Journal of Theoretical and Applied Information Technology
20th October 2014. Vol. 68 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A METHOD FOR EXTRACTING INFORMATION FROM THE WEB USING DEEP LEARNING ALGORITHM

[1]J.SHARMILA, [2]A.SUBRAMANI

[1]Assistant Professor, Department of Computer Applications,
Bharathidasan University Constituent College (W),
Orathanadu, Thanjavur Dt., Tamil Nadu, India
[2]Professor & Head, Department of Computer Applications,
K.S.R. College of Engineering, Thiruchengode, Tamilnadu, India.
E-mail: sharmila0179@gmail.com , sharmi_try@yahoo.co.in

ABSTRACT

Web mining related research are getting more important now a days because of the reason that large amount of data are managed through internet. The web usage is increasing in an uncontrolled manner. A specific system is needed for controlling such large amount of data in the web space. The web mining is classified into three major divisions that are web content mining, web usage mining and web structure mining. In this paper, we propose a web content mining approach based on a deep learning algorithm. The deep learning algorithm provides the advantage over Bayesian networks because Bayesian network is not following in any learning architecture like proposed technique. In the proposed approach, three features are considered for extracting the web content. The features used are concept feature, deals with the semantic relations in the web, format feature, deals with format of the content and title feature, deals with the web tittle. The above listed feature produces some model parameters, which is given as the input to the deep learning algorithm. The experimental analysis showed that, the proposed approach is efficient in web content extraction. The average precision, recall and f-measure values are updated as 83.875%, 78.3% and 80.83% respectively.

1. INTRODUCTION

The quantity of Web information has been rising rapidly, mainly with the development of Web 2.0 environments, where the users are encouraged to provide rich content. A large amount of Web information is presented in the form of a Web document, which occurs in both detail and list. Extraction of Web information is a significant process for information integration, but many web pages may provide the same or analogous information using entirely diverse formats or syntaxes, which makes the addition of information an interesting task. In reference to the heterogeneity and lack of structure of Web information, programmed discovery of applicable information becomes a tough task [1]. The Deep Web is the content on the web not accessible by a search on general search engines, which is also called as hidden Web or invisible Web. Deep Web contents are accessed by queries submitted to Web databases and the retrieved information i.e., query results is enclosed in Web pages in the form of data records. The distinctive Web pages are made dynamically and are tough to index by conventional crawler based search engines, namely Google and Yahoo. In this paper, we describe this kind of special Web pages as deep Web pages [12]. In general, Web information extraction tools are divided into three categories: (i) Web directories, (ii) Meta search engines, and (iii) Search engines.

Normally, a Web page covers several contents or parts, like main content areas, navigation areas, advertisements, etc. A block is a semantic part of a web page that has its own text content, style and functionality. Generally, a web page comprises two blocks: main content blocks and noise blocks. Only the main content blocks describe the informative portion that most users are interested in. Even though other blocks are supportive in improving functionality and superintendent browsing, they harmfully affect such web mining jobs as web page clustering and classification by reducing the precision of mined results and speed of processing.

Hence, these blocks are called noise blocks in this context. For instance, a CNN web page contains a sports news report in the middle of the page, which is the main content of this page. Also, there are advertisements, navigation bars, and others, situated around the main content, which are called as noise blocks [2]. In addition to main content, web pages usually have image-maps, logos, advertisements, search boxes, headers and footers, navigational links, related links and copyright information in conjunction with the main content. Though these items are required by web site owners, they will obstruct the web data mining and decrease the performance of the search engines [14], [15].

The purpose of Web mining is to develop methods and systems for discovering models of objects and processes on the World Wide Web and for web-based systems that show adaptive performance. Web Mining integrates three parent areas: Data Mining (we use this term here also for the closely related areas of Machine Learning and Knowledge Discovery), Internet technology and World Wide Web, and for the more recent Semantic Web. The World Wide Web has made an enormous amount of information electronically accessible. The use of email, news and markup languages like HTML allows users to publish and read documents at a world-wide scale and to communicate via chat connections, including information in the form of images and voice records [20]. The current web is a web designed for finding documents. The semantic web is a web designed for finding data. Data is best found through structured relationships when accuracy and context are desired. Data sharing in science is the type of exercise where accuracy and context is required. Scientific patterns of information exchange require standards, and the semantic web can provide useful tools for structuring data according to standard structured relationships. The structure that defines contextual relationships on the semantic web is known as ontology, which is a hierarchical organization of a knowledge domain that contains entities and their relations [21].

In this paper, a method is proposed on web content mining approach based on a deep learning algorithm. The deep learning algorithm provides the advantage over Bayesian networks because Bayesian network is not following in any learning architecture like proposed technique. In the proposed approach, three features are considered for extracting the web content. The features used are *concept feature*, deals with the semantic relations in the web, *format feature,* deals with format of the content and *title feature,* deals with the web tittle. The above listed feature produces some model parameters, which is given as the input to the deep learning algorithm. The process continues according to the deep learning algorithm and finally extracts content according to the input provided

The main contributions of the proposed approach are,

- A deep learning based approach is used for web content extraction
- Three characteristic features are used for identifying important blocks
- The features are listed as concept feature, tittle feature and format feature

The rest of the paper is organised as the second section include the review of related works, the problem description is given in the third section. The 4th section includes the proposed methodology and 5th section consists of the experimental evaluation and discussion. The conclusion of the work is given in the 6th section

2. REVIEW OF RELATED WORKS

Our proposed method concentrates on web content extraction based on deep learning algorithm based web data extraction. Many Researchers have developed several approaches for web content extraction based on different methods. Among them, a handful of significant researches that performs web content extraction are presented in this section.

Ashraf F *et al* [13] have proposed a system, where clustering techniques have been used for automatic IE from HTML documents having semi-structured data. By means of domain-specific information provided by the user, the proposed system has parsed and tokenized the data from an HTML document, divided it into clusters having analogous elements, and estimated an extraction rule based on the pattern of occurrence of data tokens. Then, the extraction rule has been utilized to refine clusters, and finally, the output has been demonstrated. Moreover, a multi-objective genetic algorithm-based clustering method has been used for finding the number of clusters and the most natural clustering. It is complex and even impossible to employ a manual approach to mine the data records from web pages in deep web. Thus, *Chen Hong-ping et al* [9] have proposed a LBDRF algorithm to solve the problem of automatic data records extraction from Web pages in deep Web. Experimental result has shown that the proposed technique has performed well.

Zhang Pei-ying and Li Cun-he [10] have proposed a text summarization approach based on sentences clustering and extraction. The proposed approach includes three steps: (i) the sentences in the document have been clustered based on the semantic distance, (ii) the accumulative sentence similarity on each cluster has been calculated based on the multi-features combination technique, and (iii) the topic sentences has been selected via some extraction rules. The goal of their research is to exhibit that the summarization result was not only depends on the sentence features, but also depends on the sentence similarity measure. Qingshui Li and Kai Wu [6] have developed a Web Page Information extraction algorithm based on vision character. A vision character rule of web page has been employed, regarding the detailed problem of coarse-grained web page segmentation and the restructure problem of the smallest web page segmentation. Then, the vision character of page block has been analyzed and finally determined the topic data region accurately. They have proved that after using the information extraction technology of web page, the information block of web page content has been reduced and thus the cost of index generating has been decreased, as well as increased the hit rate of search engine.

Yan Liu ***et al*** [1] have proposed a document summarization framework via deep learning model, which has demonstrated distinguished extraction ability in document summarization. The framework consists of concepts extraction, summary generation and reconstruction validation. A query-oriented extraction technique has been concentrated information distributed in multiple documents to hidden units layer by layer. Then, the whole deep architecture was fine-turned by minimizing the information loss in reconstruction validation part. According to the concepts extracted from deep architecture, dynamic programming was used to seek most informative set of sentences as the summary. Experiments on three benchmark dataset demonstrate the effectiveness of the framework and algorithms.

Theoretical results suggest that in order to learn the kind of complicated functions that can represent high-level abstractions (e.g., in vision, language, and other AI-level tasks), one may need deep architectures. Deep architectures are composed of multiple levels of non-linear operations, such as in neural nets with many hidden layers or in complicated propositional formulae re-using many sub-formulae. Searching the parameter space of deep architectures is a difficult task, but learning algorithms such as those for Deep Belief Networks have recently been proposed to tackle this problem with notable success, beating the state- of-the-art in certain areas. This monograph discusses the motivations and principles regarding learning algorithms for deep architectures, in particular those exploiting as building blocks unsupervised learning of single-layer models such as Restricted Boltzmann Machines, used to construct deeper models such as Deep Belief Networks.

3. PROBLEM DESCRIPTION

In the modern era, the websites are considered as the one touch sources of all kind of information needed by an individual. The data stored in the web spaces are numerous and one can refer any kind of information with the help of websites. Recently, information is extracted from the web using programmed methods because of the need of information. As the extraction process become viral, the websites are become sources of redundant information. The duplication becomes a major issue. Thus, a method is needed to extract information from the websites by identifying the relevant information. The main problem faced by extractors is that, a single website contains same content a number of times and possesses other irrelevant information also. In according to that, Tak-Lam Wong and Wai Lam [1] have proposed a web content mining approach in the research with the help of Bayesian networks. In their approach, they have done learning on extracting the web information and attribute discovery based on the Bayesian approach. Inspired from the research, a method is proposed for web content mining approach based on a deep learning algorithm. The deep learning algorithm provides the advantage over Bayesian networks because Bayesian network is not following in any learning architecture like proposed technique. The proposed deep learning approach helps to identify the relevant content from the websites through the layer by layer approach of the deep leaning architecture.

4. PROPOSED WEB CONTENT EXTRACTION APPROACH THROUGH

Web mining related research are getting more important now a days because of the reason that large amount of data are managed through internet. The web usage is increasing in an uncontrolled manner. A specific system is needed for controlling such large amount of data in the web space. The web mining is classified into three major divisions that are web content mining, web usage mining and web structure mining. In this paper, a discussion about web content mining and a method used for web content extraction are given. The proposed approach deals with a web content extraction method through deep learning architecture. The usual web content extraction methods concentrate only on the extracting content without checking, whether it is relevant or not. The proposed approach uses a tri phase processing to identify and extract the relevant content from the websites. The three phases of the proposed approach includes,

- Pre-processing phase
- Training of deep learning algorithm
- Testing

The above listed three processes are considered as the major processing phases by the proposed approach. The pre-processing phase concentrates on the preparing the websites for deep learning algorithm to process. The training phase makes the deep learning algorithm for identifying the relevant information. The testing phase is initiated to extract relevant information from some set of websites or web documents. The procedural approach of these three phases give more efficiency to the propose web content extraction method.

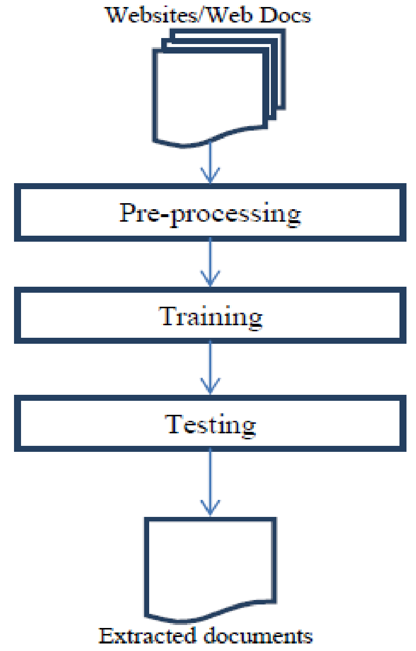


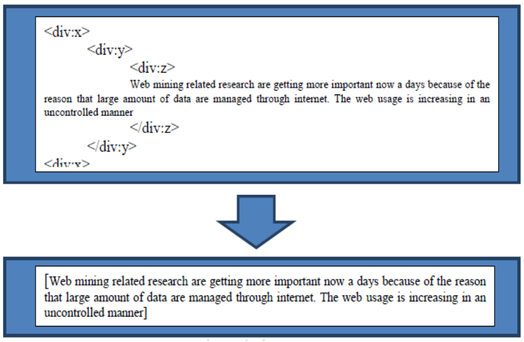
Fig.1. Overview Of Web Content Extraction

4.1. Pre-Processing Phase

The initial phase of the proposed approach is the pre-processing of websites or web documents through different document processing methods. The pre-processing phase is intended to produce input to the deep learning algorithm. A series of processing are considered in the pre-processing phase for preparing the web documents or websites for content extraction. The preprocessing phase includes the following processes,

Tag Separator: In this process, the supplied URL is processed and the particular website is retrieved from the internet. The website is then processed from its source level for identifying the tags. The tag removal process is then subjected on the websites. All the tags except <div> tags are removed from the website's source page. Since <div> tags are considered as the tags, which hold contents of the websites, the <div> tags are not removed. The contents left after the tag separator process are then subjected to block building process.

Block separator: This process is applied to extract content blocks inside particular <div> tags. Every web page contains nested <div> tags, which contain important data of the web sites. So, the contents rom the website are extracted by processing from the inner most <div> tags. The contents are extracted by opening the <div> tags level by level, i.e. from inner most <div> tags to the top most <div> tags. The top most div tag is transformed to block brackets, which act as a separator to the content under consideration.



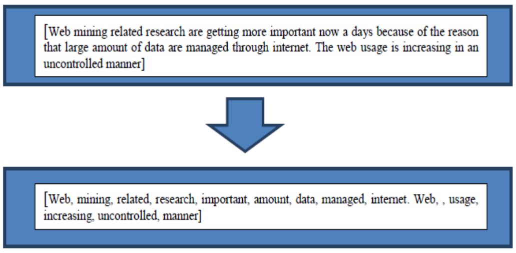
```
<div:x>
        <div:y>
                <div:z>
                Web mining related research are getting more important now a days because of the reason that large amount of data are managed through internet. The web usage is increasing in an uncontrolled manner
                </div:z>
        </div:y>
<div:x>
```

[Web mining related research are getting more important now a days because of the reason that large amount of data are managed through internet. The web usage is increasing in an uncontrolled manner]

Fig.2. Block Separator

The fig 2 represents an example of block separation process. As discussed above, the content of the websites are separated by block brackets. The importance of each block is calculated in the further sections.

Stop word removal: The block separator phase generates the contents in the web page with a separator, i.e. each content are listed as blocks. The contents contains important keywords and unwanted connective words such as is, a, as, was, etc. The proposed approach uses a stop word removal process to separate the contents from the connective words. The stop word removal phase helps the proposed approach to identify the keywords separately.



[Web mining related research are getting more important now a days because of the reason that large amount of data are managed through internet. The web usage is increasing in an uncontrolled manner]

[Web, mining, related, research, important, amount, data, managed, internet. Web, , usage, increasing, uncontrolled, manner]

Fig.3. Stop Word Removal

The fig 3 represents the processing of contents in the block using the stop word removal method. After the stop word removal, the keywords are listed in the block separated by commas.

Feature calculation based on contents

The feature calculation is the main process in the proposed web content extraction approach. The features are calculated for the content in the blocks, which are formatted by the above processing methods. The feature calculations are done to ease the processing in the deep learning algorithm. After the feature calculations, each block will possess a particular value. The value enabled blocks are given as input to the training phase of the deep learning algorithm. The blocks are identified by the deep learning training phase as important or non-important based on these feature values. Consider the following feature value calculations,

Concept feature: A major component in the feature value set, the concept feature is defined over the frequent keywords in the blocks. The concept feature is calculated based on the count of frequent keywords in sentences in the content blocks. Initially, all the blocks are scanned and unique keywords are extracted and stored in a set.

$$Uq = [k_1, k_2, \ldots, k_n]$$

Here, *Uq* represents the set of unique keywords possessed by the whole set of blocks. K_i represents the individual keywords. These key words are then scanned over the blocks of contents to extract the frequency of the keywords. A list is then created with keywords and their frequency values. A threshold value is applied to the sort out the keywords, thus the list contains only the most frequent keywords.

$$MFK_{list} = [k_1 : f_{count}, k_2 : f_{count}, \ldots, k_n : f_{count}]$$

The value MFK represents the most frequent keyword and f is represented as the count of frequency values of a particular keywords. The concept feature of a particular block is calculated by a sentence based scanning method. The most frequent keywords are traversed through each block, if a sentence contain two or more keywords, and then a value is assigned to the concept feature of that block. As the process forwards through other sentences in the blocks, the value get incremented accordingly.

$$[Content\_block \xleftarrow{traversal} MFK_{list}] \Rightarrow Concept\_feature$$

Thus, the concept feature is calculated as the sum of feature value assigned to each sentences in the block of contents. The sentence score is calculated by the presence of MFK in it.

$$Concept\_feature = \sum S_i$$

Here, S_i represents the sentence score of a particular block and the summation of the sentence score gives concept feature.

Format feature

The format feature value of a block is associated with format in which the contents are arranged in the block. The main features considered for the format feature calculation are the size of the text, the bold characters, the italic characters, line separation, captions, tittles, etc. Assessing all these values sentence by sentence, a sentence score is created.

$$Sentence\_score = \sum size, bold, italic, line, tittle, etc$$

If the feature is present in the sentence, a value of is assigned to the sentence score and if that particular feature is not present, a value of 0 is assigned to the sentence score. Similar to the concept feature, format feature is also calculated as the sum of sentence score in a particular block.

$$format\_feature = \sum S_j$$

Here, S_i represents the sentence score of a particular block and the summation of the sentence score gives concept feature.

Title feature

A sentence is considered important if it's similar to the title of text document. Here similarity is considered on the basis of occurrence of common words in title and sentence. A sentence has good feature score if it has maximum number of words common to the title. The ratio of the number of words in the sentence that occur in title to the total number of words in the title helps to calculate the score of a sentence for this feature. It is calculated by

$$Tf = \frac{t \cap w}{w}$$

Here, the term T_f indicates the tittle feature, term t indicates the tittle of the content and the term w indicates the terms present in the content. The tittle feature of the entire block is calculated by the sum of tittle feature extracted for the sentences present in the contents.

$$Tittle\_feature = \sum Tf$$

Thus, the proposed approach defines three parameters for each block present in the websites. These three parameters are the deciding parameters in the proposed deep learning algorithm. The values of these three features defined over the block define the important blocks.

4.2. Training the Deep Leaning Algorithm

In this section, we discuss the various steps involved in the training phase of the deep learning algorithm. The deep learning algorithm adopted for the text summarization phase by the proposed approach is derived from the Boltzmann algorithm. The procedure is of Boltzmann algorithm works based on the RBM architecture. The architecture defines visible layers and hidden layers, passing values through these layers a deep learning method is evolved. According to the definition of the proposed approach, the feature matrix from the prior step is given as input to the deep learning algorithm.

	F_1	F_2	F_3
B1			
B1			
….			
Bn			

Fig. 4. Input Feature Matrix

Here, the columns represent the sentences of extracted from the documents and each sentence is associated with a set of three. Each, B_i= [F_1, F_2, F_3], where i <=n where n is the number of sentences in the document. Restricted Boltzmann machine contains two hidden layers and for them two set of bias value is selected namely H_0, H_1. Each values of H_i contain corresponding attribute values, which are similar to the feature values.

$$H_0 = [h_1, h_2, ..., h_n]$$
$$H_1 = [h_1, h_2, ..., h_n]$$

The values of the attributes in the defined hidden layers are assigned randomly based on the relevance of the problem space. A repeated learning process is executed between the hidden layers and active layers of the RBM architecture. .The whole operation with RBM starts with giving the feature matrix as input. Considering the proposed approach, two hidden layers are utilized for solving the problem under discussion. RBM works in two step .The input to first step is our set of blocks defined as B=[B_1, B_2,…, B_n] which is having the five features of sentence as element of each sentence set. Similar to every learning algorithms, the adopted deep learning tech also have cyclic processing, so during the first cycle a newly refined sentence set is constructed by associating the attributes from the hidden layers.

$$B_i^{'} = \sum_{i=1}^{n} B_i + h_i \Rightarrow [B_1^{'}, B_2^{'}, ..., B_n^{'}]$$

In proceeding forward the same procedure will be applied to the obtained refined set S_i', in order to obtain a more refined sentence matrix set with the help ofH_1 and which is given by,

$$B'' = [B_1'', B_2'', ..., B_n'']$$

After obtaining the refined sentence matrix from the RBM it is further tested on a particular randomly generated threshold value for each feature we have calculated .For example we select threshold thr_c as a threshold value for the extracted concept-feature. If for any sentence f_i<*thr* then it will be filtered and will become member of new set of feature vector. The following section plots the block diagram of the proposed learning algorithm.

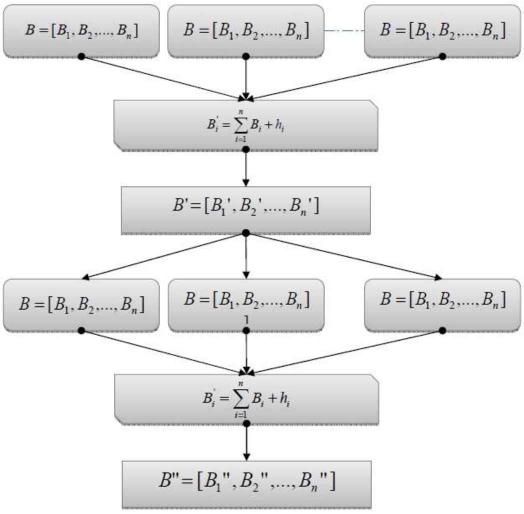

$B=[B_1,B_2,....,B_n]$
$B_i'=\sum_{i=1}^{n} B_i+h_i$
$B'=[B_1',B_2',....,B_n']$
$B''=[B_1'',B_2'',....,B_n'']$

Fig.5.Deep Leaning Training Process

The fig 5 depicts the step by step processing of the blocks through the layers of deep learning algorithm. The initial layer is filed with the obtained blocks and then it is processed with the attributes of hidden layer H_0. The new hidden layer attributes is set as a benchmarking set, which will be used in the testing phase of the learning algorithm. Now, a new sentence set $S^{'}$ is created and is again processed with the hidden layer H_1 to provide the layer H_1 with an updated attributes set. Similarly, the hidden layer is processed with features from the entire sentence in the feature matrix. The next phase of the learning algorithm is the testing phase, which involves the testing of blocks, to find the relevant contents for the web content extraction.

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1. Experimental Setup

The experimental results of the proposed method web data extraction for web document clustering are presented in this section. The proposed approach has been implemented in java (jdk 1.7) and the experimentation is performed on a 3.0 GHz core i5 PC machine with 4 GB main memory. For experimentation, we have taken many web pages which contained all the noises such as Navigation bars, Panels and Frames, Page Headers and Footers, Copyright and Privacy Notices, Advertisements and Other Uninteresting Data. The webpages are then subjected to process through the proposed deep learning network to web content extraction.

5.2. Performance Measures

Precision is the percentage of the relevant data records identified from the web page.

$$\text{Precision} = \frac{DR_c}{DR_e}$$

Recall defines the correctness of the data records identified.

$$\text{Recall} = \frac{DR_c}{DR_r}$$

Where,

DR_c is the total number of correctly extracted data records

DR_e is the total number of data records on the page

DR_r is the total number of data records extracted

F-measure is the ratio of product of precision and recall to the sum of precision and recall. The f-measure can be calculated as,

$$F-measure = \frac{2\times precision \times recall}{precision + recall}$$

5.3. Performance Analysis

This section consists of the evaluation of performance of the proposed web content extraction based on the deep learning methodology. As discussed in the prior section that, the documents are collected from the internet and the web pages collected contains both relevant and irrelevant contents as per the need of the user. The role of the proposed deep learning methodology is to extract the relevant contents from the web pages by identifying them accurately. The evaluation measures used by the proposed approach are precision, recall and f-measure. The detailed explanations about the evaluation parameters are given in the above section. The following section depicts the performance of the proposed deep learning method in responses to the given dataset. The proposed approach selected a set of web documents from the internet and manually extracted its blocks based on div tag. Then they are analysed and the important blocks are identified. After the manual calculations, the proposed deep learning method is subjected to process the extracted webpages. The content extracted by the proposed deep learning method is compare with content identified manually for evaluating the accuracy of the proposed deep learning based web content extraction.

5.3.1.Performance Analysis Based on Precision

The proposed approach has selected a set of web documents for evaluation process. The different blocks are evaluated here based on the precision parameter. As discussed above, the precision defines the relevance of the extracted blocks by the proposed deep learning based web content extraction.

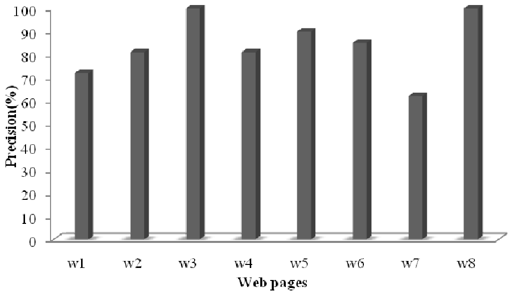


Fig.6. Precision Values

The fig 6 depicts the precision values obtained by processing the set of web documents extracted from the internet through the deep learning algorithm. Here, w1, w2, etc. are webpages used for the experimentation process. The blocks from eight different web pages are selected for the processing. The precision value is calculated by comparing the important blocks between the manually identified important blocks and the program generated important blocks. The analysis from the fig 5 shows that, the responses of the proposed approach is different for different web pages. The approach gives different value based on the depth of the blocks in the web pages. The average precision value obtained for the proposed deep learning methodology is 83.9%, which states that the proposed approach has good ability in identifying the important blocks from the web pages.

5.3.2. Performance Analysis Based on Recall

Similar web data is considered for the recall analysis also by the proposed approach. As discussed above, the recall defines the correctness of the extracted blocks by the proposed deep learning based web content extraction.

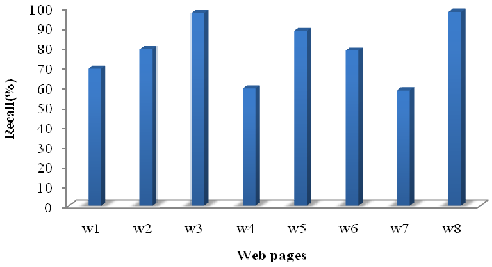


Fig.7.Recall Values

The fig 7 depicts the recall values obtained by processing the set of web documents extracted from the internet through the deep learning algorithm. In recall calculation also, blocks from eight different web pages are selected for the processing. The recall is calculated by comparing the extracted blocks between the manually identified blocks and the program generated important blocks. The analysis from the graph shows that, the responses of the proposed approach are different for different web pages. The approach gives different value based on the depth of the blocks in the web pages. The average recall value obtained for the proposed deep learning methodology is 78.9%, which states the correctness of the proposed approach in identifying the important blocks.

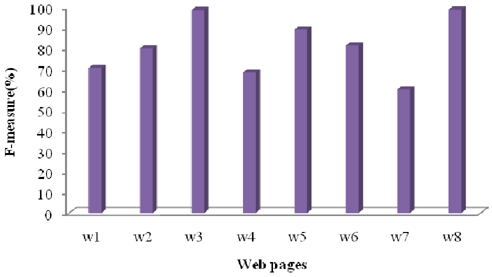


Fig.8. F-Measure Values

The fig 8 illustrates the f-measure values of the proposed approach based on the different web pages. The F-measure is calculated based on the values of precision and recall. The f-measure is considered as the harmonic mean values by comparing the precision and recall values. The analysis from the graph shows that, the response of recall and precision are reflected in the f-measure values. Considering the f-measure values, the average obtained is 80.9%. The different analysis states that, the proposed approach has good ability to extract relevant contents from the web pages.

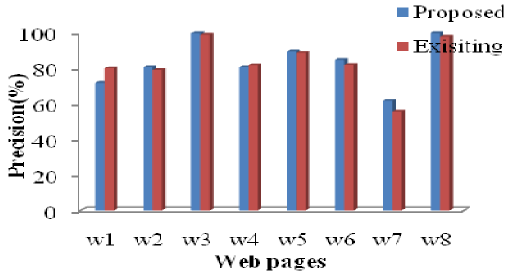
5.4. Comparative Analysis

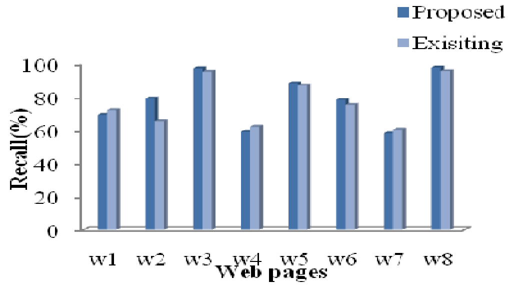
The comparative analysis portions states the comparison of proposed approach with an existing web content extraction method. Here, the proposed deep learning method is compared with a web content extraction based on Bayesian networks. In the comparison process, three parameters of the both algorithms are considered, the precision, the recall and the f-measure. The following table provides the analysis of both the approaches for a set of web pages and their content blocks.

Table.1. Comparative Analysis

	Proposed			Existing		
Web pages	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
w1	72	69	70.46809	80.3	72.1	75.9794
w2	81	79	79.9875	79.2	65.32	71.59347
w3	100	97.2	98.58012	99.2	95.2	97.15885
w4	81	59	68.27143	82.1	61.8	70.51814
w5	90	88.2	89.09091	89.17	86.7	87.91766
w6	85	78.2	81.45833	82.3	75.23	78.60635
w7	62	58.1	59.98668	55.9	60.2	57.97037
w8	100	97.7	98.83662	98.2	95.6	96.88256

The table 1 represents the value obtained for the two algorithms after processing with the same set of data. The precision, recall and f measure values are listed in the above table.





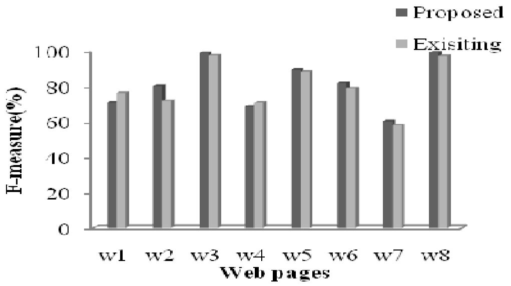


Fig.9. Comparative Analysis

The fig 9 shows the comparative analysis of the proposed deep learning based approach with the existing Bayesian based approach. The different analysis showed that the proposed approach and the existing approach are head to head in extracting the web contents. On considering the average values the proposed approach has upper hand over the existing Bayesian approach.

Table.2. Average Values

	Avg.P(%)	Avg.R(%)	Avg.F(%)
Deep learning method	83.875	78.3	80.83
Bayesian method	83.296	76.5	79.5

Thus it can be stated that the proposed deep learning algorithm based approach is efficient in extracting the web contents by identifying the important blocks.

6. CONCLUSION

Web mining related research are getting more important now a days because of the reason that large amount of data are managed through internet. The web usage is increasing in an uncontrolled manner. A specific system is needed for controlling such large amount of data in the web space. The web mining is classified into three major divisions that are web content mining, web usage mining and web structure mining. The proposed approach uses an algorithm based on deep learning methodology. The deep learning algorithm is efficient in identifying important features in content. The proposed approach uses three parameters for the evaluation of the web pages through deep learning algorithm. The deep learning algorithm is executed in two ways the training phase and the testing phase. In training phase, the deep learning algorithm is trained with known data and the relevance of the trained deep learning network is evaluated in the testing phase. The experimental analysis showed that, the proposed approach is efficient in web content extraction. The average precision, recall and f-measure values are updated as 83.875%, 78.3% and 80.83% respectively. In future, we can incorporate different machine learning technique as hybrid algorithms for efficient web information extraction.

REFERENCE

[1] Tak-Lam Wong and Wai Lam, “Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a Bayesian Approach”, IEEE Transactions On Knowledge And Data Engineering, vol. 22, no. 4, pp: 523-536, 2010.

[2] Yan Liu, Sheng-hua Zhong, Wen-jie Li, "Query-Oriented Unsupervised Multi-document Summarization via Deep Learning", Under review in Journal of Neural Networks (NN).

[3] Jen-Yuan Yeh , Hao-Ren K and Wei-Pang Yang "iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network", Expert Systems with Applications, vol. 35, no. 3, pp. 1451-1462, October 2008.

[4] Yatsko V., Shilov S. and Vishniakov T., “A Semi-automatic Text Summarization System”, In proceedings of the 10th International Conference on Speech and Computer, Patras, pp. 283-288, 2005.

[5] LaddaSuanmali, NaomieSalim and Mohammed Salem Binwahlan, "Automatic Text Summarization Using Feature Based Fuzzy Extraction", vol. 20, no. 2, pp. 105-115, November 2009.

[6] KaustubhPatil and PavelBrazdil, "SUMGRAPH: Text Summarization Using Centrality In The Pathfinder Network", International Journal on Computer Science and Information Systems, vol.2, no.1, pp. 18-32, 2007.

[7] Aaron Harnly, Ani Nenkova, Rebecca Passonneau and Owen Rambow, “Automation of Summary Evaluation by the Pyramid Method”, In Proceedings of the Conference of Recent Advances in Natural Language Processing, pp: 226, 2005.

[8] Rachit Arora and Balaraman Ravindran, “Latent Dirichlet Allocation Based Multi-Document Summarization”, In Proceedings of the second workshop on Analytics for noisy unstructured text data, pp:91-97, 2008.

[9] KhosrowKaikhah, “Automatic Text Summarization with Neural Networks”, Second IEEE International conference on intelligent systems, pp: 40-45, 2004.

[10] H. Edmundson, “New methods in automatic extracting”, Journal of the Association for Computing Machinery, Vol: 16, No. 2, pp: 264-285, 1969.

[11] Inderjeet Mani, “Recent Developments in Text Summarization", In Proceedings of the tenth international conference on Information and knowledge management, ACM Press, pp: 529 - 531, 2001

[12] ShiyanOu, Christopher S.G. Khoo and Dion H. Goh, "Design and development of a concept-based multidocument summarization system for research abstracts", Journal of Information Science, vol. 34 , no. 3, pp. 308-326 , June 2008.

[13] Jade Goldstein, Vibhu Mittal, Jaime Carbonell and Mark Kantrowitzt, "Multi-Document Summarization by Sentence Extraction", NAACL-ANLP 2000 Workshop on Automatic summarization, Seattle, Washington, vol. 4, pp. 40 - 48, 2000.

[14] Breck Baldwin and Thomas S. Morton, “Dynamic coreference-based summarization”, in Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3), Granada, Spain, June 1998.

[15] Mohammed Salem Binwahlan, Naomie Salim and Ladda Suanmali, "Swarm Based Features Selection for Text Summarization", IJCSNS International Journal of Computer Science and Network Security, vol. 9, no.1, January 2009.

[16] Ashraf, F.; Ozyer, T.; Alhajj, R.; "Employing Clustering Techniques for Automatic Information Extraction from HTML Documents," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol.38, no.5, pp.660-673, 2008.

[17] Zhang Pei-ying, Li Cun-he, "Automatic text summarization based on sentences clustering and extraction,"2nd IEEE International Conference on Computer Science and Information Technology, pp.167-170, 2009.

[18] Chen Hong-ping; Fang Wei; Yang Zhou; Zhuo Lin; Cui Zhi-Ming; "Automatic Data Records Extraction from List Page in Deep Web Sources, “Asia-Pacific Conference on Information Processingvol.1,pp.370-373, 2009.

[19] Qingshui Li; Kai Wu; "Study of Web Page Information topic extraction technology based on vision," IEEE International Conference on Computer Science and Information Technology (ICCSIT), vol.9, pp.781-784, 2010.

[20] Bettina Berendt, Andreas Hotho, Dunja Mladenic, “A Roadmap for Web Mining From Web to Semantic Web”, Institute of Technical and Business Information Systems, Otto–von–Guericke–University Magdeburg, Vol. 18 , pp. 1-21, 2008.

[21] Andrew Clearwater, “The new ontologies: the effect of copyright protection on public scientific data sharing using semantic web ontologies”, Vol. 10, pp. 182-205, 2010.