PERFORMANCE ANALYSIS OF VISION-BASED DEEP WEB DATA EXTRACTION FOR WEB DOCUMENT CLUSTERING

[1]M. Lavanya and [2]M. Usha Rani

[1]Assistant Professor [SL], Department of Master of Computer Applications
Sree Vidyanikethan Engineering College, A.Rangampet, Tirupati, Andhra Pradesh, INDIA-517102

[2]Associate Professor, Department of Computer Science
Sri Padmavati Mahila Viswavidyalayam, (SPMVV Woman's' University), Tirupati
Andhra Pradesh, INDIA-517501

Abstract

Web Data Extraction is a critical task by applying various scientific tools and in a broad range of application domains. To extract data from multiple web sites are becoming more obscure, as well to design of web information extraction systems becomes more complex and time-consuming. We also present in this paper so far various risks in web data extraction. Identifying data region from web is a noteworthy crisis for information extraction from the web page. In this paper, performance of vision-based deep web data extraction for web document clustering is presented with experimental result. The proposed approach comprises of two phases: 1) Vision-based web data extraction, where output of phase I is given to second phase and 2) web document clustering. In phase 1, the web page information is segmented into various chunks. From which, surplus noise and duplicate chunks are removed using three parameters, such as hyperlink percentage, noise score and cosine similarity. To identify the relevant chunk, three parameters such as Title word Relevancy, Keyword frequency-based chunk selection, Position features are used and then, a set of keywords are extracted from those main chunks. Finally, the extracted keywords are subjected to web document clustering using Fuzzy c-means clustering (FCM). The experimentation has been performed on two different datasets and the results showed that the proposed VDEC method can achieve stable and good results of about 99.2% and 99.1% precision value in both datasets.

KEYWORDS: FEATURES, RISKS, PROBLEMS, VDEC, FRAMEWORK, POSITION FEATURES, FUZZY C-MEANS CLUSTERING (FCM)

1. Introduction

Now-a-days, information can be retrieved from web as a main source for the required applications. The content of information from the web is in the form of unstructured text, a huge amount of semi-structured objects, called data records, are enclosed on the Web [5]. Vision-based Web Data Extraction system can be done with various web sources using different techniques and extract the data regions stored in the deep web page. Consider, if the source is a HTML Web page, the extracted information could consist of elements in the page as well as the full-text of the page itself. The deep web data region has to be again convert into a Structured format.[Zhao 2007; Irmak and Suel 2006].Vision-based web data extraction has useful data extraction from the deep web pages which are hidden web pages. The consequence of Vision based Web Data Extraction systems depends large (and quickly growing) amount of information is continuously produced, shared and consumed online: Web Data Extraction systems allow to efficiently collect this information with a limited human effort. Huge web information is presented in the form of a Web record, which consist of whole pages as well as catalog pages. Combining all the information which is extracted from the web, but many web pages may provide the same or related information using entirely diverse formats or syntaxes, which makes the integration of information a challenging task.

For instance, a NITL web page contains details of liberal information in the center 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].

Many of the noisy items are required by web site owners; they will obstruct the web data mining and decrease the performance of the search engines [14], [15]. Based on their different levels of abstraction, Web noise can be classified into two categories: Global noise and Local noise (intra-page). **Global noise** is the surplus objects with large granularities, which are in no means smaller than the individual

pages. Global noises are in mirror sites, replica Web pages and obsolete Web pages that are need to be deleted. **Local (intra-page) noise** is the irrelevant items inside a Web page. Normally, local noise is irrational with the primary content of the page. Such noise encompass banner commercials, navigational guides, garnishing images, etc [16], [17], [18]. Hence, having a method that automatically discovers the information in a web page and allots substantial measures for different areas in the web page is of an immense advantage [19], [20]. It is imperative to distinguish relevant information from noisy content because the noisy content may deceive users' concentration within a solitary web page, and users only pay attention to the commercials or copyright when they search a web page. Thus, different information within a web page will have diverse significance weight based on its location, occupied regions, subject, and more [19], [20].

Within web, semantically related content is usually grouped together and the whole page is divided into regions for diverse contents by means of explicit or implicit visual separators namely lines, blank areas, images, font sizes, colors, and more [4]. Web pages are often chaotic with disquieting features around the body of an object that distract the interest of the user from the actual content they are interested in. These "features" may comprise pop-up advertisement, showy banner advertisements, search and filtering panel, superfluous images, or links scattered around the screen. However, these noisy data are present in various patterns in diverse Web sites. Such irrelevant items should be removed for extracting only the significant information [3].

In this paper we are experimenting on an approach [10] to extract data items from the deep web pages automatically. It consists of two stages of execution: (1) Identification and Extraction of the data extraction for deep web page (2) Web clustering using FCM algorithm. Firstly in a web page, the irrelevant data such as advertisements, images, audio, etc are removed using chunk segmentation operation. The result we will obtain is a set of chunks. From which, the surplus noise and the duplicate chunks are removed by computing the three parameters, such as *Hyperlink percentage*, *Noise score* and *cosine similarity*. For each chunk, three parameters such as *Title word Relevancy*, Keyword *frequency based chunk selection* and *Position feature* are computed. Using these parameters, the sub-chunk weightage of each and every sub-chunk is calculated. The weightage of one or more sub-chunks will be greater than other sub-chunks. These sub-chunks consider as the main chunk and the keywords are extracted from those main chunk. To cluster documents, we have to select right the type of the characteristics or attributes (e.g. words, phrases or links) of the documents on which the clustering algorithm will be based and their representation.

The vision-based web document clustering approaches characterize each document according to its content, i.e. the words (or sometimes phrases) contained in it. The basic idea is that if two documents contain many common words then it is likely that the two documents are very similar. The vision-based approaches can be further classified according to the clustering method used into the following categories: *partitional*, *hierarchical*, *graph-based*, *and neural network-based* and *probabilistic*. Fuzzy clustering approaches, on the other hand, are non-exclusive, in the sense that each document can belong to more than one clusters. Fuzzy algorithms usually try to find the best clustering by optimizing a certain criterion function. The fact that a document can belong to more than one clusters is described by a *membership function*. The membership function computes for each document a membership vector, in which the I i^{th} element indicates the degree of membership of the document in the i-th cluster. The most widely used fuzzy clustering algorithm is Fuzzy c-means (Bezdek, 1984), a variation of the partitional k-means algorithm. In fuzzy c-means each cluster is represented by a *cluster prototype* (the center of the cluster) and the membership degree of a document to each cluster depends on the distance between the document and each cluster prototype. The closest the document is to a cluster prototype, the greater is the membership degree of the document in the cluster.

The paper is organized as follows. Section 2 presents the various risks to overcome in web data extraction methods. The features of visual deep web page is described in section 3 and vision –based deep web data extraction for web document clustering which is an approach is presented in section 4. An efficient approach web document clustering based on vision-based deep web is discussed in section 5. The experimental results are reported in Section 6. Section 7 explains conclusion of the paper.

2. Various Risks To Overcome In Web Data Extraction Methods

Web Data Extraction Systems has various perceptions and it influences on various disciplines like Machine Learning, Logic and Natural Language Processing in design and implementation. Many aspects are taken into consideration, where some of the aspects are independent of particular domain to plan to extract web data in the design of the web data extraction method. Other factors, depend on some of the instead, heavily depend on the exacting characteristics of the application domain where some of the technological solutions which appear to be effective in some application contexts are not suitable in other ones. Some of the approaches use static HTML web pages where tags containing a page level-by-level organizing to retrieve information. All the techniques are only to increase the correctness over huge document collections, where the results obtained are not redundant. The risk involved is extracting data from the web is tough due to number of requirements has to be met.

2.1 The major problems raised in the design of a Web Data Extraction system can be listed as follows:

Specialist help is used for Web Data Extraction techniques. A problem comprises of providing a elevated extent of mechanization by sinking human efforts as much as possible. Specialist feedbacks, however, may involve key role in increase the intensity of accuracy accomplished by a Web Data Extraction system. A associated problem is, therefore, to make out a sensible between the need of building highly automated Web Data Extraction procedures and the requirement of achieving highly accurate performance. Web Data Extraction techniques should be able to process large volumes of data in relatively short time. Such a need is particularly urgent in the field of Business and Competitive Intelligence because a firm needs to perform timely analysis of market conditions. In some of the issues, a Web Data Extraction tool has to regularly extract data from a Web Data source which can evolve over time. Web sources are continuously budding and structural changes happen with no notification thus are irregular. Eventually, in real-world situations it appear the need of maintaining these systems, that might stop functioning correctly if lacking of edibility to detect and face structural modifications of related Web sources. Searching for information on the Web is not an easy task. Searching for personal information is sometimes even more complicated. Below are several common problems we face when trying to get personal details from the web: Majority of the Information is distributed between different sites.

- It is not updated.
- Multi-Referent ambiguity – two or more people with the same name.
- Multi-morphic ambiguity which is because one name may be referred to in different forms.

In the most popular search engine Google, one can set the target name and based on the extremely limited facilities to narrow down the search, still the user has 100% feasibility of receiving irrelevant information in the output search hits. Not only this, the user has to manually see, open, and then download their respective file which is extremely time consuming. The major reason behind this is that there is no uniform format for personal information.

3. Features Of Visual Deep Web Pages

- Users use information from Web pages which are useful for various applications.
- The designers of web page associate different types of information with distinct visual characteristics to make the information on Web pages easy to understand.
- Visual features are important for identifying special information on Web pages.
- Deep Web pages are special Web pages that contain data records retrieved from Web databases, and we imagine that there are some distinct visual features for data records and data items.
- Based on observation based on a large number of deep Web pages is consistent with this hypothesis.
- The main visual features in this section and show the statistics about the accuracy of these features at the end. Position features (PFs). These features indicate the location of the data region on a deep Web page. PF1: Data regions are always centered horizontally.PF2: The size of the data region is usually large relative to the area size of the whole page. Since the data records are the contents in focus on deep Web pages, Web page designers always have the region containing the data records centrally and conspicuously placed on pages to capture the user's attention.

- By investigating a large number of deep Web pages, we found two interes ting facts. First, data regions are always located in the middle section horizontally on deep Web pages. Second, the size of a data region is usually large when there are enough data records in the data region. The actual size of a data region may change greatly because it is not only influenced by the number of data records retrieved, but also by what information is included in each data record.
- Therefore, our VDEC [10] approach uses the ratio of the size of the data region to the size of whole deep Web page instead of the actual size.

4. Vision-Based Deep Web Data Extraction For Web Document

We present new approach for deep web clustering based capture the actual data of the deep web pages. We achieve this in the following two phases. (1) Vision based Data relevant identification (2) Deep web pages clustering[11].

In the first phase,

- A data extraction based measure is also introduced to evaluate the importance of each leaf chunk in the tree, which in turn helps us to eliminate noises in a deep Web page. In this measure, remove the surplus noise and duplicate chunk using three parameters such as hyperlink percentage, Noise score and cosine similarity. Finally, obtain the main chunk extraction process using three parameters such as Title word Relevancy, Keyword frequency based chunk selection, Position features and set of keywords are extracted from those main chunks.

In the second phase,

- By using Fuzzy c-means clustering (FCM), the set of keywords were clustered for all deep web pages.

5. Performance analysis of Vision-Based Deep Web Data Extraction For Web Document

Information extraction from web pages is an active research area. Recently, web information extraction has become more challenging due to the complexity and the diversity of web structures and representation. This is an expectable phenomenon since the Internet has been so popular and there are now many types of web contents, including text, videos, images, speeches, or flashes. The HTML structure of a web document has also become more complicated, making it harder to extract the target content. Until now, a large number of techniques have been proposed to address this problem, but all of them have inherent limitations because they are Web-page-programming-language dependent. In this paper, we present new approach for detection and removal of noisy data to extract main content information and deep web clustering that is both fast and accurate[10]. The two phases and its sub-steps are given as follows.

- **Phase 1:** Vision-based deep web data identification
  - Deep web page extraction
  - Chunk segmentation
  - Noisy chunk Removal
  - Extraction of main chunk using chunk weightage
- **Phase 2:** Web document clustering
  - Clustering process using FCM

5.1 Phase 1: Vision-Based Deep Web Data Extraction

1) Deep Web Page Extraction

The Deep web is usually defined as the content on the Web not accessible through a search on general search engines. This content is sometimes also referred to as the hidden or invisible web. The Web is a complex entity that contains information from a variety of source types and includes an evolving mix of different file types and media. It is much more than static, self-contained Web pages. In our work, the deep web pages are collected from Complete Planet (www.completeplanet.com), which is currently the largest deep web repository with more than 70,000 entries of web databases.

2) Chunk Segmentation

Web pages are constructed not only main contents information like product information in shopping domain, job information in a job domain but also advertisements bar, static content like navigation panels, copyright sections, etc. In many web pages, the main content information exists in the middle chunk and the rest of page contains advertisements, navigation links, and privacy statements as noisy data. Removing these noises will help in improving the mining of web. To assign importance to a region in a web page (W_P), we

first need to segment a web page into a set of chunks. Hence, to clean a web page, a preprocessing step called Chunk Splitting Operation (fig.2) is performed.

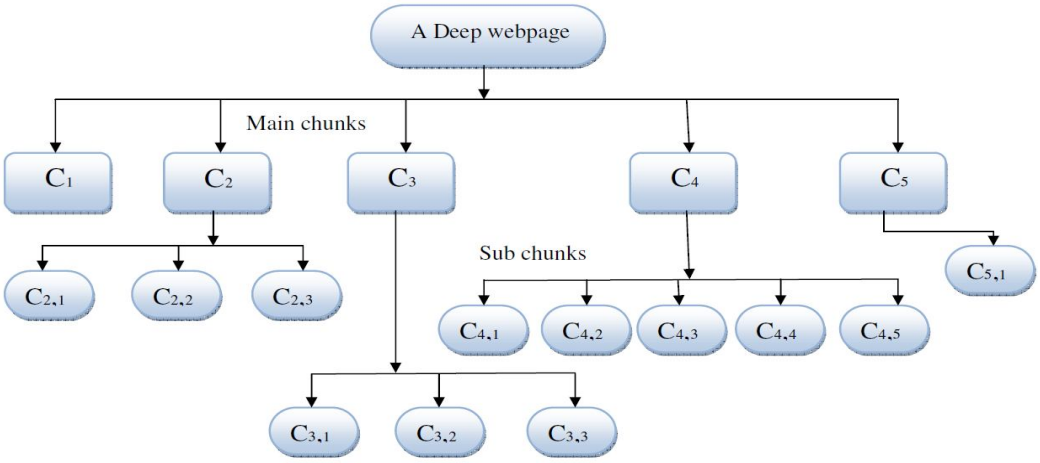

Fig. 1. The tree model of the deep web page

Basically, the layout of many web pages follows a similar pattern in such a way that the main content is enclosed in one big or <td> element which is HTML tags. In our paper, we are concentrating only the content inside the "div" tag. The tag defines a division or a section in an HTML document and it is often used to group chunk-elements. In our approach tag is consider as chunk. Normally, a tag separated by many sub tags based on the content of the deep web page. If there is no tag in the sub tag, the last tag is consider as leaf node. The Chunk Splitting Process aims at cleaning the local noises by considering only the main content of a web page enclosed in div tag. The main contents are segmented into various chunks. The resultant of this process can be represented as follows:

$$C = \{C_1, C_2, C_3, \cdots, C_n\},\ C \in DW_P$$

Where, C → A set of chunks in the deep web page DW_P

n → Number of chunks in a deep web page DW_P

In fig.1 we have taken an example of a tree sample which consists of main chunks and sub chunks. The main chunks are segmented into chunks C_1, C_2 and C_3 using Chunk Splitting Operation and sub-chunks are segmented into $C_{2,1}, C_{2,2} \ldots C_{5,1}$ in fig 2.

3) Noisy Chunk Removal

Surplus Noise Removal: A deep web page W_P usually contains main content chunks and noise chunks. Only the main content chunks represent the informative part that most users are interested in. Although other chunks are helpful in enriching functionality and guiding browsing, they negatively affect such web mining tasks as web page clustering and classification by reducing the accuracy of mined results as well as speed of processing. Thus, these chunks are called noise chunks. Removing these chunks in our research work, we have concentrated on two parameters; they are Hyperlink Percentage (HL_p) and Noise score (N_s) which is very significant. The main objective for removing noise from a Web Page is to improve the performance of the search engine.

4) Extraction of Main Chunk

Chunk Weightage for Sub-Chunk: In the previous step, we obtained a set of chunks after removing the noise chunks and duplicate chunks present in a deep web page. Web page designers tend to organize their content in a reasonable way: giving prominence to important things and deemphasizing the unimportant parts with proper features such as position, size, color, word, image, link, etc. A chunk importance model is a function to map from features to importance for each chunk, and can be formalized as:

$$\langle chunk\ \ features \rangle \Rightarrow chunk\ \text{importance}$$

The preprocessing for computation is to extract essential keywords for the calculation of Chunk Importance. Many researchers have given importance to different information inside a webpage for instance location, position, occupied area, content, etc. In our research work, we have concentrated on the three parameters Title word relevancy, keyword frequency based chunk selection, and position features which are very significant. Each parameter has its own significance for calculating sub-chunk weightage. The following equation computes the sub-chunk weightage of all noiseless chunks.

$$C_w = \alpha T_k + \beta K_f + \gamma PF_r \quad (1)$$

Where,

$\alpha, \beta, \gamma \rightarrow$ Constants

For each noiseless chunk, we have to calculate these unknown parameters T_K, K_f and PF_r. The representation of each parameter is as follows:

5.2 Phase II: Deep Web Documnet Clustering Using FCM

Let DB be a dataset of web documents, where the set of keywords is denoted by $k = \{k_1, k_2, ..., k_n\}$. Let $X = \{x_1, x_2, ..., x_N\}$ be the set of N web documents, where $x_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$. Each $x_{ij} (i = 1,...,N; j = 1,...,n)$ corresponds to the frequency of keyword x_i on web document. Fuzzy c-means [29] partitions set of N web documents in R^d dimensional space into c $(1 < c < n)$ fuzzy clusters with $Z = \{z_1, z_2, ..., z_c\}$ cluster centers or centroids. The fuzzy clustering of keywords is described by a fuzzy matrix μ with n rows and c columns in which n is the number of keywords and c is the number of clusters. μ_{ij}, the element in the i^{th} row and j^{th} column in μ, indicates the degree of association or membership function of the i^{th} object with the j^{th} cluster. The characters of μ are as follows:

$$\mu_{i,j} \in [0,1] \qquad \forall i = 1,2,...,n; \quad \forall j = 1,2,...,c; \quad (6)$$

$$\sum_{j=1}^{c} \mu_{ij} = 1 \qquad \forall i = 1,2,...,n; \quad (7)$$

$$0 < \sum_{i=1}^{n} \mu_{ij} < n \quad \forall j = 1,2,...,c; \quad (8)$$

The objective function of FCM algorithm is to minimize the Eq. (9):

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij}^m d_{ij} \quad (9)$$

Where

$$d_{ij} = \| k_i - z_j \| \quad (10)$$

in which, m(m >1) is a scalar termed the weighting exponent and controls the fuzziness of the resulting clusters and d_{ij} is the Euclidian distance from k_i to the cluster center z_i.The z_j,centroid of the j^{th} cluster, is obtained using Eq. (11)

$$z_j = \frac{\sum_{i=1}^{n} \mu_{ij}^m k_i}{\sum_{i=1}^{n} \mu_{ij}^m} \quad (11)$$

The FCM algorithm is iterative and can be applied.

6. Results and Discussion

6.1 Experimental set up

The experimental results of the proposed method for vision-based deep web data extraction for web document clustering are presented in this section. The proposed approach has been implemented in java (jdk 1.6) and the experimentation is performed on a 3.0 GHz Pentium PC machine with 2 GB main memory. For experimentation, we have taken many deep 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. These pages are then

applied to the proposed method for removing the different noises. The removal of noise blocks and extracting of useful content chunks are explained in this sub-section. Finally, extracting the useful content keywords are clustered using Fuzzy c-means clustering.

6.2. Data Sets

GDS: Our data set is collected from the complete planet web site (www.completeplanet.com). Complete-planet is currently the largest depository for deep web, which has collected the search entries of more than 70,000 web databases and search engines. These Web databases are classified into 42 categories covering most domains in the real world. GDS contains 1,000 available Web databases. For each Web database, we submit five queries and gather five deep Web pages with each containing at least three data records. **SDS:** Special data set (SDS). During the process of obtaining GDS, we noticed that the data records from two-thirds of the Web databases have less than five data items on average. To test the robustness of our approaches, we select 100 Web databases whose data records contain more than 10 data items from GDS as SDS.

6.3. Performance Measures

1) Data extraction evaluation

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

Revision is defined to be the percentage of the Web databases whose data records or data items are not perfectly extracted, i.e., either precision or recall is not 100 percent.

$$\text{Revision} = \frac{WDB_t - WDB_c}{WDB_t}$$

Where,

WDB_c is the total number of web sites whose precision and recall are both 100%.

WDB_t is total number of web sites processed.

2) Clustering evaluation

Clustering Accuracy, $CA = \frac{1}{N}\sum_{i=1}^{T} X_i$

Where, $N \rightarrow$ Number of data points in the dataset

$T \rightarrow$ Number of resultant cluster

$X_i \rightarrow$ Number of data points occurring in both cluster i and its corresponding class.

6.4 Experimental results

The sample of results obtained by the proposed approach is given in this sub-section. A sample of deep webpage considered for experimentation is shown in Fig. 2. Then, the sample web page is subjected to the proposed approach to identify the relevant web data region. Data region having description of some products is extracted by our data extraction method after removing the noises. The filtered data region is shown in Fig. 3.

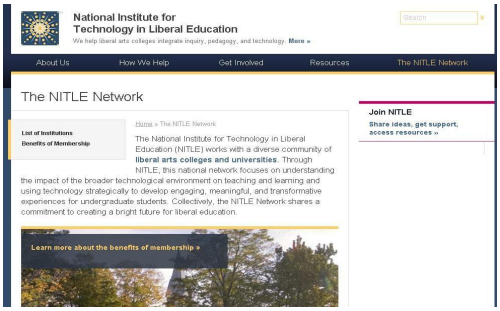

Fig. 2 A Sample Deep Web Page

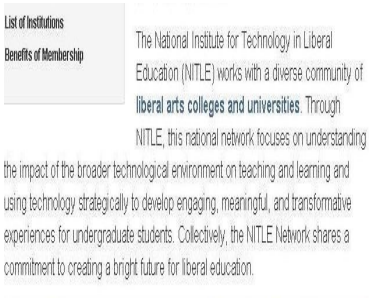

Fig. 3 Filter Data Region

6.5 Performance analysis of phase 1 of our technique

The performance analyses of the three methods on GDS and SDS datasets are presented in this section. Table 1 shows the experimental results on both GDS and SDS. Totally, VDEC performs significantly better than MDR on both GDS and SDS. But in another case, VDEC slightly get slipped in performance evaluation with ViDRE. *Precision:* The precision values of three methods are plotted as a graph shown in Fig 5, in which our proposed method VDEC performs better precision value (99.2% and 99.1%) compared with MDR in both datasets. *Recall:* The recall values obtained for three different methods are plotted in the figure 6, in which our VDEC performs better recall value (98.4% and 97.4%) compared with MDR in both datasets. *Revision:* By analyzing the figure 7, VDEC is better revision value (12% and 8%) compared with MDR in both datasets.

Table 1: Performance comparison of ViDRE, MDR and VDEC on two data sets.

	Data set		Precision	Recall	Revision
ViDRE	GDS		98.7%	97.2%	12.4%
	SDS		98.5%	97.8%	10.9%
MDR	GDS		85.3%	53.2%	55.2%
	SDS		78.7%	47.3%	63.8%
VDEC	GDS	$\alpha = .5, \beta = .5, \gamma = .5$	92.2%	91.9%	24%
	SDS	$\alpha = .5, \beta = .5, \gamma = .5$	93.6%	90.1%	16%
	GDS	$\alpha = .5, \beta = .5, \gamma = .1$	99.2%	97.4%	8%
	SDS	$\alpha = .5, \beta = .5, \gamma = .1$	99.1%	98.4%	12%

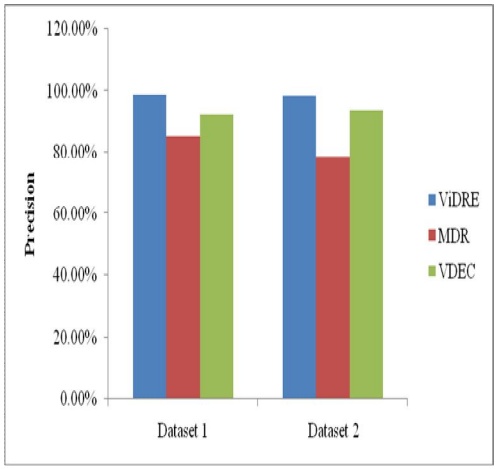

Figure 4.Precision graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .5$

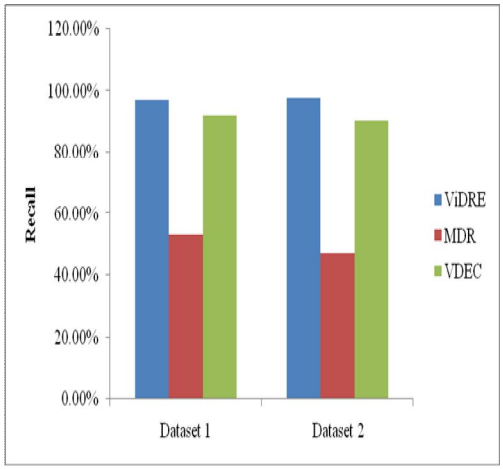

Figure 5.Recall graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .5$

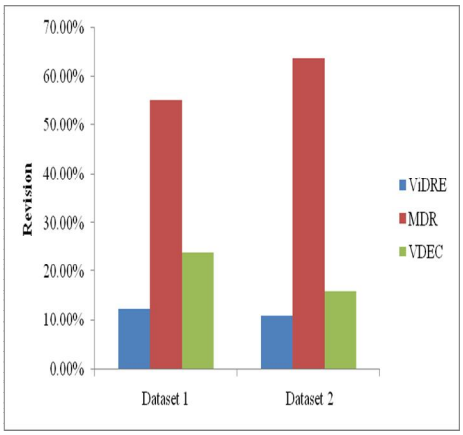

Figure 6.Revision graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .5$

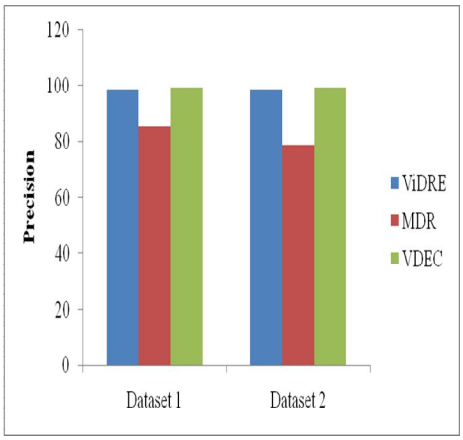

Figure 7.Precision graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .1$

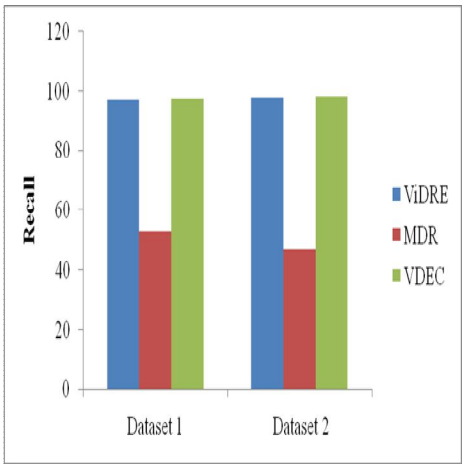

Figure 8.Recall graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .1$

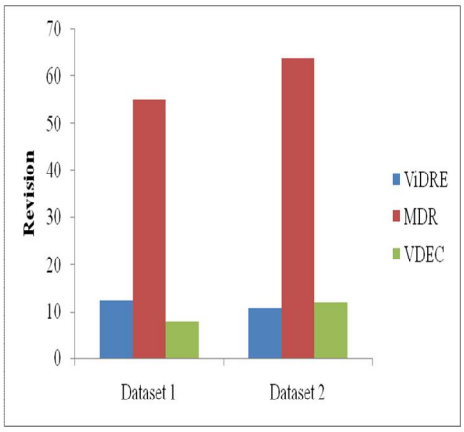

Figure 9.Revision graph of dataset 1 and dataset 2 for parameters $\alpha = .5, \beta = .5, \gamma = .1$

8. Conclusion

In this paper, we have implemented a new approach called vision-based deep web data extraction for web document clustering. In this paper, an approach to vision-based deep web data extraction is proposed for web document clustering. The proposed approach comprises of two phases: 1) Vision-based web data extraction, and 2) web document clustering. In phase 1, the web page information is classified into various chunks. From which, surplus noise and duplicate chunks are removed using three parameters, such as hyperlink percentage, noise score and cosine similarity. To identify the relevant chunk, three parameters such as Title word Relevancy, Keyword frequency-based chunk selection, Position features are used and then, a set of keywords are extracted from those main chunks. Finally, the extracted keywords are subjected to web document clustering using Fuzzy c-means clustering (FCM). Our experimental results showed that the proposed VDEC method can achieve stable and good results of about 99.2% and 99.1% precision value in both datasets.

References

[1] P S Hiremath, Siddu P Algur,"Extraction of data from web pages: a vision based approach," International Journal of Computer and Information Science and Engineering, Vol.3, pp.50-59, 2009.
[2] Jing Li, "Cleaning Web Pages for Effective Web Content Mining, "In Proceedings: DEXA, 2006.
[3] Thanda Htwe,"Cleaning Various Noise Patterns in Web Pages for Web Data Extraction," International Journal of Network and Mobile Technologies,vol.1,no.2,2010.
[4] Yang, Y. and Zhang, H., "HTML Page Analysis Based on Visual Cues," In 6th International
Conference on Document Analysis and Recognition, Seattle, Washington, USA, 2001.

[5] Longzhuang Li, Yonghuai Liu, Abel Obregon," Visual Segmentation-Based Data Record Extraction from Web Documents,"IEEE International Conference on Information Reuse and Integration, pp.502 – 507, 2007.
[6] 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.
[7] R. B. Yates and B. R. Neto,"Modern Information Retrieval," Addison-Wesley, New York, 1999.
[8] B. Larsen and C. Aone. "Fast and effective text mining using linear-time document clustering," In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.
[9] 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.
[10] M.Lavanya, Dr.M.Usha rani. "vision-based deep web data extraction for web document clustering" Global Journals Inc..,March 2012.

[11] M.Lavanya, Dr.M.Usha rani. "A Frame Work For Vision-Based Deep Web Data Extraction For Web Document Clustering ", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 7, September - 2012 ISSN: 2278-0181.
[12] Wei Liu, Xiaofeng Meng, Weiyi Meng,"ViDE: A Vision-Based Approach for Deep Web Data Extraction,"IEEE Transactions on Knowledge and Data Engineering, vol.22, no.3, pp.447-460, 2010.
[13] 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.
[14] Manisha Marathe, Dr. S.H.Patil, G.V.Garje,M.S.Bewoor, "Extracting Content Blocks from Web Pages", International Journal of Recent Trends in Engineering, Vol .2, No. 4, November 2009.
[15] Sandip Debnath,Prasenjit Mitra,C. Lee Giles, "Automatic Extraction of Informative Blocks from WebPages", In Proceedings of the ACM symposium on Applied computing, Santa Fe, New Mexico, pp. 1722 – 1726,2005.
[16] Lan Yi ,Bing Liu, "Web page cleaning for web mining through feature weighting", In Proceedings of the 18th international joint conference on Artificial intelligence, pp. 43-48 , August 09 - 15 ,Acapulco, Mexico, 2003
[17] A. K. Tripathy , A. K. Singh , "An Efficient Method of Eliminating Noisy Information in Web Pages for Data Mining", In Proceedings of the Fourth International Conference on Computer and Information Technology, pp. 978 – 985, 2004.
[18] Zhao Cheng-li and Yi Dong-yun, "A method of eliminating noises in Web pages by style tree model and its applications", Wuhan University Journal of Natural Sciences, Wuhan University, co-published with Springer Vol.9, No.5, pp. 611-616, 2004.
[19] Ruihua Song,Haifeng Liu,Ji-Rong Wen, Wei-Ying Ma, "Learning Block Importance Models for Web Pages", Proceedings of the 13th international conference on World Wide Web, pp. 203 - 211 , New York, NY, USA, 2004.
[20] Ruihua Song, Haifeng Liu, Ji-Rong Wen, Wei-Ying Ma, "Learning Important Models for Web Page Blocks based on Layout and Content Analysis", ACM SIGKDD Explorations Newsletter, Vol. 6 , No. 2,pp. 14 - 23 ,2004.
[21] Liu, B., Grossman, R. and Zhai, Y," Mining Data Records in Web Pages," KDD-03, pp. 49-55, 2003.
[22] Zhai, Y., Liu, B,"Web Data Extraction Based on Partial Tree Alignment," Proceedings of the 14th international conference on World Wide Web,pp.76-85,2005.
[23] J. Hammer, H. Garcia Molina, J. Cho, and A. Crespo. "Extracting semistructured information from the web". In Proc.of the Workshop on the Management of Semi-structured Data, 1997.
[24] Hesam Izakian, Ajith Abraham,"Fuzzy C-means and fuzzy swarm for fuzzy clustering problem,"Computer and Information Science, vol.38, no.3, pp.1835-1838, 2011.
[25] Webb, A." Statistical pattern recognition," New Jersey: John Wiley & Sons, 2002.
[26] Tan, P. N., Steinbach, M., & Kumar, V." Introduction to data mining, "Boston: Addison-Wesley, 2005.
[27] Pang, W., Wang, K., Zhou, C., and Dong, L." Fuzzy discrete particle swarm optimization for solving traveling salesman problem," In Proceedings of the fourth international conference on computer and information technology,pp. 796–800, IEEE CS Press,2004.
[28] Hathway, R. J., & Bezdek, J."Optimization of clustering criteria by reformulation," IEEE transactions on Fuzzy Systems, 241–245, 1995.
[29] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics, Vol. 3, pp.32-57, 1973.

Ms. M. Lavanya obtained Bachelor's degree in Sciences (Computer Science) from S.V.University, Tirupathi. Then she obtained her Master's degree in Computer Applications from S.V.University. She is working as Assistant Professor[SL] in the Department of Master of Computer Applications at Sree Vidyanikethan Engineering College, A.Rangampet, Tirupati. She is pursuing her Ph.D. in Computer Science in the area of Data Warehousing and Data Mining. She is in teaching since 2003. She presented many papers at National and Internal Conferences and published articles in National & International journals.

Dr. M. Usha Rani is an Associate Professor in the Department of Computer Science and HOD for MCA, Sri Padmavati Mahila Viswavidyalayam (SPMVV Woman's' University), Tirupati. She did her Ph.D. in Computer Science in the area of Artificial Intelligence and Expert Systems. She is in teaching since 1992. She presented many papers at National and Internal Conferences and published articles in national & international journals. She also has written 4 books like Data Mining - Applications: Opportunities and Challenges, Superficial Overview of Data Mining Tools, Data Warehousing & Data Mining and Intelligent Systems & Communications. She is guiding M.Phil. and Ph.D. in the areas like Artificial Intelligence, Data Warehousing and Data Mining, Computer Networks and Network Security etc.