Vision-Based Deep Web Data Extraction for Web Document Clustering

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

[1] Sree Vidyanikethan Engineering College

Received: 7 April 2012 Accepted: 1 May 2012 Published: 15 May 2012

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).

Index terms— Noise Chunk, cosine similarity, Title word Relevancy, Keyword frequency-based chunk selection, Fuzzy c-means clustering (FCM)

oday, 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 Author ? : Sr. Lecturer, SVEC, Tirupati Andhrapradesh, INDIA-51 E-mail : lavanya4_79@rediffmail.com Author ? : Assoc .Prof, Dept. of C S, SPMVV, Tirupati , Andhrapradesh, INDIA-517102 E-mail : musha_rohan@yahoo.com 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 visionbased deep web is discussed in section 6. The experimental results are reported in Section 7. Section 8 explains conclusion of the paper.

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 algorithmbased 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 multifeatures 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.

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. 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.

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 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.

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 (P W), 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 The representation of each parameter is as follows:

1. Hyperlink Keyword p HL -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. 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.

3 Hyperlink word

Given two chunks C and C , their cosinesimilarity is Cosine Similarity C C C C C C SIM c

Where,

4 C , C

Weight of keywords in C , C iv.

5 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:

6 chunk features chunk

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.r f k w PF K T C (1)

Where Constants

For each noiseless chunk, we have to calculate these unknown parameters K T , f K and r PF . 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.

7 Title word Relevancy;

k m i i k k k K m F m m T (2)

Where, 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. Let DB be a dataset of web documents, where the set of keywords is denoted by

8 Otherwise

n k k k k . Let N x x x X

be the set of N web documents, where in i i i x x x x .Each n j N i x ij

9 c j

The objective function of FCM algorithm is to minimize the Eq. (**??**): cluster, is obtained using Eq. (**??**1)n i m ij n i i m ij j k z (11)

The FCM algorithm is iterative and can be stated as follows Algorithm 2.Fuzzy c-means: 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. [1] [2] [3] [4]

Figure 1: N

Figure 2: k

Figure 3:

166

[1] © 2012 Global Journals Inc. (US)
[2] © 2012 Global Journals Inc. (US)
[3] © 2012 Global Journals Inc. (US) Global Journal of Computer Science and Technology Volume XII Issue V Version I
[4] © 2012 Global Journals Inc. (US)

Figure 4: -

Figure 5: k

446 II.

Figure 6: 4 . 4 6 .

Figure 7: x

Figure 8:

Figure 9:

Figure 10:

24 R

Figure 11: Figure 2 :Figure 4 :

specified by W is a trip
W

W p W p 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 cont
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

have
sam
weig

is the relationship of every two blocks in , wh
is

represented as:

Figure 12:

[Li] , Qingshui Li .

[Hong and Zhou] , Chen Hong , -Ping; Fang Wei; Yang Zhou . (Zhuo Lin)

[Cheng et al. ()] *A method of eliminating noises in Web pages by style tree model and its applications*, Zhao Cheng , - Li , Yi Dong-Yun . 2004. Springer. 9 p. . Wuhan University Journal of Natural Sciences, Wuhan University

[Tripathy and Singh ()] 'An Efficient Method of Eliminating Noisy Information in Web Pages for Data Mining'. K Tripathy , A K Singh . *Proceedings of the Fourth International Conference on Computer and Information Technology*, (the Fourth International Conference on Computer and Information Technology) 2004. p. .

[Zhi-Ming ()] 'Automatic Data Records Extraction from List Page in Deep Web Sources'. Cui Zhi-Ming . *Asia-Pacific Conference on Information Processingvol* 2009. 1 p. .

[Debnath et al. ()] 'Automatic Extraction of Informative Blocks from WebPages'. Sandip Debnath , Prasenjit Mitra , C Lee Giles . *Proceedings of the ACM symposium on Applied computing*, (the ACM symposium on Applied computingSanta Fe, New Mexico) 2005. p. .

[Pei-Ying and Cun-He ()] 'Automatic text summarization based on sentences clustering and extraction'. Zhang Pei-Ying , Li Cun-He . *2nd IEEE International Conference on Computer Science and Information Technology*, 2009. p. .

[Guo et al. (2010)] 'ECON: An Approach to Extract Content from Web News Page'. Yan Guo , Huifeng Tang , Linhai Song , Yu Wang , Guodong Ding . *Proceedings of the 12th International Asia-Pacific Web Conference (APWEB)*, (the 12th International Asia-Pacific Web Conference (APWEB)) April 06-April 08 Buscan, Korea, 2010. p. .

[Ashraf et al. (2008)] 'Employing Clustering Techniques for Automatic Information Extraction from HTML Documents'. F Ashraf , T Ozyer , R Alhajj . *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 14. Manisha Marathe, Dr. S.H.Patil, G.V.Garje,M.S.Bewoor (ed.) 2008. November 2009. 38 (5) p. . (International Journal of Recent Trends in Engineering)

[Larsen and Aone ()] 'Fast and effective text mining using linear-time document clustering'. B Larsen , C Aone . *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining) 1999.

[Yang and Zhang ()] 'HTML Page Analysis Based on Visual Cues'. Y Yang , H Zhang . *6th International Conference on Document Analysis and Recognition*, (Seattle, Washington, USA) 2001.

[Song et al. ()] 'Learning Block Importance Models for Web Pages'. Ruihua Song , Haifeng Liu , Ji-Rong Wen , Wei-Ying Ma . *Proceedings of the 13th international conference on World Wide Web*, (the 13th international conference on World Wide WebNew York, NY, USA) 2004. p. .

[Song et al. ()] 'Learning Important Models for Web Page Blocks based on Layout and Content Analysis'. Ruihua Song , Haifeng Liu , Ji-Rong Wen , Wei-Ying Ma . *ACM SIGKDD Explorations Newsletter* 2004. 2-57, 1973. 6 (2) p. .

[Yates and Neto ()] *Modern Information Retrieval*, R B Yates , B R Neto . 1999. New York: Addison-Wesley.

[of Network and Mobile Technologies ()] *of Network and Mobile Technologies*, 2010. 1.

[Wu ()] 'Study of Web Page Information topic extraction technology based on vision'. Kai Wu . *IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, 2010. 9 p. .

[Liu et al. ()] 'ViDE: A Vision-Based Approach for Deep Web Data Extraction'. Wei Liu , Xiaofeng Meng , Weiyi Meng . *IEEE Transactions on Knowledge and Data Engineering* 2010. 22 (3) p. .

[Li et al. ()] 'Visual Segmentation-Based Data Record Extraction from Web Documents'. Longzhuang Li , Yonghuai Liu , Abel Obregon . *IEEE International Conference on Information Reuse and Integration*, 2007. p. .

[Yi and Liu ()] 'Web page cleaning for web mining through feature weighting'. Lan Yi , Bing Liu . *Proceedings of the 18th international joint conference on Artificial intelligence*, (the 18th international joint conference on Artificial intelligenceAcapulco, Mexico) August 09 -15. 2003. p. .