Global Journals Inc.

Global Journal of Computer Science and Technology
Volume 12 Issue 5 Version 1.0 March 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Vision-Based Deep Web Data Extraction for Web Document Clustering

By M. Lavanya & Dr.M.Usha Rani
*Sree Vidyanikethan Engineering College*

*Abstract* - The design of web information extraction systems becomes more complex and time-consuming. Detection of data region is a significant problem for information extraction from the web page. 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 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. Finally, the extracted keywords are subjected to web document clustering using Fuzzy c-means clustering (FCM).

*Keywords :* *Noise Chunk, cosine similarity, Title word Relevancy, Keyword frequency-based chunk selection, Fuzzy c-means clustering (FCM)*

*GJCST Classification:* *I.4.m, H.2.8*

# Vision-Based Deep Web Data Extraction for Web Document Clustering

M. Lavanya [α] & Dr.M.Usha Rani [σ]

*Abstract* - The design of web information extraction systems becomes more complex and time-consuming. Detection of data region is a significant problem for information extraction from the web page. 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 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. Finally, the extracted keywords are subjected to web document clustering using Fuzzy c-means clustering (FCM).

*Keywords : Noise Chunk, cosine similarity, Title word Relevancy, Keyword frequency-based chunk selection, Fuzzy c-means clustering (FCM)*

## I. Introduction

Today, World Wide Web has become one of the most significant information resources. Though most of the information is in the form of unstructured text, a huge amount of semi-structured objects, called data records, are enclosed on the Web [5]. Due to the heterogeneity and lack of structure of Web information, automated discovery of relevant information becomes a difficult 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[4]. 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. These special Web pages are generated dynamically and are difficult 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.

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]. 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.

Clustering is a technique, in which the data objects are given into a set of disjoint groups called clusters so that objects in each cluster are more analogous to each other than the objects from different clusters. Clustering techniques are used in several application areas such as pattern recognition (Webb, 2002), data mining (Tan, Steinbach, & Kumar, 2005), machine learning (Alpaydin, 2004), and so on. Generally, clustering algorithms can be classified as Hard, Fuzzy, Possibilistic, and Probabilistic[2] (Hathway & Bezdek, 1995).

In this paper a novel method to extract data items from the deep web pages automatically is proposed. It comprises of two steps: (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[3]. 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. These sub-chunks consider as the main chunk and the keywords are extracted from those main chunk. Secondly, the set of keywords are clustered using Fuzzy c-means clustering.

The paper is organized as follows. Section 2 presents the related works. The problem statement is described in section 3 and the contribution of this paper is given in section 4. The definition of terms used in the proposed approach given in section 5. An efficient approach web document clustering based on vision-based deep web is discussed in section 6. The

Author [α] : Sr. Lecturer, SVEC, Tirupati Andhrapradesh, INDIA-51
E-mail : lavanya4_79@rediffmail.com
Author [σ] : Assoc .Prof, Dept. of C S, SPMVV, Tirupati , Andhrapradesh, INDIA- 517102
E-mail : musha_rohan@yahoo.com

experimental results are reported in Section 7. Section 8 explains conclusion of the paper.

## II. Review of Related Works

Our proposed method concentrates on web document clustering based on vision-based deep web data extraction. Many Researchers have developed several approaches for web document clustering based on vision-based deep web data[7]. Among them, a handful of significant researches that performs web clustering and data extraction are presented in this section.

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[8]. Then, the vision character of page block has been analyzed and finally determined the topic data region accurately.

ECON can be applied to Web news pages written in several well known languages namely Chinese, English, French, German, Italian, Japanese, Portuguese, Russian, Spanish, and Arabic. Also, ECON can been implemented without any difficulty. *Wei Liu et al [12]* have introduced a vision-based approach that is Web-page programming- language-independent for deep web data extraction. Mainly, the proposed approach has used the visual features on the deep Web pages to implement deep Web data extraction, such as data record extraction and data item extraction[11]. They have also proposed an evaluation measure revision to gather the amount of human effort required to produce proper extraction.

## III. Problem Statement

In a web page, there are numerous immaterial components related with the descriptions of data objects. These items comprise advertisement bar, product category, search panel, navigator bar, and copyright statement, etc. Generally, a web page $W_P$ is specified by a triple $W_P = (\omega, \phi, \eta)$. $\varpi = \{W_{p1}, W_{p2} ... W_{pn}\}$ is a finite set of objects or sub-web pages. All these objects are not overlapped. Each web page can be recursively viewed as a sub-web-page and has a subsidiary content structure. $\phi = \{\phi_1, \phi_2 ... \phi_n\}$ is a finite set of visual separators, such as horizontal separators and vertical separators. Every separator has a weight representing its visibility, and all the separators in the same $\phi$ have same weight. $\eta$ is the relationship of every two blocks in $\varpi$, which is represented as: $\eta = \varpi \times \varpi \rightarrow \phi \cup \{NULL\}$. In several web pages, there are normally more than one data object entwined together in a data region, which makes it complex to find the attributes for each page. Also, since the raw source of the web page for representing the objects is non-contiguous one, the problem becomes more complicated. In real applications, the users necessitate from complex web pages is the description of individual data object derived from the partitioning of data region.

## VI. Contribution of The Paper

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.

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.

## VII. Definitions of Terms Used In the Proposed Approach

**Definition (chunk $C$):** Consider a deep web page $DW_P$ is segmented by blocks. These each blocks are known as chunk.

For example the web page is represented as, $DW_P = C_1, C_2, C_3 ... C_n$ ,Where the main chunk , $C_1 = C_{1,1}, C_{1,2} ... C_{m,n}$ .

**Definition** (Hyperlink $(HL_P)$): A hyperlink has an anchor, which is the location within a document from which the hyperlink can be followed; the document having a hyperlink is called as its source document to web pages.

Hyperlink percentage $HL_P = \frac{n_l}{N}$

Where,

$N$ → Number of Keywords in a chunk

$n_l$ → Number of Link Keywords in a chunk

**Definition** (Noise score $(N_s)$): Noise score is defined as the ratio of number of images to total number of chunks.

Noise score, $N_s = \frac{n_I}{N_B}$

Where, $n_I$ → Number of images in a chunk

$N_B$ → Total number of images

**Definition (Cosine similarity):** Cosine similarity means calculating the similarity of two chunks. The inner product of the two vectors i.e., sum of the pairwise multiplied elements, is divided by the product of their vector lengths.

Cosine Similarity, $SIM_c C_1, C_2 = \frac{|C_1.C_2|}{|C_1| \times |C_2|}$ where

$C_1, C_2$ →Weight of keywords in $C_1$, $C_2$

**Definition (Position feature):** Position features (PFs) that indicate the location of the data region on a deep web page. To compute the position feature score, the ratio $(T)$ is computed and then, the following equation is used to find the score for the chunk.

$$PF_r = \begin{cases} 1 & 0.7 \geq T \\ 0 & Otherwise \end{cases} \qquad (4)$$

Where,

$$T \rightarrow \frac{\text{Number of keywords in Dtata Re gion chunk}}{\text{Number of keywords in Whole web page}}$$

$PF_r$ → Position features

**Definition** (Title word relevancy): A web page title is the name or heading of a Web site or a Web page. If there is more number of title words in a certain block, then it means that the corresponding block is of more importance.

Title word relevancy, $T_K = 1 - \left[ \frac{m_k}{\left( m_k + \sum_{i=1}^{|m_k|} F(m_k^{(i)}) \right)} \right]$

Where,

$m_k$ → Number of Title Keywords

$F(m_k^{(i)})$ → Frequency of the title keyword $m_k$ in a chunk

**Definition (Keyword frequency):** Keyword frequency is the number of times the keyword phrase appears on a deep Web page chunk relative to the total number of words on the deep web page.

Keyword frequency based chunk selection, $K_f = \sum_{k=1}^{K} \frac{f_k}{N}$

Where,

$f_k$ → Frequency of top ten keywords

$N$ → Number of keywords

$k$ → Number of Top-K Keywords

## VIII. Proposed Approach to Vision-Based Deep Web Data Extraction for Web Document Clustering

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

## IX. Block Diagram

In the first phase, we are mainly concentrating to remove the following noises in stages: (1) Navigation bars, Panels and Frames, Page Headers and Footers, Copyright and Privacy Notices, Advertisements and Other Uninteresting Data. (2) Duplicate Contents and (3) Unimportant Contents according to chunk importance. The removal of these noises is done by performing three operations. Firstly, using the chunk segmentation process, the noises such as the advertisements, images, audio, video, multiple links etc. are removed and only the useful text contents are segmented into chunks. Secondly, using three parameters such as *hyperlink percentage, Noise score* and *cosine similarity*, the surplus noise and duplicate chunks are removed to obtain the noiseless sub-chunks. And lastly, for each noiseless sub-chunk, we considered three parameters such as *Title word Relevancy ,Keyword frequency based chunk selection,* and *Position features*, using which we calculated the Sub-chunk weightage of each and every chunk. The high importance of the sub-chunks weightage consider as main-chunk weightage and the keywords are extracted from those main chunk. In the second phase, the set of keywords extracted are subjected to Fuzzy c-means clustering (FCM). The system model of the proposed technique which is extracting the important chunks and deep web clustering is shown schematically in Fig 1.

### a) *Phase 1: Vision-Based Deep Web Data Extraction*

#### i. *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.

#### ii. *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.

extract main content information and deep web clustering that is both fast and accurate. 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

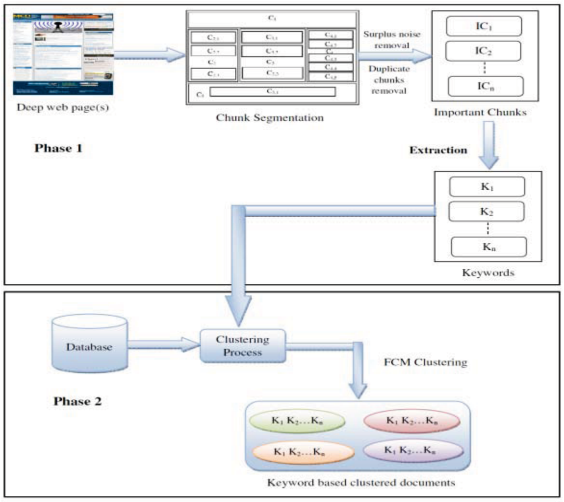

*Fig. 1.* Proposed method for extracting the important chunks and web clustering

Normally, a <div>tag separated by many sub <div> tags based on the content of the deep web page. If there is no <div>tag in the sub <div>tag, the last <div>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 $C_{2,1}, C_{2,2} ... C_{5,1}$ in fig 2.

#### iii. *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.

The representation of each parameter is as follows:

1. *Hyperlink Keyword* $(HL_p)$ – A hyperlink has an anchor, which is the location within a document from which the hyperlink can be followed; the document containing a hyperlink is known as its source document to web pages. Hyperlink Keywords are the keywords which are present in a chunk such that it directs to another page. If there are more links in a particular chunk then it means the corresponding chunk has less importance. The parameter Hyperlink Keyword Retrieval calculates the percentage of all the hyperlink keywords present in a chunk and is computed using following equation.

$$\textit{Hyperlink word Percentage, } HL_P = \frac{n_l}{N}$$

Where,

$N$ → Number of Keywords in a chunk

$n_l$ → Number of Link Keywords in a chunk

2. *Noise score* $(N_s)$ – The information on Web page $W_P$ consists of both texts and images (static pictures, flash, video, etc.). Many Internet sites draw income from third-party advertisements, usually in the form of images sprinkled throughout the site's pages. In our work, the parameter Noise score calculates the percentage of all the images present in a chunk and is computed using following equation. $\textit{Noise score, } N_s = \dfrac{n_I}{N_B}$

Where,

$n_I$ → Number of images in a chunk

$N_B$ → Total number of images

***Duplicate Chunk Removal Using Cosine Similarity:*** *Cosine Similarity*: Cosine similarity is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval applications [7] and clustering too [8]. Here, duplication detection among the chunk is done with the help of cosine similarity.

Given two chunks $C_1$ and $C_2$, their cosine similarity is

$$\text{Cosine Similarity} \quad SIM_c(C_1, C_2) = \frac{|C_1.C_2|}{|C_1| \times |C_2|}$$

Where,

$C_1$ , $C_2$ →Weight of keywords in $C_1$, $C_2$

iv. *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: $\langle \textit{chunk features} \rangle \Rightarrow \textit{chunk}\ \text{importance}$

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.

$$C_w = \alpha T_k + \beta\ K_f + \ \gamma\ PF_r \tag{1}$$

Where $\alpha, \beta, \gamma$ → 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:

1. *Title Keyword* – Primarily, a web page title is the name or title of a Web site or a Web page. If there is more number of title words in a particular block then it means the corresponding block is of more importance. This parameter Title Keyword calculates the percentage of all the title keywords present in a block. It is computed using following equation.

   *Title word Relevancy*; (2)

$$T_K = 1 - \left[ \frac{m_k}{\left( m_k + \sum_{i=1}^{|m_k|} F({m_k}^{(i)}) \right)} \right]$$

Where, $m_k$ → Number of Title Keywords

$T_k$ → Title word relevancy, $F({m_k}^{(i)})$ → Frequency of the title keyword $n_t$ in a chunk

2. *Keyword Frequency based chunk selection:* Basically, Keyword frequency is the number of times the keyword phrase appears on a deep Web page chunk relative to the total number of words on the deep web page. In our work, the top-K keywords of each and every chunk were selected and then their frequencies were calculated. The parameter

keyword frequency based chunk selection calculates for all sub-chunks and is computed using following equation.

*Keyword Frequency based chunk selection*

$$K_f = \sum_{k=1}^{K} \frac{f_k}{N} \qquad (3)$$

Where,

$f_k$ → Frequency of top ten keywords

$K_f$ → Keyword Frequency based chunk selection

$k$ → Number of Top-K Keywords

3. *Position features (PFs):* Generally, these data regions are always centered horizontally and for calculating, we need the ratio $(T)$ of the size of the data region to the size of whole deep Web page instead of the actual size. In our experiments, the threshold of the ratio is set at 0.7, that is, if the ratio of the horizontally centered region is greater than or equal to 0.7, then the region is recognized as the data region. The parameter position features calculates the important sub chunk from all sub chunk and is computed using following equation.

$$PF_r = \begin{cases} 1 & 0.7 \geq T \\ 0 & Otherwise \end{cases} \qquad (4)$$

Where,

$$T \rightarrow \frac{\textit{Number of keywords in Dtata Re gion chunk}}{\textit{Number of keywords in Whole web page}}$$

$PF_r$ → Position features

Thus, we have obtained the values of $T_K$, $K_f$ and $PF_r$ by substituting the above mentioned equation. By substituting the values of $T_K$, $K_f$ and $PF_r$ in eq.1, we obtain the sub-chunk weightage.

***Chunk Weightage for Main Chunk:*** We have obtained sub-chunk weightage of all noiseless chunks from the above process. Then, the main chunks weightage are selected from the following equation

$$C_i = \sum_{i=1}^{n} \alpha \, c_w^{(i)} \qquad (5)$$

Where, $c_w^{(i)}$ → $i^{th}$ Sub-chunk weightage of Main-chunk. $\alpha$ → Constant, $C_i$ → Main chunk weightage

Thus, finally we obtain a set of important chunks and we extract the keywords from the above obtained important chunks for effective web document clustering mining.

```
Input   :  Deep web page $W_p$ , Input Constants $\alpha, \beta, \gamma$
Output  :  Set of cluster with relevant keywords $CS_k$
Pseudo code

1. Input the deep web pages, $W_P$
2. Deep web page segmentation using <div> tag
3. Compute Noise chunk removal value for each leaf nodes of deep web pages
    3.1 Compute surplus noise removal for all leaf nodes using Hyperlink
        Percentage and noise score.
        3.1.1 Compute Hyperlink word Percentage, $HL_P = \frac{n_l}{N}$
        3.1.2 Compute Noise score, $N_s = \frac{n_I}{N_B}$
    3.2 Compute duplicate noise removal for all leaf nodes using cosine similarity
        $SIM_c C_1, C_2 = \frac{|C_1.C_2|}{|C_1| \times |C_2|}$
4. Compute sub-chunk weightage value $SC_w$ for each Noiseless chunk of deep Web pages
    4.1 Compute Title word relevancy for noiseless chunk
        $T_K = 1 - \left[ \frac{n_t}{\left( n_t + \sum_{i=1}^{|n_t|} F(n_t^{(i)}) \right)} \right]$
    4.2 Compute Keyword frequency based chunk importance
        $K_f = \frac{\sum_{k=1}^{TopK} f_k}{N}$
    4.3 Compute Position features based chunk importance
        $PF_r = \begin{cases} 1 & 0.7 \geq T \\ 0 & Otherwise \end{cases}$
    4.4 Compute sub-chunk weightage
```

$$SC_w = \alpha H_f + \beta F_f + \gamma PF_r$$

5. Compute main-chunk weightage value $M_i$ , i.e. set of keywords

$$M_i = \sum_{i=1}^{n} \alpha e_w^{(i)}$$

6. After computing the extraction of keywords for all deep web pages, set of

Keywords were clustered using Fuzzy c-means clustering.

$$CS_k = \{ CS_{k1}, CS_{k2} ... CS_{kn} \}$$

### b) 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 $\mu$ with $n$ rows and $c$ columns in which $n$ is the number of keywords and $c$ is the number of clusters. $\mu_{ij}$ , the element in the $i^{th}$ row and $j^{th}$ column in $\mu$ , indicates the degree of association or membership function of the $i^{th}$ object with the $j^{th}$ cluster. The characters of $\mu$ are as follows:

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

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

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

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

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

Where

$$d_{ij} = \left\| k_i - z_j \right\| \tag{10}$$

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)

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

The FCM algorithm is iterative and can be stated as follows

Algorithm 2.Fuzzy c-means:

1. Select m (m > 1); initialize the membership function values $\mu_{ij}$ , $i = 1,2,...,n$ ; $j = 1,2,...,c.$
2. Compute the cluster centers $z_j$ , $j = 1,2,...,c,$ according to Eq. (11).
3. Compute Euclidian distance $d_{ij}$ , $i = 1,2,...,n$ ; $j = 1,2,...,c.$
4. Update the membership function $\mu_{ij}$ , $i = 1,2,...,n; j = 1,2,...,c$ according to Eq. (12).

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}}} \tag{12}$$

## X. Results and Discussion

### a) 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.

### *b) 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.
Experimental results

### *c) Performance Analysis of Phase 2 of Our Technique*

While analyzing the results of GDS and SDS datasets, the accuracy, computation time and the memory usage is evaluated in clustering process. Accuracy: The accuracy values obtained for two different datasets are plotted in the figure 8, in which the dataset 1 and dataset 2 are achieves better accuracy (40%) in Fig 8. Execution Time: The run time performance of the methods is plotted as a graph shown in Fig 9, in which the dataset 2 achieves better execution time (0.906 sec) compared with data set 1. Memory usage: By analyzing the figure 10, the dataset 1 are achieves more memory usage (2990 KB) compared with dataset 2.

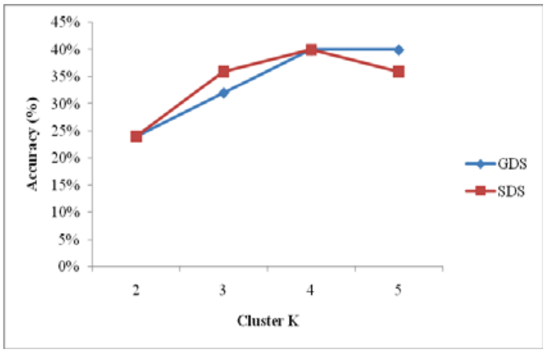

*Figure 2 :* clustering accuracy of dataset 1(GDS) and data set 2(SDS)

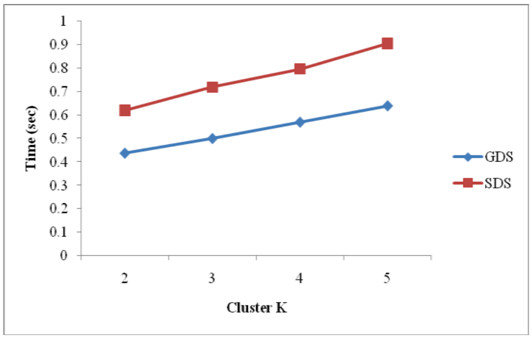

*Figure 3 :* Time of dataset 1(GDS) and data set 2(SDS)

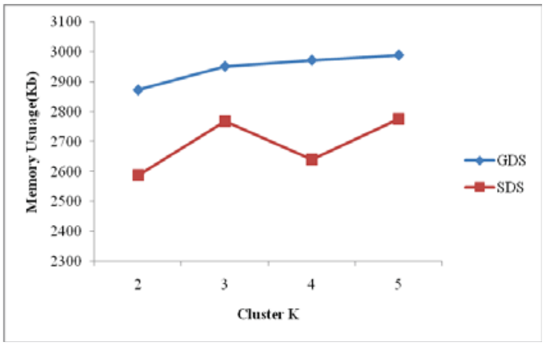

*Figure 4 :* Memory usage of dataset 1(GDS) and dataset 2(SDS)

## XI. Conclusion

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 for both datasets.

## References Références Referencias

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. 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.
11. Yan Guo, Huifeng Tang, Linhai Song, Yu Wang, Guodong Ding, “ECON: An Approach to Extract Content from Web News Page”, In Proceedings of the 12th International Asia-Pacific Web Conference (APWEB), pp. 314 – 320, April 06-April 08 Buscan, Korea, 2010
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. 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. 2-57, 1973.

This page is intentionally left blank