ENHANCEMENT IN GENERALIZED APPROACH FOR PROCESSING UNSTRUCTURED DATA USING SEMANTIC STRUCTURE OF TEXT

Piyush Gusani(B.E.(C.E.)),NGI Junagadh
Prof. Piyush Gohel(M.E.(C.E.)),Prof. Daxa Vekariya(M.E.(C.S.E.)),CE Department NGI,Junagadh

Abstract: We know that most of professional and personal places have computers in today’s life and they contain webpage text documents as well as other offline text documents. Among these documents to differentiate documents of a particular interest manually is a very tedious and time consuming process. There are several approaches available for the above mention problem that is text classification. This paper focuses on enhancement of conventional approach that uses words only for classification rather than semantics of the text. In this paper we use semantic structure of text and then use it for classification process.

Keywords: Text Classification, semantic structure

I. INTRODUCTION

The amount of text documents online as well as offline is increasing day by day. It is very complex problem to manage these text documents, choose some documents of a particular class among hundreds even thousands of documents. About 75% of web page is html page that contains most of text data[2].

This paper focuses on enhancement of conventional approach for text classification that uses Bag of Words approach for classification purpose[2]. Bag of words approach use words for classification purpose without knowing the position of words in document. We know that according to the different position of words in a document meaning or intention of that word may be different. so to improve this we proposed enhancement in conventional approach. In Our enhanced approach semantic structure of text is used for processing of unstructured data.

There are four phase in our proposed approach
a) Reading webpage text from url and check for any update in webpage text.
c) Converting Discourse representation of Documents into graphs and measuring similarity/dissimilarity between document graphs using graph distance measures.
d) Document classification on the basis of graph distance measures.

II. LITERATURE SURVEY

Robert Blumberg and Shaku Atre[1] uncovered the problems and issues with unstructured data. In this paper they have raised issues that come in way while dealing with unstructured data.

Vaishali Ingle[2] proposed an Algorithm for Webpage classification that uses conventional approach like bag of words. This algorithm gives good result when text is simple text without having nagations, implications, prepositions etc but when text document contains complex grammar than it affects on classification of documents.


Jacob[5] ’s book on NLP and use of NLTK in Python is very helpful for converting our theoretical ideas for text processing into practical way.

Hense, Joseph and Uwe[6] written a book of Discourse Representation Theory which is very helpful for understanding the discourse representation and implementing it actually.

III. PROPOSED WORK

From above literature we found two issues
1) To use better option rather than Bag of words
2) Updating of web document text and effect of update in Classification process

To solve above stated issues we proposed an Enhanced approach for Processing Unstructure Data.
Fig. 1 proposed work Architecture.

Phase I
As shown in figure when a text document (webpage/offline text document) comes in the system, it is checked whether it is already processed or not. If document is already processed then check for update in document. If document does not have any updates than use already summarized text for further processing. In case of update in document, extract the updated text and use it for summarization process along with existing summarized text.

Phase II
Now Discourse Representations structure will be generated from the summarized text of documents using C&C and Boxer tool. Summarization is done in a manner that sentences having words with more frequency in document are arranged in top down manner. For example “adam drinks water.” Here, x0, x1, and x2 are the referents and male(x0), water(x1), drink(x2), event(x2), agent(x2, x0), patient(x2, x1) are the conditions. 
\{x0, x1, x2: male(x0), water(x1), drink(x2), event(x3), agent(x3, x1), patient(x3, x2)\} is Discourse Representation of above sentence.

Phase III
In this phase DRS are Converted into graphs using Graph generation tools. For sentence “adam drinks water” graph is generated in a manner that all referents x0, x1, x2 are treated as nodes of graph and conditions like agent(x2, x0) treated as edges of graph.

Fig. 2 Graph generated from DRS for Sentence “adam drinks water”[6]

Graphs generated from different documents are used for measuring similarity or dissimilarity using between pairs of documents using Graph Distance measuring. MCS-minimum common Supergraph and mcs-Maximum common subgraph will be computed.

Above process will be done using two pairs of text documents testing and training data.

Phase IV
By using these Distance matrices the classification will be done using K-NN classifier.

**BRIEF DISCRIPTION OF TOOLS & TECHNIQUES**

A) Python 2.7 and NLTK[3]
Python is an open source object oriented language, especially available for developers and researchers. It provides lots of support packages for different needs of programming. Like for Natural language Processing NLTK is there, Which provides hundreds of classes and function for NLP.

NLTK also have a class named drt that is useful for discourse representations.

B) C&C and Boxer[7][8]
For generating Discourse structures of a document Combinational & Categorical paeser and Boxer tool is available. It is Compitable with python so we can use it easily.
A sequence of sentences is given as input to algorithm than first sentences is analysed and it’s sements work as contex for next sentence. For example Sentence S1 semantics are in K1 than K1 will be contex for S2 and than K1,2 will be generated. This process continues till the last sentence in sequence. As output of this process Discorse Representation Structure of the document is generated.

C)Naivesumpy
It is a readily available python code for tex summerization. It takes input as Text document /Text and generates sequence of sentences having words with higher Term Frequency . In this summerization process frequent words like is, the, a, an etc are not considered.

D)MySql
It is useful for database related operations. It can be easily connet to python programm. In our system the summerized text generated from text documents can be stored in database for further use with help of My Sql.

CONCLUSION
By implementing our proposed enhanced approach we can improve the unstructured data processing with parameters accuracy and time.

FUTURE WORK
After implementaion of our proposed work we will compare the classification results with bag of word approch and mark the difference.

REFERENCES
[3] Steven, Ewan and Uwe, “Natural Language Processing with Python”.