Journal of Engineering and Applied Sciences 12 (Special Issue 8): 8270-8277, 2017
ISSN: 1816-949X

Proposed Method for Web Pages Clustering using Latent Semantic Analysis

[1]Tawfiq A. Al-Asadi, [2]Ahmed J. Obaid, [2]Ahmad A. Al-Khayatt
[1]College of Information Technology, Babylon University, Babel, Iraq
[2]Department of Computer Science, College of Education, Kufa University, Najaf, Iraq

Abstract: With explosive growth of information in World Wide Web and large host numbers continuously added to the internet which has different types of content such as text, images, audio, videos and others made the process of analysis web data very complex and difficult process. Analysis of web data also is very important task which required by many organizations, academic centers, companie's, agencies, etc., for various task such as enhanced searching process, monitoring and business application. Therefore, many algorithms has proposed in recent research works to construct a model used for knowledge discovering from a corpus of web pages which might exist in different structured forms. The process of extracting web pages features and grouping similar web pages having similar interesting features called web page clustering. In this study, latent semantic analysis has been considered which has applied successfully in many text documents applications by first applying web pages mining process to our selected web pages corpus then text mining process are applied to represent web pages in a Vector Space Model (VSM), finally, K-means algorithm proposed to grouping similar web pages which are exist in semantic space.

Key words: Web Content Mining (WCM), web page clustering, Latent Semantic Analysis (LSA), K-means, mining, successfully

INTRODUCTION

Web pages clustering process can be considered one of the most complex and important task in web mining which require to types of data to be studied and analyzed successfully to enhanced services provided on the web. Web users made large number of requests to various web pages which either belong to single web site or different web sites hosted in many web servers on the world. Structure and appearance of web sites contents are differ depend on the design technique used where many web sites designed by using CMS scripts or other languages such as Asp, PHP, HTML, etc. The application of data mining in web data is called web mining. Web data Mining classified into three categories are: Web Usage Mining (WUM), Web Content Mining (WCM) and Web Structure Mining (WSM) based on data type used in mining process. Web usage mining is the application of data mining to understanding user's behaviors and their interesting towards various web sites contents. Web content mining is the task of applying data mining techniques to extracting useful information from the contents of web documents while web structure mining is the process of extracting knowledge about hyperlinks among web pages and the travers paths of users among large number of web pages which may belong to same web site or others sites (Cooley *et al*., 1997; Al-Asdi and Obaid, 2016).

A starting point to perform web pages clustering process is to create a model for representing web pages by its features, one of the most usage model for representing text documents as well as web pages is VSM where web pages are parsed to extract unique words from set of web pages which has considered as unique features that has a discriminative power to recognize pages among each other's. The most usage weighting schema which is commonly used to weighting features and its important in web pages is by use term frequency inverse document frequency approach (Anastasiu *et al*., 2013). When all words that extracted from web corpus has been considered, then the VSM become very sparse and most values in VSM are 0's, this phenomena called high-dimensional sparsity (Curse of Dimensionality) and this issue made data mining techniques such as clustering a very difficult process (Charu, 2015a, b). Therefore, most of mining results doesn't have a meaning with present lot of noise features in high-dimensional space. Feature selection/transformation are a pre-processing methods used to reduce high-dimensional space into lower dimension which lead to improve the performance of data mining techniques such as clustering (Shah and Mahajan, 2012). However, reduction of dimension space not only eliminate the noise (irrelevant features) and redundant attributes but also improve the computation space and storage requirement (Alelyani *et al*., 2014; Aggarwal and Zhai, 2012).

Corresponding Author: Tawfiq A. Al-asadi1, College of Information Technology, Babylon University, Babel, Iraq

The most commonly method used in dimension reduction where new features are linear combination to the features in original space is Latent Semantic Analysis (LSA). LSA is widely used in document clustering, classification, indexing and information retrieval and is a quite appropriate for a sparse data as in text documents. LSA which is also called latent semantic indexing (when used in information retrieval) employed by search engines to retrieve the relevant result according to user query (Charu, 2015a, b). In this study, web sites of KUFA University has been considered to be processed, the study proposed start crawling the web pages corpus from university web server to local repository then applying pre-processing and analyzing steps where the textual features has been extracted from web pages after pre-processed it then using LSA to extract and represent corpus in semantic space. Finally, point assignment clustering algorithm (K-means) proposed to partition similar group of web pages depend on its correlation to semantic concepts.

Problem statement: Web content mining techniques has a number of research issues regarding to extract and discovering knowledge from web pages contents that comes in different forms, web clustering currently considered most critical problems in web application and it's highly important for categories, organize, discovering and extracting knowledge from dynamic web resource contents. There are practical challenges in clustering process for web sites clustering a new algorithm needed to combine different type of data forms and preprocessing varies web resources (Yadav and Mittal, 2013; Vakali *et al*., 2007). There is another kind of challenges related to representation of extracted features from large collection of web pages. Extracted textual feature from web pages faces several additional problems related to high dimensional feature representation and variant length of web pages. Unlike traditional data mining techniques, clustering of web pages required to use sophisticated algorithm doesn't use in any other application and can deal with sparse features which may fail to group interesting content when use another type of algorithms (Anastasiu *et al*., 2013). Therefore, there is always a persistent need to analyses web pages from time to time to discover similar web pages that belongs to similar or different web sites which can give a knowledge about what are web pages speaking about.

Literature review: The explosive growth of web pages in the world make the indexing process of pages by search engines is difficult in order to represent requested knowledge from users in a proper way, also to group similar web pages that share a meaningful data. Therefore, clustering process is an unsupervised technique used to group similar objects that are similar to each other within similar group and are varied from those in other corpuses or groups. Thus, many researchers are working from this points by Jain *et al*. (1999) provide an spacious survey of various techniques used for web document clustering process. In 2004, Hammouda and Kamel (2004) proposed a clustering algorithm based on Document Index Graph (DIG). Andrews and Fox (2007) illustrate a development approaches used for web document clustering while (Hung and Xiaotie, 2007) measure similarity among web documents by using Suffix-tree similarity measure then used Hierarchal Clustering Algorithm (HCA) to group similar web pages. Srinivas and Mohan (2010) proposed a clustering algorithm by using incremental hierarchical clustering approach. Jensi and Jiji (2014) illustrate an optimization approaches for clustering text documents including LSA concepts which used for indexing similar documents. Nirkhi and Hande (2008) discuss an overview of most used web document clustering algorithms and its application and limitation exist in web environment. Xiao (2010) presented an overview of semantic techniques used in web document clustering and its limitation when applying in real time web data. In this study, we employee LSA concept to extracting a knowledge from real time web pages that collected from Kufa University web sites which has a different sources and content and illustrate the result of applying semantic analysis on our selected data.

MATERIALS AND METHODS

Proposed models: Implementation of our proposed system consist from several steps including crawling web data till clustering analysis model, Fig. 1 shows our proposed model.

Crawling data: Web crawler system also called web spider software are the programs can download web pages, data on the web is sparse on million on pages serve by hug number of servers when users follow the hyperlinks to access information from one page to other. Crawler can visit many web sites to collect web data and used for mining and storage purpose. In our proposed model, we use crawling software for catching website pages and images and store it in repository. There are many crawling software that used for different purpose such as gathering information and links updating for web site content, monitoring and performance test and other tasks. One of the most usage crawling software Teleport Pro., it's possible to apply configuration of it for catching the web pages and images and avoiding other files such as CSS files, java files and so on. In images it's also

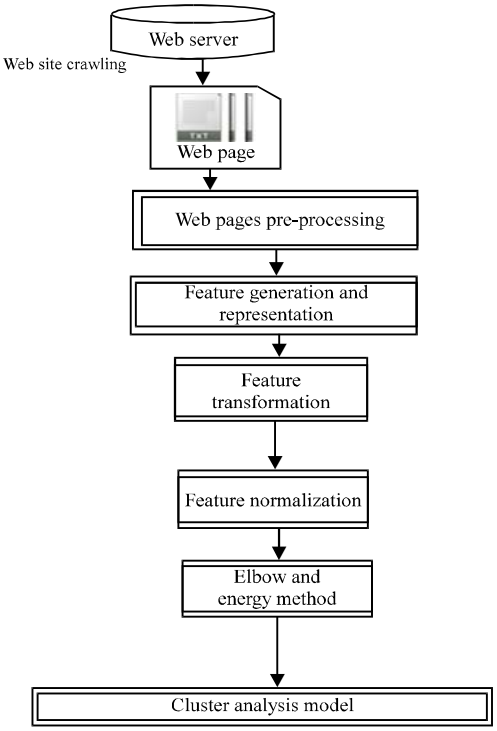

Fig. 1: Proposed model diagram

possible also catching images and filtering it by avoiding such as gif and small size images. The main feature in this software is able to catch any page or content for that web site and store it in static HTML page, some web sites design by use CMS (Content Management Systems), so, articles are store also as a static HTML pages for this reason, we use it for our web site.

Web pages preprocessing: Web page like any text document need to be preprocessed before applying mining algorithms such as extracting the terms and counting words in it, the difference in web document such as HTML page is by tag removal and identified the main content blocks in that page and this required carefully and nested procedures when applying preprocessing on HTML pages (Srivastava *et al*., 2005).

Preprocessing steps: Web pages preprocessing include some preprocessing steps as follow.

Identifying text fields: There are different text fields in HTML pages such as title, body and metadata. In many search engine search it consider the title of article or keywords when searching about related content in our model we assign weight for title let say 0.5 for the words that appear in title and remaining for words that in body.

Removing HTML tags: By removing HTML tags may cause problems by joining text that separated by tags, there are many software can be used for converting HTML pages to text files and then preprocessing text files for extracting text features from it. We introduce this problems in HTML pages and solve it by used new tools for extracting HTML text this tool is called "Html AgilityPack". By using this tools text in HTML tags can be extracted then store it in file that carried its own name.

Anchor text identification: Anchor text is the text that associate with hyperlinks, anchor text contain information that pointed to other pages. In many HTML pages we found many pages contain same anchor text that linkage pages among others for this reason these words were eliminated in text files.

Main content blocks text identification: Commercial web pages contain many information regarding to description, copy right and search notes to other related topics. This issue happen with a commercial and different content pages, when using nested pages design and static content, blocks of text content in these HTML pages will be simply identified.

Text preprocessing: To preprocess text files need to do again many processes due web pages can be varied and contain many contents some of these pages may be news articles, blogs, information about particular topics, lectures, notes and others while some pages might written in many languages. Clustering or classification of these pages by topics could be a challenges problem. There are many approaches for dealing with document indexing and information retrieval such as Vector Space Model (VSM) and Latent Semantic Indexing (LSA). A practical method for web document representation is by used VSM representation, in this study, we discussed briefly about how to preprocess text of web documents as shownin the following steps (Liu, 2011).

Stop word removal: Stop words are commonly occurred words in documents, these words provide little information about documents and it's actual content, typically stop words commonly available around 300-400 words and its available by Yadav and Mittal (2013).

Stemming: Variation of the same word need to be consider when preprocess text files, for example, the variation of the word "Computer" can be found as" Computers" for English languages and for Arabic can be found as "كو مبيو تر", "حاسب الي", "ا له حاسبة", "الحاسوب" and others all of these words refer to one general word

category " computer". In the model we consider status of two languages due the web site for KUFA University has been designed by using two languages Arabic and English. Articles has also been written by those languages. The main drawback in stemming is in some cases there are many words refer to different meaning, this challenge face us if we try to clustering or classification text document based on its content.

Punctuation marks: After stemming process has been done, punctuation marks such as comma, semicolons, numbers, dates, hyphens and special characters are eliminated. Typically hyphenated word can be treated as separated word or they merged into single word. Many common used words may added to dictionary that used for feature extraction process. Result of this step are text documents that contain distinct and meaningful words and will be treated as a bog-of-words which the ordering of the words irrelevant. Use bag-of-words representation is suddenly effective way for fast document labeling and classification.

Web (pages) documents representation: In this step each web document is represented by vectors where the rows represent web documents and columns are unique words. The entries of web document-word matrix are frequencies of unique word in every web document. A popular weight for word frequency is by use Tf×Idf schema. In this method, the weight of word (term)t_{iin} web document d_j is the number of times that word (term) t_i appears in web document d_j and it's denoted by f_{ij}. Idf is the inverse document frequency, this can reflect the importance of given words (terms) that appears in web document this schema sometimes used for text classification (Joachims, 1998). The equations for a given schema as follow:

$$tf_{ij} = \frac{f_{ij}}{Max\{f_1, f_2, f_3, ..., f_{|v|}\}} \quad (1)$$

Where:

tf_{ij} = The normalized-term frequency

$|v|$ = The vocabulary size of the collection

The inverse document frequency is given by:

$$idf_i = \frac{\log N}{df_i} \quad (2)$$

Where:

N = Total number of web documents

df_i = The number of web document that contain the term t_i

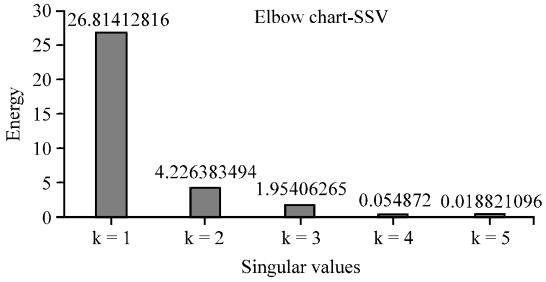

Fig. 2: SSV of word (term) web pages matrix

Then, the weight of term t_i is given by:

$$w_{ij} = tf_{ij} \times idf \quad (3)$$

If the term t_i does not appear in web document d_j, then, otherwise refer to number of occurrence that term t_i in web document d_j. In order to obtain better result for web document with different length then the following Eq. 4:

$$w_{ij} = tf_{ij} \times idf = \frac{f_{ij}}{\sum_{s=1}^{N}(tf_{ij}\, idf(w_{sj})^2} \times \log_2\left(\frac{N}{df_i}\right) \quad (4)$$

Finally, the result as shown in Fig. 2 given to the next step which is called "feature transformation" which is used for reduce the high dimensionality of extracted features.

Feature transformation: Feature transformation also called (feature reduction) aim to mapping the features in original space to new feature space has lower dimension and can't linked to the features in the original space. The goal for applying feature transformation to creating intelligent features to avoid the problems in original features such as polysemy, synonymy and homonymy which present in web documents. The most commonly method used in dimension reduction in which the new features are linear combination to the features in original space is Latent Semantic Analysis (LSA). LSA is widely used in document clustering, classification, indexing and information retrieval and is a quite appropriate for a sparse data as in text documents. LSA which is also called latent semantic indexing (used with information retrieval) by search engines to retrieve the relevant result according to user query.

Normalization and elbow curve plotting: The result of pre-processing steps and feature transformation using LSA method for input word (term) web pages matrix which is called (A) is three matrices as shown in the following Eq. 5:

$$A = U \sum V^T \quad (5)$$

Where:
U = The mapping of words (terms) to concept view
V^T = The mapping of web pages to extracted concepts

Therefore, for large collection of data set that contain hundred thousands of web pages (or web documents), recommended dimensions based on recent studies between 100-500 dimensions. Hence, with this number of dimension a new contribution work can be extracted is by applying clustering algorithm on extracted semantic space as in the proposed system where partitioning clustering algorithm proposed to cover the problem of finding similarities among large number of web pages. Elbow curve plotting is a method used in LSA for finding the optimal truncated value in reduction step where K in LSA implementation represent the new truncated space for the matrices of $\{U\sum V^T\}$ to reduction space of (K) where $A'_k = \{U_k \sum_k V_k^T\}$, elbow method start by plot the Squared of Singular Values (SSV) as shown in Fig. 3 and 4.

In Fig. 3, the first value seemsadded more significant for information retrieval or analysis content than other values, therefore, the first dimension give an absolute value for the web pages collection. However, this value correspond to the length of web pages in matrix V^T and correspond to the number of times words used in all web pages in matrix U.

Cluster analysis model: One of the most used algorithm especially in text mining is K-means algorithm which used Euclidian distance as default dissimilarity measure. In text mining where number of dimension become very high which is called curse of dimensionality, most feature transformation method used in text mining is LSA which

DJ	DK	DL	DM	DN	DO	DP	DQ	DR	DS	DT
0.043323481	0.022321262	0.129598408	0.062038643	0.129598408	0.129598408	0.216617403	0.178707714	0.129598408	0.064799204	0.078100575
0.047077634	0.024255489	0.140828628	0.06741454	0.281657257	0.211242943	0.141232901	0.194193453	0.140828628	0.070414314	0.084868302
0.070997332	0.036579472	0.106191158	0	0.106191158	0.106191158	0.141994664	0.292861301	0.212382316	0.106191158	0.06399454
0.033301275	0.017157589	0.049808927	0	0.049808927	0.049808927	0.06660255	0.183155327	0.099617854	0.049808927	0.060033235
0.034091102	0.017564526	0	0	0	0	0	0	0	0	0
0.042945736	0.022126639	0	0	0	0	0	0	0	0	0
0.045922905	0.023660545	0	0	0	0	0	0	0	0	0
0.046558964	0.023988258	0	0	0	0	0	0	0	0	0
0.044476786	0.022915471	0	0	0	0	0	0	0	0	0
0.043498156	0.022411259	0	0	0	0	0	0	0	0	0
0.021939515	0.011303747	0	0	0	0	0	0	0	0	0.03955104
0.021952906	0.011310646	0	0	0	0	0	0	0	0	0
0.036790543	0.018955341	0	0	0	0	0	0	0	0	0
0.055550305	0.028620805	0	0.079547291	0	0	0	0	0	0	0
0.035027037	0.018046742	0	0	0	0	0	0	0	0	0
0.035831796	0.018461373	0	0	0	0	0	0	0	0	0
0.034012079	0.017523812	0	0	0	0	0	0	0	0	0.030657312
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.058436236	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.154301563	0.141441799	0	0	0.046360698
0	0	0	0	0	0	0	0	0	0	0.055440228
0	0.033811897	0	0	0	0	0	0	0	0	0
0	0.035073677	0	0	0	0	0	0	0	0	0

VSM-TF-IDF

Fig. 3: Web pages in VSM representation

topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	topic10	topic11	topic12	topic13	topic14	topic15	topic16	topic17
-0.00621	0.00041	-3.3E-05	0.002121	0.001122	-0.00064	-3.1E-06	0.000288	-0.00035	-0.00165	-0.00019	-1.3E-05	0.001513	0.001823	-0.00327	-0.00077	-9.9E-05
-0.00352	0.000232	0.000332	0.000781	0.000982	0.001737	-0.00049	-0.00061	-0.00065	-0.0012	0.000398	0.000637	0.000215	0.001915	-0.00194	-0.00117	8.96E-05
0.021633	-0.00169	0.002189	-0.02331	-0.01011	-0.14008	0.018151	-0.01995	0.012774	0.001013	-0.00436	-0.00477	-0.02646	0.009333	-0.00549	-0.02509	-0.00613
-0.00964	-0.00068	0.000302	0.000479	0.00322	-0.00079	-9.6E-06	0.000221	-0.00207	-0.00084	-0.00236	-0.0004	-0.00127	0.008237	-0.00433	-0.0004	-0.00112
-0.00204	-6.7E-06	0.000327	0.000854	0.000557	-0.00031	-6.5E-05	0.001224	8.45E-05	-0.00093	-0.00054	2.1E-05	-0.00058	0.00099	-0.00097	-0.00016	3.26E-05
-0.00152	-2.2E-05	9.01E-05	0.00044	0.000691	-0.0002	-0.00019	0.001262	-0.00168	-0.00162	0.002365	0.000806	-0.00054	0.00104	-0.00178	-0.00107	-0.00066
-0.00128	4.04E-06	0.000119	0.000314	0.000492	-9.6E-05	-0.00029	0.000609	-0.00118	-0.00102	0.001482	0.000563	-0.00046	0.001154	-0.00153	-0.00099	-0.0008
-0.00247	-6.7E-06	0.000637	0.000698	0.000489	-9.8E-05	-0.00035	0.0009	-0.00083	-0.00111	0.001149	0.000734	0.00021	0.002007	-0.00216	-0.00124	-0.00077
-0.00126	3.06E-05	0.000137	0.00035	0.0004	-5.5E-05	-0.00018	0.000469	-0.00064	-0.00062	0.000738	0.000378	-8.6E-05	0.000891	-0.00117	-0.00052	-0.00048
-0.00338	-0.00031	0.00191	0.001192	0.001328	-0.00032	-0.00092	0.000549	0.000962	0.000605	-0.00099	-2.5E-05	0.000169	0.0024	-0.00168	-0.00082	0.00033
-0.00276	-6.2E-05	0.000384	0.000287	0.00061	-0.00015	-0.00048	0.000869	-0.00058	-0.00082	-3.7E-06	0.000104	-0.00037	0.002662	-0.00206	-0.00106	-0.00102
-0.00122	3.28E-05	0.000163	0.000296	0.000311	6.94E-06	7.68E-05	0.000266	-3.4E-05	-7.7E-05	-0.00048	3.88E-05	-7.4E-05	0.000928	-0.00063	-0.00018	-6E-05
-0.00111	3.85E-05	0.000167	0.000317	0.000354	-2.9E-05	-0.00043	-0.00031	-0.00014	-9.5E-05	-0.00062	7.49E-05	0.000423	0.00142	-0.00126	-0.00012	8.67E-05
-0.00286	8.85E-05	0.000253	0.000992	0.001272	-0.00011	-0.00119	-0.00077	-2.5E-05	-0.00041	-0.0016	6.83E-05	0.001549	0.003109	-0.00315	-0.00046	-0.00028
-0.00279	-0.0004	-0.00074	-0.00129	-0.0003	0.000133	0.00174	-0.00081	-0.00035	0.00063	0.000151	-0.00022	0.000487	0.000689	-0.00087	-0.00048	-0.00028
-0.00633	-0.00038	0.002017	0.000146	-0.00031	0.000141	-0.00127	5.7E-05	-0.00014	-0.00041	-0.0005	0.000669	0.002047	0.00446	-0.00563	-0.00221	0.000375
-0.00469	-0.00011	0.001706	0.00169	-0.0014	-0.00036	-0.00188	-0.00043	0.001035	-0.00026	-0.0001	0.000463	0.000934	0.002983	-0.00368	-0.00165	0.000373
-0.00194	7.96E-05	0.000462	0.000401	8.68E-05	0.000152	-0.00039	-0.0002	8.32E-05	2.12E-05	-0.00029	0.000349	0.000671	0.001808	-0.00228	-0.00093	-3.4E-05
-0.00308	0.000365	0.000697	0.001084	-4E-05	0.00011	-0.00096	-0.00043	0.000365	-0.00035	-0.00052	0.000701	0.001483	0.002397	-0.00359	-0.00164	0.000456
-0.001	4.65E-05	0.000194	0.000208	0.000115	3.55E-05	-0.00034	-1.3E-05	0.00014	0.000109	-8.2E-05	0.000175	0.000421	0.000869	-0.00118	-0.00048	-1.8E-05
-0.00095	5.65E-05	0.00022	0.000246	5.85E-05	5.86E-05	-0.00027	-0.0001	8.2E-05	-6.2E-05	-0.00016	0.000199	0.000451	0.000902	-0.00113	-0.00047	4.87E-05
-0.0011	0.00023	0.000205	0.000609	-0.00041	-5.2E-05	-0.00039	-7.6E-05	-0.00015	0.000205	-0.00013	0.000127	0.000391	0.000975	-0.00117	-0.00057	3.91E-05
-0.00093	4.84E-05	0.000224	0.000226	6.75E-05	6.16E-05	-0.00027	-4.9E-05	-2E-05	-2.6E-05	-0.00021	0.000176	0.000486	0.000906	-0.00119	-0.00045	1.49E-05
-0.00098	4.76E-05	0.000221	0.000255	0.000138	5.36E-05	-0.00026	-7.6E-05	0.000173	0.000193	-5.8E-05	0.00022	0.000367	0.000934	-0.00123	-0.00052	-2.4E-05
-0.00094	4.52E-05	0.000212	0.00019	4.12E-05	5.18E-05	-0.00029	-3.7E-05	1.49E-05	-2.4E-05	-0.00021	0.000154	0.00047	0.000941	-0.00122	-0.00045	8.79E-07
-0.00097	3.25E-05	0.000264	0.000217	4.17E-05	5.54E-05	-0.00022	-0.00011	3.59E-05	-5E-05	-0.0002	0.000199	0.000464	0.000977	-0.00125	-0.00046	2.76E-05
-0.00396	0.000302	0.000933	0.001036	0.000282	0.000225	-0.00113	-0.00029	0.000325	-0.00037	-0.00057	0.000871	0.001713	0.003456	-0.00456	-0.00145	-0.00021
-0.00138	0.000231	0.00014	0.000696	0.000142	-0.00014	-0.00056	-0.00036	-2.3E-05	0.000214	-0.00043	-2.7E-05	0.000389	0.00112	-0.00158	-0.00029	-0.00017
-0.00103	6.43E-05	0.000196	0.000264	6.04E-05	3.04E-05	-0.00033	-6.8E-05	0.000259	-0.00022	-0.00013	0.000284	0.000553	0.000908	-0.00101	-0.00032	-2.2E-05

Fig. 4: Matrix (U) words (terms) concept similarity

is used by search engine, information retrieval, text noise reduction and irrelevant feature elimination tasks. The primary goal of K-means algorithm is to partition data set into k-clusters where K here its differ from K used in reduction step, by represent each cluster with representative points that summarize cluster and assign each point to the appropriate cluster (Langhnoja *et al*., 2013), normalization can give better clustering result due to the various ranges in our data point in each space.

RESULTS AND DISCUSSION

After completing pre-processing steps the result of representing web pages corpus is a matrix of web page-words matrix as shown in the following, where the rows representing web pages, unique words (terms) in columns and entries are weights of every word (term) in web corpus. Clustering process, Fig. 5 represents the clusters for our data set and centroids for each cluster, in these clusters contain some sites which has been browsed by users each cluster represent a different space from others.

After pre-processing steps completed, LSA has implemented which is used a method called Singular Value Decomposition (SVD)on result matrix from Fig. 3, the result of applying LSA produce three matrices as discussed previously which are U, $\sum$ and V^T as shown in Fig. 6, respectively.

Figure 5 shows the result of singular values representation in descending orders which means the extracted semantic concepts in the beginning have more discriminative power than others, truncated can be done

Singular values - decreasing order															
2.888724	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	2.164182	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	2.126129	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1.962022	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1.8829	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1.841416	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1.755195	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1.656033	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1.539783	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1.486281	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1.425049	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1.393406	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1.286158	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1.246761	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.230262	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.16609
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 5: Matrix ($\sum$) represent singular values in descending order

topics/w-p	W-page1	W-page2	W-page3	W-page4	W-page5	W-page6	W-page7	W-page8	W-page9	W-page10	W-page11	W-page12	W-page13	W-page14	W-page15	W-page16	W-page17
topic1	0.032285	-0.00586	0.003132	-0.04594	0.00598	-0.40326	0.04704	-0.04867	0.024556	-0.01458	0.024279	-0.02409	0.022427	-0.02717	0.006515	-0.00666	0.014782
topic2	0.036054	-0.00965	0.001775	-0.04761	0.010194	-0.37705	0.041665	-0.04036	0.001664	0.001434	0.009921	0.004131	0.021377	-0.01949	-0.00572	0.012485	0.039459
topic3	0.04156	-0.0065	0.003333	-0.04307	0.008829	-0.40005	0.044395	-0.063	0.016468	-0.01303	0.018055	0.008453	0.024516	-0.04064	0.012011	-0.02592	0.00774
topic4	0.033507	-0.00708	0.004827	-0.04141	0.008106	-0.38073	0.044626	-0.04747	0.016642	-0.01117	0.015785	0.005993	0.024559	-0.02304	0.021556	-0.01646	0.014499
topic5	0.03368	-0.00611	-0.00267	-0.04965	0.017015	-0.33138	0.038174	-0.05133	0.013903	0.004649	0.000841	0.001062	-0.02544	-0.02097	0.022831	0.047931	-0.03646
topic6	0.04723	-0.00187	0.003629	-0.017	0.029182	-0.25455	0.029559	-0.00966	-0.01685	0.016574	-0.01171	0.033687	0.014308	-0.01055	0.003943	0.034495	0.011791
topic7	0.033549	-0.29558	-0.07818	0.075341	-0.00279	-3.3E-05	-0.01118	-0.01473	0.029299	-0.01931	0.003018	0.008106	0.002952	0.002322	-0.00185	-0.01047	-0.01107
topic8	0.039817	-0.36348	-0.09697	0.091464	-0.00474	0.000136	-0.01297	-0.00695	0.025292	-0.02855	-0.00101	0.00309	0.007906	-0.00705	0.00908	0.014238	-0.00978
topic9	0.039616	-0.37163	-0.10025	0.098349	-0.00732	0.001871	-0.01124	-0.01397	0.026632	-0.02997	0.000517	0.003162	0.003163	-0.01398	0.007381	0.003588	-0.00417
topic10	0.043359	-0.37444	-0.09918	0.09174	-0.0061	0.000974	-0.01714	-0.01697	0.027393	-0.0323	0.001528	0.000766	-0.00052	-0.00879	0.005696	0.005754	-0.0093
topic11	0.047146	-0.3695	-0.09843	0.087858	-0.0047	-0.00451	-0.00894	-0.00414	0.015569	-0.00968	-0.00078	0.00122	0.010551	-0.00674	-0.00639	0.014652	-0.01226
topic12	0.043365	-0.35194	-0.09289	0.081698	-0.00126	0.00054	-0.00903	-0.00166	0.008165	-0.00898	-0.01167	-0.00386	0.001416	0.000574	-0.0003	-0.00201	0.001562
topic13	0.074609	-0.00669	-0.01599	-0.08226	-0.13788	0.025595	0.059864	-0.08889	-0.35133	-0.31963	-0.24025	-0.0541	0.134007	-0.0193	-0.00142	0.026656	0.015144
topic14	0.058264	-0.00372	-0.01556	-0.05394	-0.07763	0.024271	-0.02132	-0.12214	-0.049	0.038986	0.170748	0.045634	0.183506	-0.15896	-0.20309	0.046447	-0.0932
topic15	0.082985	-0.01328	0.002608	-0.07077	-0.07368	0.027087	-0.05473	-0.1897	-0.06525	0.089783	0.152061	0.031837	0.161303	0.203723	0.084108	0.014395	-0.01364
topic16	0.076143	-0.00754	-0.00102	-0.05117	-0.09705	0.029275	-0.04209	-0.16751	-0.05915	0.041337	0.167914	0.041858	0.145823	-0.14712	-0.15404	-0.04698	0.027777
topic17	0.100048	-0.00229	0.014977	-0.04235	-0.07175	0.026828	-0.08443	-0.20306	-0.03148	0.020169	0.037824	-0.02294	-0.04872	-0.04635	-0.2082	-0.1559	-0.06046
topic18	0.076263	0.012103	-0.03899	-0.0415	-0.06989	0.030439	-0.03143	-0.15476	-0.03922	0.05612	0.097853	0.050624	0.052144	-0.1055	-0.12607	-0.02788	-0.09537
topic19	0.086499	0.00474	-0.00053	-0.06697	-0.05496	0.035099	-0.07457	-0.16061	-0.02348	0.032842	0.153805	0.050939	0.18998	0.197728	-0.00299	0.037682	-0.07395
topic20	0.062563	-0.01916	-0.0011	-0.06362	-0.04368	0.025228	-0.06374	-0.14271	-0.08563	0.107029	0.189254	0.071461	0.170809	0.104056	0.093138	0.101468	0.115626
topic21	0.045254	-0.0048	-0.0083	-0.04408	-0.03894	0.022073	-0.02042	-0.08612	-0.0696	0.049567	0.12905	0.041348	0.111702	-0.08666	0.014058	0.109258	0.168501
topic22	0.098998	0.018421	0.020066	0.003297	0.020812	0.001298	-0.22899	-0.03086	0.014516	-0.03924	-0.03539	-0.03621	-0.18695	0.004471	-0.03089	-0.11833	0.047373
topic23	0.103124	0.001903	0.011484	-0.04923	-0.00138	0.014264	-0.19718	-0.1514	-0.01681	-0.01122	-0.0594	-0.01626	-0.21746	0.004335	-0.05027	-0.17072	0.056113
topic24	0.097176	0.002879	0.025095	-0.00845	0.004188	-0.01822	-0.21853	-0.02895	-0.03928	-0.08973	-0.04936	0.015392	-0.14349	0.011883	0.017986	-0.09393	0.063021
topic25	0.083856	0.023364	0.010806	0.029872	0.013954	-0.01301	-0.25325	0.099996	0.034907	-0.07916	-0.02579	0.06066	-0.04958	-0.03989	0.050621	0.181608	0.150153
topic26	0.100839	0.000605	0.022238	-0.03657	0.006421	-0.07128	-0.1573	-0.08267	-0.03821	-0.07836	-0.01929	-0.01684	-0.16429	0.030697	-0.07608	-0.11995	0.079577
topic27	0.088558	0.000962	0.030987	-0.01117	0.056587	-0.02053	-0.06678	0.06496	-0.12562	0.034841	-0.06017	0.026911	0.005542	0.015103	-0.16926	0.196887	-0.07634
topic28	0.098634	0.036869	0.0712	0.186104	0.094081	0.011069	0.086941	-0.08424	-0.02836	-0.04864	0.089039	-0.16365	-0.07089	0.09636	-0.10451	0.075156	0.099815
topic29	0.099329	0.018506	0.070751	0.15529	0.098484	0.005453	0.095442	-0.09221	-0.0622	0.017564	-0.02241	0.113038	0.017452	-0.0128	0.032053	0.030571	-0.024

Fig. 6: Matrix (VT) represent web pages-concepts correlation

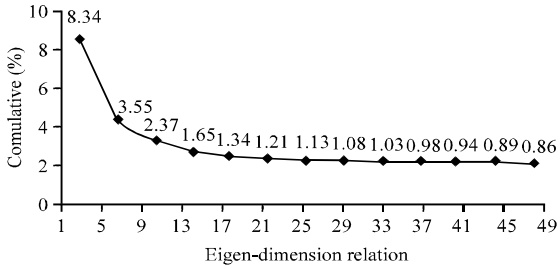

Fig. 7: Elbow Graph of SQR (σ) dimensions values

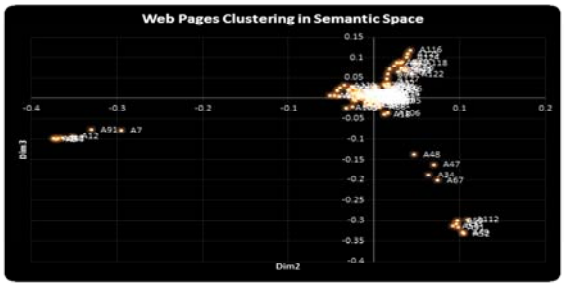

Fig. 8: Web pages clustering in semantic space using K-means

based on the result of this matrix where K has selected based on the number of web pages in the corpus. Figure 6 shows the mapping of web pages in the corpus to new extracted semantic concepts in semantic space, when K has been selected the truncated of this matrix start from bottom of matrix V^T.

To select appropriate truncated Value for the matrix (V^T) which represent the correlation of web pages to the new concepts derived in semantic space, square singular values of matrix ($\sum$) used for this purpose which reflect the importance of concepts in descending order that means the highest value assigned to concepts has higher discriminative power than others which might doesn't have that property, Fig. 7 shows the importance of extracted semantic concepts towards 50 dimensions based on singular values importance.

By using reduction of feature space in semantic analysis lot of noise, irrelevant and synonymy issuesare eliminated when extracting concept in new concept space. Therefore clustering of web pages in semantic space provide many meaningful results as shown in Fig. 8 when applying partitioning based algorithm such as K-means.

CONCLUSION

This study deal with the problem of clustering web sites pages based on discovering hidden information from the whole pages itself, clustering of web pages (documents) are differ from traditional text document clustering which due its contain lot of embedded information in its structure design as well as the extraction of pure text content from these pages is a practical challenges when corpus contain thousands of web pages comes in different design and forms. Therefore, in this study, we introduce a solution for labeling web pages by combination of using semantic analysis of our web pages corpus then applying K-means algorithm which is successfully partitioned our selected data in to set of clusters which given to us what our web pages speaking about and find similar web pages as well as similar web sites specially when labeling property absent from our selected data.

RECOMMENDATIONS

For a future research, we plan to find group of similar words together by implementing LDA concept, we searching for enhanced topic modeling process by employing non matrix factorization by selecting a proposer truncated values to produce a clusters with real topic labels.

REFERENCES

Aggarwal, C.C. and C.X. Zhai, 2012. A Survey of Text Clustering Algorithms. In: Mining Text Data, Charu, C.A., and X.Z. Cheng (Eds.). Springer, New York, USA., ISBN:978-1-4614-3222-7, pp: 77-128.

Al-Asdi, T.A. and A.J. Obaid, 2016. An efficient web usage mining algorithm based on log file data. J. Theor. Appl. Inf. Technol., 92: 215-224.

Alelyani, S., T. Jiliang, L. Huan, 2014. Feature Selection for Clustering: A Review. In: Data Clustering: Algorithms and Application, Charu, C.A. and C.K. Reddy (Eds.). CRC Press, Boca Raton, Florida, ISBN-13:978-1-4665-5822-9, pp: 29-60.

Anastasiu, D.C., A. Tagarelli and G. Karypis, 2013. Document Clustering: The Next Frontier. In: Data Clustering: Algorithms and Applications, Charu, C.A. and K.R. Chandan (Eds.). CRC Press, Boca Raton, Florida, ISBN-13: 978-1-4665-5822-9, pp: 305-338.

Andrews, N.O. and E.A. Fox, 2007. Recent developments in document clustering. MSc Thesis, Virginia Tech, Blacksburg, Virginia.

Charu, C.A., 2015a. Mining Text Data. In: Data Mining the Text Book, Aggaral, C.C. (Ed.). Springer, Switzerlands, Europe, ISBN:978-3-319-14141-1, pp: 429-456.

Charu, C.A., 2015b. Cluster Analysis. In: Data Mining the Text Book, Aggaral, C.C. (Ed.). Springer, Switzerlands, Europe, ISBN:978-3-319-14141-1, pp: 153-204.

Cooley, R., B. Mobasher and J. Srivatsava, 1997. Web mining: Information and pattern discovery on the world wide web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, November 3-8, 1997, Newport Beach, CA., pp: 558-567.

Hammouda, K.M. and M.S. Kamel, 2004. Efficient phrase-based document indexing for web document clustering. IEEE. Trans. Knowl. Data Eng., 16: 1279-1296.

Hung, C. and D. Xiaotie, 2007. New suffix tree similarity measure for document clustering. Proceedings of the 16th International Conference on World Wide Web, May 08-12, 2007, ACM, Banff, Alberta, ISBN:978-1-59593-654-7, pp: 121-130.

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Comput. Surv., 31: 264-323.

Jensi, R. and D.G.W. Jiji, 2014. A survey on optimization approaches to text document clustering. Intl. J. Comput. Sci. Appl., 3: 31-44.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Springer, Berlin, Heidelberg, pp: 137-142.

Langhnoja, S.G., M.P. Barot and D.B. Mehta, 2013. Web usage mining using association rule mining on clustered data for pattern discovery. Intl. J. Data Min. Tech. Appl., 2: 141-150.

Liu, B., 2011. Information Retrieval and Web Search. In: Web Data Mining, Liu, B. (Ed.). Springer, Berlin, Germany, ISBN:978-3-642-19459-7, pp: 211-268.

Nirkhi, S. and K. Hande, 2008. A survey on clustering algorithms for web applications. Proceedings of the 2008 International Conference on Semantic Web & Web Services (SWWS), July 14-17, 2008, CSREA Press, Las Vegas, Nevada, pp: 124-129.

Shah, N. and S. Mahajan, 2012. Document clustering: A detailed review. Intl. J. Appl. Inf. Syst., 4: 30-38.

Srinivas, M. and C.K. Mohan, 2010. Efficient clustering approach using incremental and hierarchical clustering methods. Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN), July 18-23, 2010, IEEE, Barcelona, Spain, ISBN:978-1-4244-6916-1, pp: 1-7.

Srivastava, T., P. Desikan and V. Kumar, 2005. Web Mining Concepts, Applications and Research Directions. In: Foundations and Advances in Data Mining, Wesley, C. and Y.L. Tsau (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-25057-9, pp: 275-307.

Vakali, A., G. Pallis and L. Angelis, 2007. Clustering Web Information Sources. In: Web Data Management Practices; Emerging Techniques and Technologies, Vakali, A. and P. George (Eds.). Idea Group Publishing, Hershey, Pennsylvania, USA., ISBN:1-59904-230-4, pp: 34-55.

Xiao, Y., 2010. A survey of document clustering techniques and comparison of LDA and moVMF. Master Thesis, North Carolina State University, Raleigh, North Carolina.

Yadav, M. and M.P. Mittal, 2013. Web mining: An introduction. Intl. J. Adv. Res. Comput. Sci. Software Eng., 3: 683-687.