SURVEY ON DATA EXTRACTION AND PRE-PROCESSING TECHNIQUES FOR UNDERSTANDING THE MOOD OF A PERSON

Naveen D. Chandavarkar
Assistant Professor
Dept. of Computer Science,
SPRDP, Govt. First Grade Degree College, Mudgal
Lingasugur Taluk, Raichur Dist., Karnataka, India

Abstract : Data extraction is an integral part of data analysis system, where the large amount of data is pre-processed by some specific Pre-processing technique, so as to use the pre-processed data for analyzing mood of a person who tells the reviews about the products. This paper describes the various methods of data extraction and describes how to analyze the mood of a person using a 3 tuple mechanism that includes a Data Extractor, Mood Analyzer and Final Decider with a set of predetermined moods. The speed at which the data is generated requires efficient techniques to extract and pre-process the data. Tokenization, stop word removal, stemming are the steps involved in pre-processing the data. Data from web page is extracted by the use of HTML structure of the web page and by generating DOM trees. Its accuracy varies from person to person as mood is not a stable quantity but is reliable.

***IndexTerms*: - Mood Analyser, Product reviews, Pre-processed, Tokenization, Predetermined Moods.**

I. INTRODUCTION

Over the past two decades there has been a significant increase in the field of psychology dealing with the study of mood, especially using modern technologies as they advance drastically.

For the purpose of mood analysis the data is extracted from websites and various sources of reviews about a product [1]. Data extraction is either through a static web page that displays the same data or dynamic web page where the data content changes.

Web pages generally contain information hidden. This hidden information contains content from the database or data warehouse that can be encountered on the web. Search engine can be used that results in static or dynamic data. Web data extraction is a process of retrieving the data from the web. The data extracted will be sent to database or to excel page or for processing in another application. In these papers various techniques of data extraction like Depta, DeLa, NET, Fivatech, ExaLG and Xwrap has been encountered [1].

Any form of real world data is incomplete, noisy and inconsistent. These data lack some attributes, so these data needs to be pre-processed. Pre-processing involves data cleaning, data integration, data transformation, data reduction and data discretization. Pre-processing is said to improve efficiency of the mining algorithm on unstructured data.

Many techniques have been developed to assign positive, negative or neutral mood states experimentally using complex text classification methods or unsupervised semantic orientation computation from POS (parts of speech) [14]. Those techniques have helped us understand the relations and insights. However doubts have been arising about the effectiveness and reliability of these standard procedures for mood Analysis.

The problems faced in mood analysis or opinion mining is:

- ❖ Moods and emotions are relative terms, i.e. the meaning of each mood keeps changing from person to person.
- ❖ Emotions last for about 6 seconds only, i.e. if you are going to find mood of a person based on present emotion(s) of that person then it might be inaccurate.
- ❖ Some people are introvert and suppose this person expresses something, they will suppress their views and may be the results become inaccurate.

So to solve these problems and to skip the last problem, i.e. partial reviews, from interfering with the accurate results, a technique is devised.

Importance of mood analysis

As this paper is based on Mood analysis of product reviews, it's very helpful for:

(a) Consumers: Sometimes people are confused and they don't know what to buy and what not to. For that people go online for reviews and comments to find a better product, basically people waste time finding those products with good and better review or comments of a particular product.

(b) Business/Entrepreneurs: Companies or Entrepreneurs always search for business insights of their products or services they have to offer. They are always in search of data analytics for insights like consumer's decision making process, consumer's choices, consumer's wants and needs etc [14][15][16].

II. DATA EXTRACTION TECHNIQUES

Various techniques on web data extraction have been worked on. Two types of data extraction technique are mainly observed as wrapper induction and automatic extraction.

As per the study done by Devika K et. al. [1] on Data Extraction and Label assignment (DeLa) that extracts data and assigns label to the data, is followed by steps during the execution of data:

- Collect the labels of the website from elements.
- Generate the regular expression using wrapper generator.
  - ✓ Data-rich section extraction.
  - ✓ C-repeated pattern.
  - ✓ Optional attributes.
- Data alignment that take place in two phases.
  - ✓ Data extraction.
  - ✓ Attribute separation.
- Assign labels to the data present in the column of the table.

Nested data Extraction based on tree matching (NET) is another technique to extract data from data records as well as nested data records. Initial step here is tag tree construction which is based on lower level nested records. The task of extraction is identified has two observations.

- The list of similar objects describes the data record in a continuous region and formatted by HTML tags.
- Under a single parent node similar data records as per the region is been represented. The traverse ( ) algorithm recursively finds the record nested if the subtree's depth from current level of the node is greater than or equal to three level. Relational table is used to have the extracted data to the user by using output ( ) algorithm.

EXALG is a technique that processes the extraction of templates in two phases i.e equivalence class generator and analysis stage.

The equivalence class generation involves:

- ✓ Differentiating the role of non tag token.
- ✓ Finding equivalence classes.
- ✓ Detection and removal of invalid LFEQs (Large and frequently occurring equivalence classes).
- ✓ Differentiate the role of tag tokens.

In the analysis stage templates are built using LFEQs. LFEQs are large equivalence classes and contains token in large number of pages.

ROADRUNNER is another web data extraction technique discussed by Kristina Lerman et. al. that generates wrapper for web pages of the same class[2]. This technique is efficient for extraction of nested web pages and pages having same structure. Union-Free regular expression is the form of wrapper that is generated by comparing the two pages of the website. The initial page will be the wrapper that will be the wrapper which will be compared with the rest of the pages, based on mismatches, i.e. tag mismatching and string mismatch, the wrapper is generated. String that is mismatched will be replaced by a symbol in the wrapper. Tag that is mismatched can be either the optional field or by iterations.

Study done by Yanhong Zhai [3] on Depta, is a partial tree alignment based extraction technique. The first step in this technique is to identify the data records by dividing the web pages. This step is next level of MDR technique.

Depta is compared with Fiva Tech (As per the study done by Mohammed kayed et. al. [4]) because same task is been shared i.e. data alignment and frequent pattern mining. The loss of data is more possible in Depta as the repetitive patterns are first found and alignment is done later. In the case of Fiva Tech loss of data is minimal as data alignment takes place, and then the frequently repetitive patterns are found. Capability of Depta is compared with Fiva Tech in handling the data i.e. Depta is capable to handle single data record and Fiva Tech is capable to handle single and multiple data record.

III. PRE-PROCESSING TECHNIQUES

As per the study done by Dr. S Vijayarani et. al. [5] on purpose of pre-processing of text data, applications of text mining and other contingencies are highlighted as well. Useful information and textual data can be extracted by the process of text mining.

By using the techniques of pattern matching, information extraction identifies keywords and relationship with in text and converts it into a relational database.

The main theme of this study was categorization by comparing the document to an already defined set of topics. This work also extends to understanding and manipulating the natural language text by the use of natural language processing performed with the help of computer. It is also specified that cross language information retrieval (CLIR), natural language text processing and summarization, speech recognition, user interface, machine translation, artificial intelligence and expert system etc. are the applications of NLP.

Extraction, stop word elimination and stemming are the part of pre-processing mechanism followed in this study. Tokenizing is the process of converting the file content into individual element by the use of extraction method. Text looks heavier and unimportant for analysis by stop word elimination to decrease the dimensionality.

Understanding of the root or stem of the words that are phonologically related (i.e. decreasing the number of words), so as to accurately match stem and remove the common suffixes by using the process of stemming. Based on the domain various stemming algorithm have been developed for providing different services at different level over the years.

As per the study done by Vairaprakash guruswamy et. al. [6], Pre-processing in text mining, natural language processing and information retrieval and its importance have been specified. They have also worked on evaluating the problem in Pre-processing methods for text archives.

Text Pre-processing in NLP system is been examined in this study. Tokenization, stemming and stop word removal are Pre-processing methods proposed in this paper. Sentences are probed by the process of tokenization and are put as list of tokens that can be supplied as input for the algorithms. Brackets hyphen and punctuation marks have been removed in this work.

Stemming process iterates around understanding common representation of words. Under stemming and over stemming are identified has flaws in this process.

Stemming method mainly discussed in this study is:

- ✓ Table look up approach
- ✓ Successor Variety
- ✓ N-gram stemmers
- ✓ Affix removal

This work also turns around eliminating the noise from the text data, trim the size of the data and perform stemming by the techniques of pre-processing. Table 1 shows the different stemming algorithms.

Table 1. stemming algorithms

Stemming Algorithm		
Truncating	**Statistical**	**Mixed**
1. Lovins	1. N gram	1. Infectional & Derivational
2. Portors	2. HMM	a. Krovetz
3. Paice/ Huisk	3. Yass	b. Xerox
4. Dawson		2. Corpus Based

C. Ramasubramanian et. al. [7] from Anna University studied pre-processing to text mining in different ways i.e. by improving the stemming techniques. Disadvantages of MF porter's algorithm i.e. the stemming algorithm has been addressed in this paper. Process to overcome the drawback of the existing approach is discussed in this paper.

To overcome the increase in accuracy and wrong matches spell check is incorporated. Processing time for understanding misspelled words can be saved. In this study a smart word list is proposed that eliminates stop words effectively without actually disturbing the important words.

The study performed by Vikram singh et. al. [8] from National institute of technology, Haryana stated few Pre-processing techniques for information retrieval systems. This paper largely explains stemming algorithm and tokenization that is adopted during data pre-processing.

This paper looks tokenization as very crucial step in text data pre-processing. Data is formally looked at as bag of words. The process of splitting this bag of words into words is said to be token. Information on frequency of token that can be used for information retrieval is formulated by the process of tokenization. Application of stop word removal and stemming algorithm results in output that contains token; this is said to the important for the pre-processing algorithm.

With few sample data sets taken in this paper to shows efficiency increase i.e. upto 68%, when compared token generated after pre-processing and token generated without pre-processing.

Work done by critian moral et. al. [9] from Spain, have experimented an assessment on information retrieval application by using stemming algorithm. The main purpose of stemmer is to clustering words according to topic; improve the effectiveness of IR algorithms. There is reduction of number of words that is needed to be processed due to the words sharing the same stem. Algorithm that is mostly used in information retrieval application is porter's stemmers, but Lovins stemmer is the most widely known algorithm.

Krovetz's is a simple stemmer algorithm that covers past tense, plurals and -ing verbs. Dictionary is been provided to verify the stemmed words.

IV. MOOD ANALYSIS

Gamallo et. al. [10] have worked on Naive-Bayes Strategy for Sentiment Analysis on English Tweets, in which they state that with a binary classifier model they achieved much more accuracy, assigning in only positive or negative categories. Identification of polarity lemma in the tweets with the basic strategy of search the polarity do understand that the tweet is positive or negative and if the polarity doesn't exist then the tweet is said to be neutral. This system developed in this paper address for four language that includes English, Spanish, Portuguese and Galician.

Kumar Singh et. al. [11] worked on the analysis of sentiments and mood using POS over a large number of blog posts on two topics, 'Regionalism' and 'Women's Reservation in India' which uses Vector Space representation and applies semantic orientation approach SO-PMI-IR algorithm for sentimental analysis. This method works more accurately with blog posts. Discussion the different applications of this method is been done in this paper.

To understand the correctness the accuracy of the results were compared by both manual labelling and by comparing outcome of multiple techniques implemented. Instead of machine learning approach that requires a huge amount of training data sets and time to train the classifier, POS tagged based semantic oriented approach is adopted that works on unsupervised approach.

James A. Russell et. al. [12] worked on a model of emotions, which is put forward in a graphical representation in 2D, one axis describes the Variance or how much the word/emotion is pleasant and the other describes how active the emotion is. The graph is a result of their factor-analysis. The study is more focused on level of variance. This work explores affects by self reported data by using huge sample of subject that result in variance unique of that particular methodology. This is like task performed in the previous studies that resulted in unique variance.

Li and Wu [13], worked using text mining and sentiment analysis for online forums hotspot detection and forecast, where they use text mining and sentimental analysis approaches on online forum hotspot detection and forecasting. Further they use SVM (support Vector Machine) and K-mean algorithm to generate unsupervised text mining approach. They extract data from a forum to achieve their goal. The algorithm mentioned in this paper is the combination of k-means clustering and SVM classification to build an integrated approach for online sports forum. By seeking the information from hotspot predicting approaches, companies can be benefited in many ways. Results generated from this approach can be employed with market basket analysis for better decision making.

Fang and Zhan [14], worked on Sentiment analysis using product review data. They discuss Opinion mining and their method to solve problem of sentiment polarity categorization and use data/reviews collected in Amazon.com. They make use of POS tagging and Tokenization. They perform this analysis on sentence level and review level. Sentence level categorization is resulted based on manually labeled sentence and machine labeled sentence. Review level categorization is based on vectors generated i.e. if reviews with 4-star, rating will be labeled with positive. Vectors are labeled negative if reviews are 1 star or 2 star. 3 star reviews is said to be neutral.

Kaur and Singla [15], worked using sentimental analysis of Flipkart reviews using Naive Bayes and Decision Tree algorithm, where they have discussed about the inefficiency in ratings(generally a star rating of 5 stars) and the development such as classifying review texts into subjective/objective and positive/negative attitude of buyers for that product. In this paper they describe text spelling correction features in the review text. Further they classify comments using Hybrid algorithm combining Decision Trees and Native Bayes Algorithm.

Chauhan and Yadav [16], worked using sentimental Analysis of Product Based Reviews Using Machine Learning Approaches, where they make use of machine learning approach (including NB and SVM) due to the highly unstructured data/reviews. They extract data (unstructured), product reviews, pre-process it and calculate polarity (namely Positive, Negative and neutral), also a graph is plotted for the result. Finally discuss about precision and accuracy.

V. COMPARISON

In comparison the common task performed by fiva tech and Depta are data alignment and frequent pattern matching [4][3]. The initial step in Depta is to find repetitive patterns and then alignment is performed, that may most probably result in loss of data. On the other side fiva tech performs data alignment, latter finds the frequently repetitive patterns that results in probably less loss of data. Fiva tech mainly meant to handle single and multiple data record whereas Depta handles only single data record [4].

Similarity in Fiva tech and EXALG a common template and data token is used[4][1]. Many extraction techniques has been mentioned that works on data record except for Roadrunner and NET [1][2].

In paper [7] disadvantage of MF porters algorithm has been addressed in the context of stemming algorithm and to overcome the drawback of the existing approach has been discussed, but in [9] experiment on assessment of information retrieval application by using stemming algorithm is performed.

Paper [6] involves the study of evaluating problems in pre-processing methods for text archives, but in [5] the text document is compared with already defined set of topics and then categorized. [5] Also extends to understand and manipulate the natural language text by the use of natural language processing, but [6] only works on examining the pre-processing of NLP system.

In [5] tokenizing the file content into individual element by the use of extraction method is worked on as sentences are probed by the process of tokenization and is put as list of tokens which can be supplied as input for the algorithms, but in [8] data is looked as bag of words that is tokenized and is mentioned as a crucial step in text data pre-processing. Paper [6] discussed mainly about stemming method like table lookup approach, successor variety, N-gram stemmer and affix removal, but as discussed in [9] the purpose of stemmer is to cluster words according to topic and improve the effectiveness of IR algorithm.

Paper [10] shows the study done on Naive-Bayes strategy for sentiment analysis on English tweets that give more accuracy when worked with binary classifier model and assigns positive or negative categories. In comparison paper [11] uses vector space representation and applies semantic orientation approach over the large no of blog post on certain topic. SVM (support vector machine) and k-mean algorithm is used to generate unsupervised text mining approach [13]. In the study of paper [15] they used Naive bayes and decision tree algorithm for sentimental analysis of flipkart review and tried to understand the inefficiency in rating system. Paper [16] used machine learning approach for sentimental analysis of product based reviews and due to highly unstructured data.

VI. CONCLUSION

This research paper results in giving out options to solve the fundamental painful problem of customers by making them help to decide which product reviews is genuine and how would those insight help them come with a very good product to buy and for businessmen / entrepreneurs by giving them insights on their products and services they have offer to their customers. Better understanding of different extraction techniques in different stages given in various papers. Usage of pre-processing techniques like stop word removal, stemming, tokenization are clearly depicted in various forms in different papers. This paper also takes through different stemming methods. Analyzed different technique for sentimental analysis on product reviews, blog post, online forum for hotspot detection and fore casting.

REFERENCES

[1] Devika K, SubuSurendran. “An Overview of Web Data Extraction Techniques”, volume 2, issue 4.

[2].Kristina Lerman, University of Southern California “Automatic wrapper generation and data extraction” August 25, 2010.

[3]. Mohammed Kayed and Chia-Hui Chang. “FivaTech: Page level web data extraction from template pages”, IEEE transactions on knowledge and data engineering, volume 22, no.2,2010.

[4]. Y. Zhai and B. Liu, “Web data extraction based on partial tree alignment”, Proc Int’l Conf. World Wide Web (WWW-14), 2005, pp. 76-85.

[5] Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya: “Pre-processing Techniques for Text Mining - An Overview” International Journal of Computer Science & Communication Networks, Vol 5(1), 7-16

[6] Vairaprakash Gurusamy, Subbu Kannan: “Preprocessing Techniques for Text Mining” October 2014.

[7] C.Ramasubramanian, R.Ramya: “Effective PreProcessing Activities in Text Mining using Improved Porter’s Stemming Algorithm” International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013

[8] Vikram Singh and Balwinder Saini “An Effective PreProcessing Algorithm For Information Retrieval Systems” International Journal of Database Management Systems (IJDMS) Vol.6, No.6, December 2014

[9] Moral, C., de Antonio, A., Imbert, R. & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval/ Information Research, 19(1) paper 605. [Available at http://InformationR.net/ir/191/paper605.html]

[10] Pablo Gamallo ,Marcos Garcia : “Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets”,- Proceedings of the 8th international Workshop on Semantic Evaluation(SemEval 2014),pp 171-175,Dublin,Ireland, August 23-24 2014.

[11] Vivek Kumar Singh, Mousmi Mukherjee and Ghanshyam Kumar Mehta : “Sentimental and Mood Analysis of Weblogs using POS Tagging based Approach” – Chapter “Contemporary Computing”, Volume 168 of series “Communications in Computer and Information Science”, pp 313 – 324.

[12] James A. Russell : "A Circumflex model of Affects" –Journal of Personality and Social Psychology, Volume 39 ,issue no. 6, pp 1161-1178, 1980.

[13] Nan Li, Desheng Dash Wu: "Using text mining and sentiment analysis for online forums hotspot detection and forecast",-Decision Support Systems 48, 17 September 2009,pp 354-368, Journal homepage: www.elsevier.com/locate/dss

[14] Xing Fang,Justin Zhan: **"**Sentiment analysis using product review data", Journal of Big Data, Springer- April 2015, 2:5 DOI: 10.1186/s40537-015-0015-2.

[15] Gurneet Kaur, Abhinash Singla: "Sentimental Analysis of Flipkart reviews using Naive Bayes and Decision Tree algorithm" - IJARCET , Volume 5 Issue 1, pp 148-153,January 2016.

[16] Manvee Chauhan, Divakar Yadav: "Sentimental Analysis of Product Based Reviews Using Machine Learning Approaches"-JNCET, Volume 5, special issue 2, pp 19-25, December 2015.