A Novel Technique for Web Pages Clustering Using LSA and K-Medoids Algorithm

Nora Omran Alkaam, Noor A. Neamah and Faris Sahib Al-Rammahi

Abstract The extensibility of various web documents available on the web made a critical challenge for many serious tasks such as information retrieval (IR), content monitoring, and indexing. Web documents could be any type of data that can be requested by user and delivered from web server through several web browsers. Most of web documents contain textual contents and are typically called web pages. However, in order to perceive and discover knowledge from these pages, novel techniques are required that have been never applied in other domains. In this paper, a new approach has been proposed by performed latent semantic analysis (LSA) on the result of VSM, which involves the correlation among web pages to their extracted features. The result of LSA involves the matrices that reflect the correlation between the web pages to their related concepts, which were used frequently for retrieving process. PAM (K-Medoids) algorithm was used with respect to semantic space, to portion the web pages into coherent groups. One of the most challenges in any clustering algorithm is to identify the correct number of clusters for the given data. Hence, two approaches are used for this manner: Elbow graph analysis to estimate the number of cluster range based on (SSE) values and clustering evaluation metrics. Calinski–Harabasz criterion (CH) and Silhouette Coefficient (SC) are the best well-known evaluation metrics commonly used in partitioning-based algorithms. UOT has been considered to evaluate the proposed system, and the results are shown in the proposed system to achieve high accuracy results to separate the similar pages into coherent groups.

Keywords Clustering · Web mining · Data mining · Web content mining · PAM · K-Medoids · Silhouette coefficient · Calinski–Harabasz criterion

N. O. Alkaam (✉)
Department of Higher Studies, Ministry of Higher Education and Scientific Research, Baghdad, Iraq
e-mail: noor.nemah12@gmail.com

N. A. Neamah · F. S. Al-Rammahi
Department of Computer Techniques Engineering, Imam Al-Kadhem College, Baghdad, Iraq

V. K. Solanki et al. (eds.), *Intelligent Computing in Engineering*,
Advances in Intelligent Systems and Computing 1125,
https://doi.org/10.1007/978-981-15-2780-7_79

1 Introduction

With the massive growth of web documents on the World Wide Web, and enormous increasing of these documents on the web introduces a big challenge to understand its contents. However, with million if not billion interconnected web documents created by millions of Authors around the world, the task of understanding it requires hundreds of years. These documents either typically includes descriptions property in their contents (e.g., titles, tags, keywords, meta-descriptions, etc.,) or not, the processing of web documents should be performed. In addition, the web documents do not exist in unique form, and frequently exist in three common forms which are structured, semi-structured, and unstructured forms. Web documents could be image, text files, videos, xml files, etc., and these contents are varied based on the type of its content. The process of discovering knowledge and expressing useful content from large raw data is called data mining. In web scenario, the process of extracting knowledge from large web documents is called web mining. Hence, the main goal of web mining is to providing techniques that make the web data more convenient and efficient for applying advanced techniques (e.g., classification, clustering, association rule, indexing, etc.) [1]. In recent years, there has been an urgent need to apply web technologies to make search engines work best, and help variant users around the world to get their interested contents [2]. Web mining is categorized into three classes based on type of data that should processed which are Web Content Mining, Web Structured Mining, and Web Usage Mining as shown in the following block diagram in the figure (Fig. 1) [3].

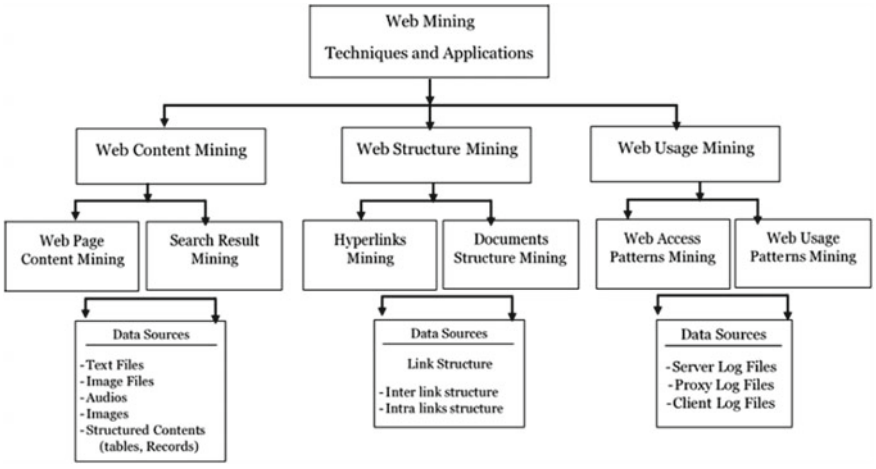

Fig. 1 Web mining categorization [2]

2 Proposed System

The main primary part in proposed system is latent semantic analysis of the text contents in web pages. VSM typically involves thousands of independent words, where every word is treated as a single feature. However, the dimensionality becomes very sparse with the present of all these words; VSM contain many (0's) values, for most of the existing words that may belong to a small subset of web pages. However, the mean values of these columns become much smaller than nonzero columns, and this produces other challenges when clustering algorithm has been used with traditional similarity/dissimilarity measures. LSA helps to avoid some critical challenges in the text, like lexical semantic issues in some words such as synonymy, polysemy, and homonymy.

2.1 Latent Semantic Analysis of the Text

LSA is a technique in Natural Language Processing (NLP), which uses a mathematical method called (SVD). SVD, actually, is a matrix factorization method for a given VSM, which aim to project the original dimensional space into lower dimension space commonly called (k-space). LSA assumes that words (features) that are close to each other in their meaning will appear in similar pieces of text. However, by reducing the dimension into semantic space with preserving the similarity structure among web pages, the words will appear into their related semantic concepts.

By assumption $WM \times N$ is a word-web page matrix where rows (M) are web pages, and (N) columns are the result Words in these web pages, the entries of the contents are the words weights measured by using (TF-IDF), then the decomposition of the matrix W is given as a three matrices as in the following equation:

$$LSA(W) \equiv SVD(W) = U \Sigma V^T \tag{1}$$

where

- $\boldsymbol{U}$ is ($\boldsymbol{M} \times \boldsymbol{r}$) matrix which represents the correlation between (M) web pages to the new projected space, which is called (semantic/concepts) space. columns of $\boldsymbol{U}$ matrix that represent ***left singular vectors and*** indicate to the degree of pertinence of each word to its extracted concepts.
- $\sum$ is ($\boldsymbol{r} \times \boldsymbol{r}$) is a diagonal matrix and is called ***singular values matrix.*** $\boldsymbol{U}$ matrix actually contains non-negative square roots of $\boldsymbol{r}$, and the values arranged are in decreasing order. However, by selecting top (σ) values, the U and V^T matrix transformed into ($\boldsymbol{K}$) space where the value of ($\boldsymbol{K}$) reflects the importance of concepts to their related words, and web pages to their concepts.
- $\boldsymbol{V^T}$ is ($\boldsymbol{N} \times \boldsymbol{r}$) matrix and typically shows the semantic correlation of words to their extracted concepts. This matrix includes (negative and positive) values; the negative values refer to the words includes either poor pertinence degree or not

related to that concept, $\boldsymbol{V^T}$ columns named ***right singular vectors***. The work shows the decomposition process of VSM matrix.

2.2 Elbow Graph Cutting (EGC) Method

Elbow graph analysis is a method typically used in semantic analysis like LSA, PLSA, LDA, etc. The main goal of EGC is to finding the optimal truncated value that is used in semantic space after reduction step. The value (K) in LSA represents the new truncated value for the matrices of $\{U \sum V^T\}$, to produce new transformed matrices $K'k = \{Uk \sum k VkT\}$. Elbow graph cutting value is detected when the singular values curve changes with insignificant value as shown in the work, where the expected number of clusters can be detected significantly as shown below.

2.3 Partitioning-Based Algorithm

Partitioning-based algorithms involve the process of assigning the points to its nearest centroid, which is also called (centroid-based clustering). However, these algorithms initially choose K points as representative points to initialize cluster centroids, and then every data point is assigned to the closest one. Hence, for a given dataset (D) with (M) data points that is presented in (N) dimensional space $\boldsymbol{D} = \{p_i\}_{i=1}^{m}$, the main goal of partitioning algorithms is to find k clusters $C = \{C_1, C_1, \ldots, C_k\}$ by partitioning the dataset $\boldsymbol{D}$ into K clusters. In initialization step, the cluster centroids are chosen randomly, and then the process is repeated for every not assigned data points and updated cluster centroids. The process will continue until the function converges or requirements are met. However, the result clusters are present with spherical shape because the data points are assigned to its closest centroid. The centroid (mean) is defined as in the following equation:

$$\mu_i = \frac{1}{n_i} \sum_{p_j \in C_i} p_j \tag{2}$$

where $n_i = |C_i|$ and refers to whole number of data points in cluster C_i. The main goal of portioning based algorithm is to find clusters with minimal sum of squared error (SSE). K-means, CLARA, PAM, and variation of K-means algorithm are some example of this type.

2.4 Partitioning Around Medoids (PAM) Algorithm

PAM is commonly known as *K*-Medoids algorithm, and it is one of the most popular partitioning-based algorithm, which is related to *K*-means algorithm. PAM attempts to minimize the distance among the points within one cluster, and designates a point to be the center (centroid) of cluster. PAM typically chooses *K*-medoids as cluster centers, and mostly works with Manhattan Norm (I) to calculate the distance among data points. Hence, to calculate the distance between two web pages (p, q) by using Manhattan Distance is as follow:

$$Dis([p_1, p_2, \ldots, p_m], [q_1, q_2, \ldots, q_m]) = \sum_{j=1}^{m} (p_j - q_j) \tag{3}$$

One of the most difficult questions in partitioning-based clustering algorithm is how many clusters can be found (correct number of clusters). However, there are some useful tools that could be used to know the correct number of clusters by considering the optimal evaluation values that correspond to the cluster number. One of most common tool used with *K*-means and its variation is Silhouette Coefficient measure. The following algorithm describes the works of *K*-Medoid clustering algorithm which is partitioned by the data points around the Medoids, and use greedy search technique to faster the searching process.

Algorithm 1: Clustering Forming based k-Medoids

K-Means Input (W, k)
Input: (***W***)*corpus of "Web pages"*
(k)*Is the number of clusters* specified by user {*K=1...k*}
Output:($\boldsymbol{\mu_k}$)*Set of Clusters Medoids*{ $\boldsymbol{\mu_c = \mu_1 \ldots \mu_k}$}
Begin

1. **Initialization**: Randomly select(K) points from (**W**) as Medoids.
2. Repeat
3. Assign each remaining points in (**W**) to its closet Medoid.
4. For each Medoids points ($\boldsymbol{\mu_k}$)
5. For Each non-Medoids points ($\boldsymbol{Nm}$)
6. Computer the cost of (**S**) of swapping$\boldsymbol{\mu_k}$with $\boldsymbol{Nm}$
7. If (S<0) then Swap $\boldsymbol{Nm}$with $\boldsymbol{\mu_k}$toform new set of Medoids**.**
8. **Until** (Convergence or **No change**).

End

2.5 Clustering Results Evaluation

To evaluate the clustering results, either there are some significant measures used with whole algorithm or certain type of algorithms. However, there are two measures used in this paper (HC) and (SC):

- **Calinski–Harabasz Criterion (CH)**

This measure is best suited for partitioning clustering algorithm, where user defines the number of clusters. CH measure is some time called variance ratio criterion (VRC) and is used to measure the cluster validity depending on the average between and within cluster sum of square, and can be calculated as in the following equation:

$$CH_k = \frac{SS_B}{SS_w} \cdot \frac{(N-K)}{(K-1)} \tag{4}$$

where

$\boldsymbol{SS_B}$ is sum of squares between clusters.
$\boldsymbol{SS_W}$ is sum of squares within clusters.
$\boldsymbol{K}$ is the number of clusters, $\boldsymbol{N}$ is the number of instances.

$$SS_w = \sum_{i=1}^{k} \sum_{P \in c_i} \left\| P - \mu_i^2 \right\| \tag{5}$$

$$SS_B = \sum_{i=1}^{k} |C_i| \left\| \mu_i - \mu^2 \right\| \tag{6}$$

- **Silhouette Coefficient (SC)**

SC can be defined as a measure for both cohesion and separation of clusters. SC measures the variation between the average distances of data points in their closest cluster to others. SC values ranged from (−1, to 1), where large positive values refer to high separating clustering and point is in optimal cluster, while lower values refer bad clustering results or separation. SC can be calculated based on the following equation:

$$S_i = \frac{M_{\text{Out}}^m(P_i) - M_{\text{in}}(P_i)}{\text{Max}\left\{M_{\text{Out}}^m(P_i), M_{\text{in}}(P_i)\right\}} \tag{7}$$

3 Experimental Results

The implementation of the proposed system, starting from preprocessing of web pages in real SGML documents, actually these documents contain massive instruction lines as shown in the following figure (Fig. 1), where every web page involves around (300–1000) lines, and the useful text content arise only in 5–10 lines.

Removing a lot of SGML content could face several challenges. However, the best way is to identify the significant SGML tags. In this paper, only three tags are considered (Title, Paragraphs in Body <p> and Meta-tag description).

After preprocessing of web pages, and follow all the web pages preprocessing steps, the result of these web pages is a sparse matrix called (VSM) matrix.

Next step in the proposed system including latent semantic analysis of VSM matrix, where the result of applying for the given above matrix is three matrices (U, Σ and V^T). U matrix actually contains the web pages correlation to their concepts and in this paper, U matrix is used to clustering of web pages. The work shows the U matrix after applying LSA analysis.

The result matrix of the work is used with PAM (K-Medoids) algorithm with cluster number ranges by considering the elbow graph result work and applying the clustering process. The following table (Table 1) shows the clustering result along with evaluation metrics as shown below:

Table 1 Clustering results and evaluation of (UOT) web pages using PAM algorithm

Cluster number	SC values	Distances means	SSB	SSW	CH values
3	0.3618	7.5482	1.9521	22.6446	259.6299
4	0.1794	5.4169	2.9289	21.6677	248.4068
5	0.5214	4.1815	3.6890	20.9077	262.8475
6	0.2336	3.3260	4.6406	19.9561	290.0365
7	0.2844	2.7239	5.5290	19.0676	304.8951
8	0.5925	2.2758	6.3903	18.2063	314.2819
9	0.3730	1.9435	7.1053	17.4913	314.5495
10	0.4515	1.6917	7.6799	16.9168	327.9745
11	0.4798	1.4066	9.1246	15.4721	352.4478
12	0.4848	1.2286	9.8538	14.7428	362.7412
13	0.3657	1.0875	10.4592	14.1375	375.9957
14	0.3859	0.9594	11.1646	13.4321	398.8142
15	0.5014	0.8456	11.9130	12.6836	424.1663
16	0.5539	0.7196	13.0830	11.5137	473.0967
17	0.5622	0.6636	13.3155	11.2811	520.3291
18	0.4171	0.5316	15.0279	9.5688	473.0729
19	0.5110	0.4538	15.9736	8.6231	470.8083
20	0.5224	0.4589	15.4177	9.1789	502.1349

The above table (Table 1) results show that the optimal number of clusters for the given web pages in UOT dataset by using LSA with PAM algorithm is (17) and it is colored with yellow to distinguish it from other clustering and evaluation results. The work shows the (SC) evaluation values across the given cluster range in PAM algorithm.

The work shows the scatter plot of web pages in UOT dataset, where the (U) matrix of LSA is used to plotting the data points in space by discarding the first columns' values in (U) matrix and considering the second and third columns as (X) and (Y) coordinates, respectively.

4 Conclusion and Future Works

Clustering web documents have received a lot of attention in recent years because of the accumulation of pages in the web servers without knowing or archiving any details in these pages. However, clustering of these documents provides many facilities for many web services and applications. In this paper, we introduce a novel methodology by using PAM algorithm, which considers one of the most popular partitioning-based clustering algorithms. UOT dataset is considered as main data to implement the proposed system, where the dataset has been collected using Teleport Software. After that, preprocessing of web pages steps has started based on proposed system sequence. The result Matrix is represented as VSM, and the weighting score of every web page is calculated by (TF-IDF). After that, LSA is applied to analysis the result matrix of VSM, which produces three matrices ($U \Sigma V^T$). U matrix which represented (Web Pages—Concept) matrix is used to clustering these pages (partitioned it into coherence groups). However, every cluster (group) involves the most related web pages based on its content. Elbow graph analysis is used to expect the number of clusters based on square of singular values, which reflect the correlation between (web pages to Concepts), and (words to concepts). Hence, LSA guarantees that similar webpages in contents fall in similar cluster. PAM algorithm is used with the given estimated cluster range. SC and CH metrics are used to evaluate the clustering results of PAM and the values show that the optimal number of clusters is (17) as shown in experimental result section. The proposed methodology shows that the proposed system achieves high accuracy rate in order to clustering real-time web pages, and after texted some clusters, we found most of the similar web pages are concluded in similar clusters. Our future works, in this manner involve improving the clustering of web pages by comparing the result of other algorithms, and improving the preprocessing of web pages to eliminate many noise features that may appear in VSM. In addition, we proposed to use labeling algorithm to find the common concepts among web pages in every cluster.

References

1. Langhnoja SG, Barot MP, Mehta DB (2013) Web usage mining ton discover visitor group with common behavior using DBSCAN clustering algorithm. Int J Eng Innov Technol 2(7):169–173
2. Sandhya, Mala chaturvedi (2013) A survey on web mining algorithms. Int J Eng Sci 2(3):25–30
3. Saini S, Pandey HM (2015) Review on web content mining techniques. Int J Comput Appl 118(18):33–36