ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 2, February 2014

SEGMENTING WEB PAGES USING CORRELATION CLUSTERING AND REDUCING NOISY DATA USING SIMPLE K-MEAN ALGORITHM

Rajdeepa B[1], Premavathi M [2]

Assistant Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore[1]

M.Phil Research Scholar, Department of Computer Science, PSG College of Arts & Science, Coimbatore[2]

Abstract: The World Wide Web is most important facet in every part of the world. All areas in IT and other industries run using this WWW which contains large amount of data. The World Wide Web includes more and more websites each containing large data and it is highly demanded. Web pages are the useful aspect for retrieving required data from internet but problem in web page data retrieval is, it sometimes contains irrelevant data. This article is intended to retrieve the relevant data by segmenting web pages and removing noise in segmented web pages via K-means Algorithm in clustering.

Keywords: Vision-Based Web page Content Structure Analysis, Correlation Clustering, Clustering, K-Mean Algorithm.

I . INTRODUCTION

For the past and upcoming years' internet is the growing trend in sharing vast amount of data. Web pages are the important area to share the information over network. But the main issue in web page data retrieval is, the information on web is mixed with irrelevant data. A web page typically contains various contents of data such as navigation, decoration, interaction and contact information, which are not related to the topic of the web-page. Web page content is analyzed using vision-based content analysis; segmented using correlation clustering and extraneous data in web page can be removed by using K-Mean algorithm of clustering by segmenting web pages. The related work focuses on segmenting the web page into small pieces and removing noisy data from the web page using the effective clustering algorithm such as K-Mean algorithm.

II. WEBPAGE CONTENT ANALYSIS

Web page contains many data which are not relevant and furthermore it contains multiple pages that are not appropriate to the pages. Such web content structure is analyzed and detected using Vision-Based Content Structure analysis approach as follows:

This article suggests the vision-based content structure analysis, in which the every node in vision-based content structure analysis is called a *block* which is a set of basic objects. It is important to note that, the nodes in the vision-based content structure do not necessarily correspond to the nodes in the DOM tree approach. The basic model of vision-based content structure analysis for web pages is described as follows.

A web page Ω is represented as a triple $\Omega = (O, \Phi, \delta)$. $O = \{\Omega^1, \Omega^2, ..., \Omega^N\}$ is a finite set of blocks. All these blocks are not overlapped. Each block can be recursively viewed as a sub-web-page associated with sub-structure induced from the whole page structure. $\Phi = \{\varphi^1, \varphi^2, ..., \varphi^T\}$ is a finite set of separators, including horizontal separators and vertical separators. Every separator has a weight indicating its visibility, and all the separators in the same Φ have the same weight. δ is the relationship of every two blocks in O and can be expressed as: $\delta = O \times O \to \Phi \cup \{NULL\}$.

For example, suppose Ω_i and Ω_j are two objects in O, $\delta(\Omega_i, \Omega_j) \neq NULL$ indicates that Ω_i and Ω_j are exactly separated by the separator $\delta(\Omega_i, \Omega_j)$ or we can say the two objects are adjacent to each other, otherwise there are other objects between the two blocks Ω_i and Ω_j.

Fig 1: Vision-Based Content Structure Analysis Example

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 2, February 2014

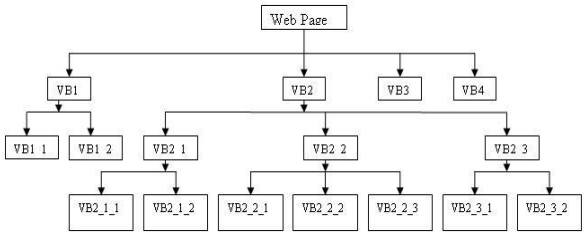

Fig 2: Web Page Segmentation Process Sample

$$O = (VB1, VB2, VB3, VB4)$$

$$\Phi = \{\varphi^1, \varphi^2, \varphi^3\}$$

$$\delta \begin{pmatrix} (VB1, VB2) \\ (VB2, VB3) \\ (VB3, VB4) \\ else \end{pmatrix} = \begin{pmatrix} \varphi^1 \\ \varphi^2 \\ \varphi^3 \\ NULL \end{pmatrix}$$

Fig (a)

$$VB2 = (VB2\_1, VB2\_2, VB2\_3)$$

$$\Phi^2 = \{\varphi_2^1, \varphi_2^2\}$$

$$\delta^2 \begin{pmatrix} (VB2\_1, VB2\_2) \\ (VB2\_2, VB2\_3) \\ else \end{pmatrix} = \begin{pmatrix} \varphi_2^1 \\ \varphi_2^2 \\ NULL \end{pmatrix}$$

Fig (b)

Fig (a) & (b): Specification of web Content Structure Analysis

Since each Ω_i is a sub-web-page of the original page, it has similar content structure as Ω. Recursively, we have $\Omega_s^t = (O_s^t, \Phi_s^t, \delta_s^t)$, $O_s^t = \{\Omega_{st}^1, \Omega_{st}^2, ..., \Omega_{st}^{N_{st}}\}$, $\Phi_s^t = \{\varphi_{st}^1, \varphi_{st}^2, ..., \varphi_{st}^{T_{st}}\}$ and $\delta_s^t = O_s^t \times O_s^t \to \Phi_s^t \cup \{NULL\}$ where Ω_s^t is the t^{th} object in the sub-web-page level s, N_{st} and T_{st} are the number of objects in O_s^t and number of separators in Φ_s^t.

The above pattern shows an example of vision-based content structure for a web page of Yahoo shopping Auctions. It shows the layout structure and the vision-based content structure analysis of the page. In the first point, the original web page has four visual blocks (objects) VB1~VB4 and three separators. Then sub content structure for each sub web page is constructed. For example, VB2 has three offspring blocks and two separators. It can be further analyzed as shown in Fig (b).

For each object, the *Degree of Coherence* (DoC) is defined to measure rationality of the web page content. DoC has the following properties:

- The greater the DoC value, the more consistent the content within the block;
- In the hierarchy tree, the DoC of the child is not smaller than that of its parent.

In this algorithm, DoC values are integers ranging from 1 to 10, although alternatively different ranges (e.g., real numbers, etc.) could be used.

The *Permitted Degree of Coherence* (PDoC) can be pre-defined to achieve different granularities of content structure for different applications. Different application can use VIPS to segment web page to a different granularity with proper PDoC.

The vision-based content structure is more likely to provide a detailed partitioning of the page. Every node of the structure is likely to convey certain explanations. For example, in Yahoo shopping web page the section VB2_1_1 denotes the category links of that particular web page, and that VB2_2_1 and VB2_2_2 shows different products for shopping in detail.

III. CORRELATION CLUSTERING FOR WEB PAGE SEGMENTATION

The correlation clustering problem starts with a complete weighted graph. The weight Vpq ∑ [0, 1] of an edge represents the cost of placing its endpoints p and q in two different segments; similarly, (1 − Vpq) represents the cost of placing p and q in the same segment. Since every edge contributes, whether it is within one segment or across segments, the segmentation cost function is automatically regularized: trivial segmentations such as one segment per node, or all nodes in one segment, typically have high costs, and the best segmentation is somewhere in the middle.
In fact, the number of segments is picked automatically by the algorithm. Note that the costs depend only on whether two nodes are in the same segment or not, and not on the labels of particular segments themselves. This imposes two constraints on using correlation clustering for segmentation. First, it precludes the use the invisible label ξ with its special properties. Hence, in order to satisfy Constraint 1, we must restrict the set of nodes to the set of leaf nodes, leaf (N), of the DOM tree. Second, pair wise costs between two nodes p and q must depend only on whether they belong to the same segment or not:

$$V_{pq}(\mathcal{S}(p), \mathcal{S}(q)) = \begin{cases} v_{pq} & \text{if } \mathcal{S}(p) \neq \mathcal{S}(q), \\ 1 - v_{pq} & \text{if } \mathcal{S}(p) = \mathcal{S}(q). \end{cases}$$

Thus, the objective function becomes,

$$\text{cclus}(\mathcal{S}) = \sum_{\substack{p,q \in \text{leaf}(\mathcal{N}) \\ \mathcal{S}(p) \neq \mathcal{S}(q)}} v_{pq} + \sum_{\substack{p,q \in \text{leaf}(\mathcal{N}) \\ \mathcal{S}(p) = \mathcal{S}(q)}} (1 - v_{pq}).$$

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 2, February 2014

The algorithm of Ailon, Charikar, and Newman is used for correlation clustering to find a segmentation whose cost is within a factor of two of the optimal. The algorithm CClus which is used in correlation clustering is iterative. At each stage, a node p in the current graph is chosen uniformly at random and removed from the graph. A new cluster is created with just p in it. Next, all the nodes q such that Vpq ≥ 1/2 are removed from the graph and placed in the cluster along with p. The process is repeated on the remaining graph. Since the algorithm is randomized, several independent trials are performed till the solution with the least objective value.

IV. CLUSTERING ALGORITHM

The clustering is the process of grouping up of data of similar types. Cluster analysis is a collection of patterns into cluster based on their similarity. The patterns in cluster analysis are usually represented as a vector measurement or as a multidimensional space. Patterns within the same cluster are more similar to the patterns related to other cluster.

The cluster holds the objects or data which are similar the dissimilar objects are compared to other objects in the cluster to make a group. The clustering algorithm is of different types which are as follows:

(i) Hierarchical Clustering (Connectivity-Based Clustering)
(ii) K-Means Clustering (Centroid-Based clustering)
(iii) Distribution-based clustering
(iv) Density-Based Clustering.

Among above all clustering algorithms the thesis focus on Centroid-based algorithm for clustering and removing noisy data in web page.

K-MEANS ALGORITHM FOR REMOVING NOISY DATA

The K-Mean algorithm is effective algorithm for grouping\ the data/objects that are similar. In this algorithm thegrouping is done by the following steps:

- K-mean takes input as set of S of objects and an integer K. the output as partition of S into subsets as S1,…..,S2, Sk.
- Data in the segmented page are classified as belonging to one of K groups.
- Cluster membership is determined by calculating the centroids for each cluster.
- Each data is assigned to cluster of the closest centroid.
- If dispersion within cluster occurs iterative reallocation of clusters are done.

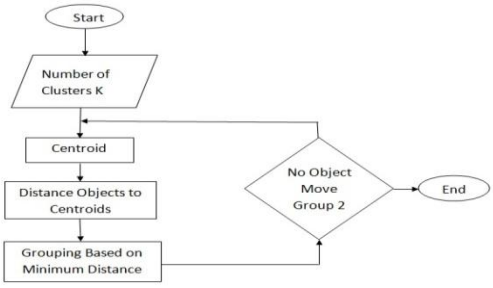

Fig 8: Process Step of K-Mean

The original data is divided into K groups or K clusters as shown in the figure 9 and figure 10. Such grouping in K-Mean is done by calculating centroid value for each cluster using the formula:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

The main objective of this is to minimize the total cost as:

$c(S_i)+\ldots\ldots\ldots+c(S_k)$

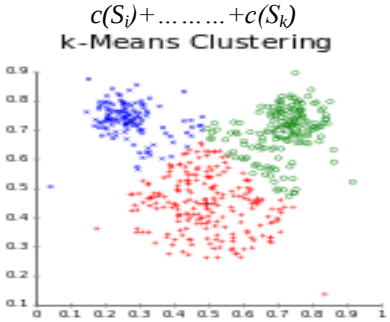

Fig 9: Clustered Data

The time complexity for calculating centroid is *O(nkl)* where *n* is number of patterns, *k* is number of clusters and *l* is number of iterations the algorithm takes. Using K-Mean algorithm relevant data is clustered and irrelevant (noisy) data is removed.

V. HIERARCHICAL CLUSTERING FOR MERGING SEGMENTED WEB PAGE CONTENT

The hierarchical clustering algorithm is used to divide the web page into different pieces or to merge the segmented web page content. In this paper, this algorithm is used to merge the data content that are segmented using Vision-Based page segmentation algorithm. The data that are segmented as S_1……, S_2, S_k are merged into single page. This is the cheapest merging algorithm used in merging web page content.

Let's take S_i and S_j. Once the contents are merged the S_i and S_j are removed from the list of sets and replaced with (S_i U S_j). This process of merging is continued until all data comes under single group.

PROCESSING STEPS OF RELATED WORK

The processing steps of this study are explained pictorially as follows:

- The YAHOO shopping web page is used for segmenting and removing noisy data.
- First the web page is analyzed for content structure using Vision-Based content structure analysis App.
- Then the analyzed web page is segmented into smaller units using VIPS algorithm.
- Segmented web pages are clustered and noisy data in the units are removed using K-Means algorithm in clustering.
- After removing noisy data the segments are again merged into single unit using hierarchical clustering method.
- The silent content of web page is stored in web page database.

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 2, February 2014

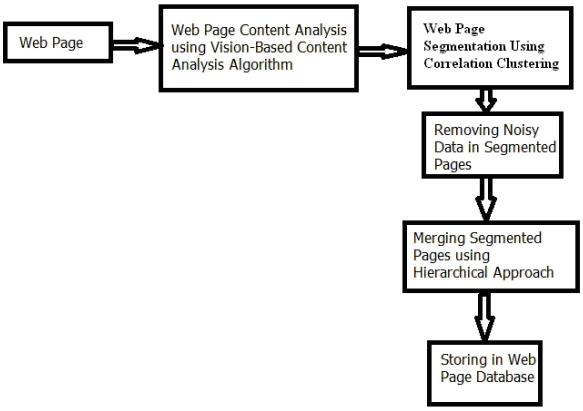

Fig 10: Process Step of related work

VI. CONCLUSION

Web mining is the recent trend in all areas where the required data is extracted from relevant database over network. Such data must be relevant to the requirement and the data should be related to the web page. The main problem in web mining is extracting the relevant and noiseless data.

Thus the study shows the about the algorithm that can be used for segmenting the web page and how to remove the noisy data from the web page using clustering K-Mean algorithm.

BIOGRAPHIES

1. Osama Abu Abbas, "Comparision Beween Cluatering Algorithms", The International arab Journal of Information Tachnology, Vol. 5, No. 3, July 2008.
2. Deepayan Chakrabarti, Ravi Kumar, Kunal Punera,"A Graphic-Theoritic Approach to Web Page Segmentation", Corpus Characterization & Search Performance.
3. Cai, D., Yu, S., Wen, J.R., Ma, " VPIS: A Vision-Based Pae Segmentation Algorithm", Microsoft Technical Report. MSR-TR-2003-79, 2003.
4. Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, , Ruth Silverman, and Angela Y. Wu," An Efficient k-Means Clustering Algorithm: Analysis and Implementation" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July 2002.
5. Marjan Kuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo, "A survey of hierarchical clustering algorithms", The Journal of Mathematics and Computer Science, Vol .5 No.3 (2012) 229-240.
6. Hui Xiong, Member, Gaurav Pandey, Michael Steinbach, Member and Vipin Kumar, Fellow," Enhancing Data Analysis with Noise Removal", IEEE Transaction on Knowledge and Data Engineering.
7. Xindong Wu, Qiang Yang, "10 Challenging problems in Data Mining", International Journal of Information Technology & Decision Making, Vol. 5, No. 4 (2006) 597–604.
8. Ricardo Baeza-Yates, "Information retrieval in the Web:beyond current search engines", International Journal of Approximate Reasoning 34 (2003) 97–104.
9. R. Cooley, B. Mobasher, J. Srivastava, and Web mining: information and pattern discovery on the World Wide Web, ICTAI, 1997, pp. 558–567.
10. Kevin Chen-Chuan Chang, Bing Liu, "Special Issue on Web Content Mining", WWW' 03, 2003.