1 Example of pattern from set of texts

In this paper, how the patterns are identified from set of input queries is found out considering the semantic information of queries queried. It is carried out by the use of ontology and Seq2seq model. There is no proper method available for

semantic relationship between different texts for domain specific data. A different method of finding semantic relationship between texts of domain data is established by use of ontology. The similar texts will have same IDs and unsimilar texts will have different IDs. The IDs are the inputs to the Encoder-Decoder model. Seq2seq is a Encoder-Decoder framework which convert a set of sequences from one domain to another domain. The input to the model will be combination of IDs which have 'n' occurrences of different IDs. Fig.1 shows an example of patterns from set of inputs and Fig.2 shows an example of no pattern from set of inputs. From the Fig.1, the ID will be assigned either "123124" or "abcabd". This ID is passed to the Seq2seq model as input to the system, which produces the repeating pattern of output as "12" or "ab". The second example shows the absence of patterns from set of inputs. If there is no semantic relationship between texts, in such cases there will be no input to the Seq2seq model and hence no repeating pattern can be found out. The working of pattern identification from texts is discussed in section 3.

Text : "How old are you?",  
"What is your name?",  
"What is your annual income?",  
"How much do you spend every month?"  
"What about your savings"

Pattern : No pattern found

Fig. 2 Example of no pattern found from set of texts

A. *Applications:*

1. *Applications of Pattern identification of text documents:*

- Knowledge mining (concept) of repeating behaviour
- Information retrieval
- Summarization
- Recommendation systems
- Story line generation

2. *Application of Pattern identification in Social media platforms:*

- Instant detection of user behavioural patterns
- Summarizing public opinions

The section 2 includes various methods of pattern identification. Section 3 includes semantic analysis of texts for domain-specific data and Section 4 discussing about Seq2seq model of identifying patterns.

## II. DIFFERENT METHODS FOR PATTERN IDENTIFICATION

This section deals with various methods for pattern identification. These methods mainly focus on the pattern identification in sentences.

A. *Methodology-1 Heling Jiang et. al 2016*

This methodology uses the concept of machine learning algorithms along with graph theory [3]. The data and pattern classification is used to classify each item in a set of cluster of data into one of predefined set of groups. Here, a binary classification approach of data is carried out. Different methods like Decision Tree (DT) based Approach, Bayesian Network (BN) based Approach and proposed method was carried out.

B. *Methodology-2 Bhaskaran Shankaran et. al 2015*

The methodology developed effective pattern matching technique [4]. It finds out the patterns and then evaluates term weights according to the distribution of terms extracted from the discovered patterns. It also solves the misinterpretation problem by considering the influence of patterns from the negative training examples to find out the noisy patterns and also tries to reduce the influence of noise for the low-frequency problem. The input documents are pre-processed and fed to the pattern taxonomy model. This model effectively solves the problem of polysemy and synonymy present in texts effectively.

C. *Methodology-3 Qin Wu et. al 2009*

This model is constructed based on both information about frequency and position of keywords present in the text [5]. First, set of relevant keywords are identified and then the relative distances of the keywords with a document is



aggregated. This is used to construct a weighted directed multi-graph. It generates vectors for each document in the high dimensional feature space which denotes the signature vector which determine similarity values for pair of documents. And the documents are classified using Quasi-Clique Merge clustering method. It clusters the documents based on special feature of multi-membership clustering. This approach was tested on the detection of fraudulent emails written by the same person, and also on plagiarized publications.

#### D. Methodology-4 Minhua Huang et.al 2009

In this method, a probabilistic graphical model is developed to recognise the patterns present in texts [6]. This model is derived from probability function for a sequence of categories given a certain sequence of symbols. The categories include the NP chunks and the semantic role of the sentence. Semantic role labelling(SRL) is the process of assigning labels to words or phrases present in that sentence which indicates the semantic role of the sentence which is similar to that of an agent, goal, or result. The newly developed model has its own mathematical representation different from existing graphical models such as CRFs, HMMs, and MEMMs.

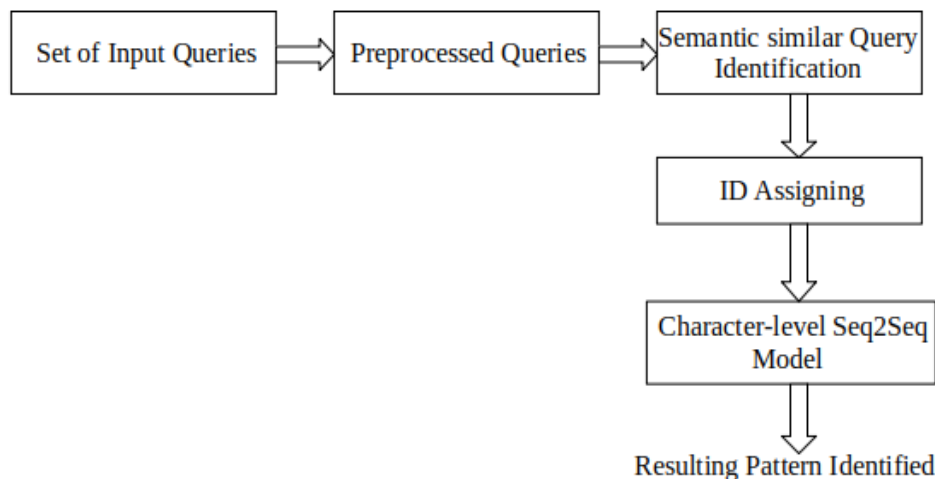


Fig. 3 System Overview

### III. SEMANTIC ANALYSIS OF TEXTS USING ONTOLOGY

System overview of the pattern identification is shown in Fig 3. After getting set of queries to the system, each of the queries is pre-processed and passed to semantic analysis module. Queries with similar semantics are provided with the unique IDs and others with different one. The combination of set of IDs is passed to the Seq2seq model to produce the resulting pattern. The semantic analysis of text data for a particular domain cannot be directly applied by the use of existing semantic matching techniques like Universal Sentence Encoder of Google, InferenceNet of Facebook and so on. To find semantics information of domain-specific data, there is no proper method available, which is the drawback of the system. This problem is addressed by the use of ontology. Ontology is a formal description of knowledge as a set of concepts within a domain and the relationships that exists between the concepts. The benefits of using ontology are automated reasoning about data, coherent and easy navigation among users, easy to extend to different domains. After pre-processing the input queries like finding n-grams of each query and retrieving the Domain-Specific Words (DSWs) from ontology, all the labels of the DSWs is considered as a whole. Those queries with identical DSWs is grouped together and assigned similar IDs. If a new input is queried other than the existing ones, new ID is assigned to it. It is clear from the example shown in Fig.1. These generated IDs are the input to the Encoder-Decoder model for identifying patterns from set of input texts.

### IV. SEQ2SEQ MODEL FOR IDENTIFYING PATTERNS

Seq2seq model is a machine translation task which passes input from user to encoder and decoder produces the resulting pattern of characters as output. Character-level Encoder-Decoder model is developed for identifying patterns. Dataset is developed for this model. It contains combination of sequence of characters as input to the model. It is trained and corresponding pattern is retrieved from decoder. The encoder encodes the characters in a sequence and when the last word is read, it passes its internal hidden state to the decoder, which then starts generating the output sequence. Much of it is similar to a language model and considers the probability distribution to choose the next word in the sequence of input. A greedy search is used which will select the next word using the highest probability in the softmax layer.

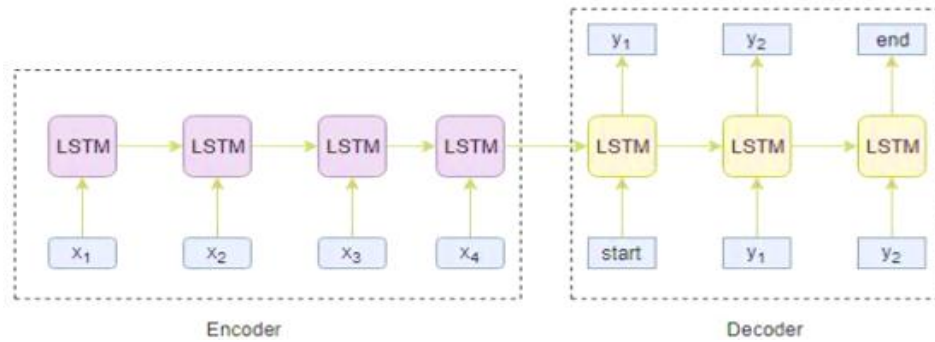


Fig. 4 Encoder-Decoder model

Consider the input sequence to pattern identification is set of combination of different characters. Decoder contains 2 tags ('start' and 'end') to identify the start and end of the decoded output. This model involves text inputs of Embedding followed by a hidden layer of LSTM. LSTM is used as it is insensitive to gap length and has ability to control and hence produce good results. The Softmax function produces a vector that represents the probability of a list of classes and uses these to generate most probable repeating characters in output layer.

After producing the repeating sequence of character from decoder, with the help of python dictionaries the IDs are reverted to its corresponding set of queries asked by user as input to the system. As it is clear from the Example 1 that the output is set of queries ("How old are you?" and "Mention the ambition of your life?").

## V. CONCLUSION AND FUTURE SCOPE

This paper focus on the method for identifying patterns from set of input queries. Also, the semantics of each text are matched by using ontology as domain data is considered. The seq2seq model helps to identify patterns of texts and produce good results. In future, the identification of patterns from very large documents considering semantic information of data can be carried out.

## REFERENCES

- [1]. Kyunghyun Cho, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", in Association of Computational Linguistics, 2014.
- [2]. M. Raj, P. Swaminarayan, J. Saini and D. Parmar, "Applications of pattern recognition algorithms in agriculture: a review", in International Journal of Advanced Networking and Applications, 2009.
- [3]. H. Jiang, A. Yang, F. Yan and H. Miao, "Research on pattern analysis and data classification methodology for data mining and knowledge discovery", in International Journal of Hybrid Information Technology, 2016.
- [4]. B. Shankaran, M. Patil, S. Suryawanshi, S. Mandhane and S. Raskar, "A novel approach for text extraction using effective pattern matching technique", in IJRET, 2015.
- [5]. Q. Wu, E. Fuller, and C. Q. Zhang, "Text document classification and pattern recognition" in IEEE International Conference on Advances in Social Network Analysis and Mining, 2009.
- [6]. M. Huang and R. M. Haralick, "Identifying patterns in texts", in IEEE International Conference on Semantic Computing, 2009.