Semantic Based Cluster Content Discovery in Description First Clustering Algorithm

MUHAMMAD WASEEM KHAN*, HAFIZ MUHAMMAD SHAHZAD ASIF*, AND YASIR SALEEM**

RECEIVED ON 16.05.2015 ACCEPTED ON 14.12.2015

ABSTRACT

In the field of data analytics grouping of like documents in textual data is a serious problem. A lot of work has been done in this field and many algorithms have purposed. One of them is a category of algorithms which firstly group the documents on the basis of similarity and then assign the meaningful labels to those groups. Description first clustering algorithm belong to the category in which the meaningful description is deduced first and then relevant documents are assigned to that description. LINGO (Label Induction Grouping Algorithm) is the algorithm of description first clustering category which is used for the automatic grouping of documents obtained from search results. It uses LSI (Latent Semantic Indexing); an IR (Information Retrieval) technique for induction of meaningful labels for clusters and VSM (Vector Space Model) for cluster content discovery. In this paper we present the LINGO while it is using LSI during cluster label induction and cluster content discovery phase. Finally, we compare results obtained from the said algorithm while it uses VSM and Latent semantic analysis during cluster content discovery phase.

Key Words: Information Retrieval, Singular Value Decomposition, Vector Space Model, Label Induction Grouping Algorithm, Term Frequency, Inverse Document Frequency.

1. INTRODUCTION

With the tremendous development of data and emergence of the Internet, situation got changed to access the data of interest impressively. A large amount of data is accessible on-line for millions of peoples in free way. But, unfortunately a small part of this population can benefit from this information available in this library without proper indications. It is a fact that more than 80% of the available data is in the text form. Many search engines have been introduced to mine data from this library which uses different algorithms to mine and analyze the text data on the basis of the internal relationship of the data. Clustering of the data is an approach that is used to group the similar data. This technique was used in the Scatter-Gather [1] system for the first time; after this many algorithms including STC (Suffix Tree Clustering) were initiated to use the concept of phrases to find the similarity among the documents [2,3]. SHOC (Semantic Hierarchical Online Clustering) [4,5] is an algorithm of the same kind. MSEEC [6] and Vivisimo are search engines that use algorithms of such type which use the idea of grouping the documents on commercial basis [7].

* Department of Computer Science & Engineering, University of Engineering & Technology, Lahore.
** Department of Computer Sciences, COMSATS Institute of Information Technology, Sahiwal.

Description first clustering algorithms are used to increase the quality of cluster labels and readability of thematic groups. Including lexical terms it also considers the phrases as well for the candidates of cluster labels.

LINGO is the algorithm of type description first. This algorithm in its existing form does use a novel IR technique LSI for the purpose of cluster label induction and VSM for cluster content discovery. In this paper we use LSI during both phases; cluster content discovery and cluster label induction.

2. THEORETICAL BACKGROUND

2.1 Vector Space Model

VSM is an IR technique in which a text document is represented as a multidimensional vector. In VSM we compare algebraic vectors instead of text documents because once we are able to represent a text document into an algebraic vector then algebraic operations can be used to find out the similarities among the vectors. In VSM each vector v represents a document j in multidimensional space. Each element v_{ij} represents a specific term of document j and value of this term in the vector represents the strength of relationship of term i to document j. The matrix of the vectors we construct is called term x documents matrix (A). In matrix A rows represent the number of terms of the documents and columns represent the number of the documents d. Element a_{ij} of the matrix A represents the relationship between term i and document j. Number of term weighting schemes including "Binary weighting", "term frequency", "term frequency inverse document frequency" can be used to measure this degree of relationship according to the requirements [8]. After the construction of matrix A various methods are available to measure the distance between vectors representing document a and b; mostly cosine similarity calculation is used. Formula is given below:

$$\cos\theta_j = \frac{a_j^T q}{\|a_j\|\|q\|} - \frac{\sum_{j=1}^{t} a_{ij} q_i}{\sqrt{\sum_{i=1}^{t} a_{ij}^a} \sqrt{\sum_{i=1}^{t} q_{ij}^a}}$$

2.2 Latent Semantic Indexing and Singular Value Decomposition

It is a new IR technique which deals with the limitations of VSM. It does not consider just the lexical terms but the underlying concepts behind the terms. It deals with the issues of terms relating to words by concepts driven by statistical methods and use as replacement of the lexical terms behind the documents [9]. Performance analysis shows that statically derived concepts are the more rich indicators of information than the individual terms in the documents.

SVD (Singular Value Decomposition) is a fundamental mathematical construct behind the LSI. It decomposes the original term-document matrix A into three sub-matrices U, S and V^T. Matrix U is of txt dimensions and its columns constitutes the left singular vectors of A. In matrix S constitutes singular values on its diagonal in the descending order. Matrix S is the matrix of t x d dimensions. V^T matrix is an orthogonal matrix which has the dimensions of dxd and its column vectors comprise the right singular vectors of original matrix A. SVD has the quality that we can reconstruct the original matrix A by the reduced dimensions of matrix U, S and V^T upto k dimensions. Detailed review is available in [10]. It depends upon the user that up to what extent he wants to eliminate the extraneous terms. The larger the value of k the larger is the proximity to get the original matrix. So k is chosen in such a way that at least 80% of information content in the original matrix should be retained.

3. OVERVIEW OF THE ALGORITHM

While designing clustering algorithms it is necessary to pay special attention on the presentation of cluster labels and contents of those clusters. It should be emphasized that Labels and contents of the clusters should have a meaning to users. Most of the algorithms of this field follow the strategy to find the contents of the clusters first and then on the basis of these contents assign the appropriate labels to them. Without considering any similarity measure among the labels and contents it might

be possible that labels might not be the proper representative of those groups. To avoid such type of problems LINGO attempts to find meaningful cluster labels first and then assign the appropriate documents to the labels. It considers the frequent phrases and lexical terms from input documents as label candidates and chooses the appropriate labels after pruning. After ensuring the description quality of labels it assigns the documents to labels.

Algorithm-1 is the pseudo-code when it uses LSI in cluster content discovery. Particular steps are also given in the following sections.

3.1 Preprocessing and Frequent Phrases Extraction

Preprocessing is the preliminary step in any IR technique. In this step we remove the stop words and other un-necessary tags from the available dataset to reduce the effect of these terms on our results. Because these terms might affect our results negatively [11].

Frequent phrases are defined as the ordered terms which occur in the input documents in a repeated manner. We consider these terms also as cluster label candidates with the lexical terms because it is a fact that good writers use the idea of synonymy to express his views and to get the attention of his readers. So by considering the idea of synonymy a sentence can be represented in variety of ways by avoiding the repetition. SVD got the potential to identify abstract concepts behind the document [12].

To be a candidate for the label of the cluster a phrase or term must occur more than a specific threshold; the term or phrase must not start or end with a stop word or tag and it must be a complete phrase or term. A complete phrase or term is defined as it should not be possible to get another term or phrase by adding or removing a substring. It cannot be extended by adding or removing an element into it. These assumptions are discussed in [2,4].

3.2 Cluster Label Induction

Once we have extracted frequent phrases from the input data; they are considered as candidates of clusters labels [13-15]. There are three steps involved; construction of txd matrix, discovering the abstract concepts and trimming of labels.

The txd matrix is constructed from the input dataset by representing each sentence as a vector of the multidimensional matrix that contains the terms of the sentence exceeding from the defined frequency threshold. A term weighting scheme *tfidf* (term frequency inverse document frequency) is used to measure the weight of each term in the vector [16-19]. To discover the abstract concepts of term-document matrix SVD is applied on the matrix to find the orthogonal basis. As we know that SVD decompose the matrix A into sub matrices U, S and V^T. Matrix U represents the abstract concepts of the matrix A. For future calculation we use the reduced U matrix upto k terms. The value of k is selected by calculating the Frobenius norms of the matrix A and matrix A_k; which is a reduced matrix upto k dimensions. Let us define a threshold q which represents a percentage value that upto what extent the original information of matrix A should be retained. So larger the value of k larger will be the information restored in the A_k matrix. The condition $\|A_k\|_F / \|A\|_F \geq q$ should be satisfied which is the Frobenius norm of matrix X.

In the step of phrase matching and label pruning the phrases and abstract concepts are represented as column vectors of the same vector space of matrix A. By doing so we would be able to use cosine similarity to calculate how close is a term and a phrase to abstract concept. Let we refer it with P a matrix of size tx(p+t). Where t is the number of terms and p is the number of phrases we have calculated from original dataset. Many tools are available to calculate phrases from the dataset. In this work we have used Maui Indexer and Kia for this purpose. Having t and p we can easily construct the matrix P by using one of the weighting schemes. Here we have used $t_i i_{df}$ and the frequency of the term is the frequency in the original

dataset. Once we have matrix P and i_{th} column vector of matrix U we calculate the cosine distance between abstract concept and phrase by the formula $m_i=U^T_i*P$. This process can be extended upto entire matrix U and matrix M; a matrix of cosines between P and U is constructed by $M = U^T_k*P$ formula. The component of matrix M with maximum value is considered as candidate of cluster labels. At the end of cluster label induction phase candidate cluster labels are pruned to induce the cluster labels. For this purpose we construct another matrix Z in which candidates for cluster labels are represented as a documents and we calculate the $Z*Z^T$ which produces a matrix of similarities among the cluster labels. From each row we pick the column which exceeds the defined threshold and leave all but the candidate with maximum score.

3.3 Cluster Content Discovery

In this phase documents in the corpus are allocated to the cluster labels inducted in the former phase. For this purpose LSI technique is used. In this process of assigning the documents to cluster labels we construct the matrix Q in which induced labels are represented as column vectors and multiply it with the matrix A_k. Matrix A_k is a matrix which we have reconstructed by the reduced dimensions of matrix U, S and V^T upto k dimensions; Let $C = Q^T*A_k$. Now in the matrix C the element c_{ij} will give the strength of relationship between document j in cluster i. We will assign the document j to cluster i where the value of c_{ij} increase from specific threshold. The remaining documents which do not fall into any cluster end up with an artificial group called topic of others.

3.4 Final Cluster Formation

For the purpose of presentation, clusters are presented in sorted orders based on their score. The score of the cluster is calculated by a simple formula given below:

Cluster score = score of label x $||C||$

This score function favors large clusters over smaller. We may say that it gives the quality of clusters.

4. RESULTS AND EVALUATION

The study is evaluated by performing experiments on several datasets. In each experiment it gives us better results as compared to existing methodology; while it is used to find cluster contents. It assigns the documents to appropriate labels and reduces the group of un-assigned documents named "Others". It can be observed clearly that our proposed methodology in which we have found the contents of clusters by using LSI has reduced the topic of others remarkably. It has also been observed in the results that our technique has grouped the documents in the most relevant cluster. In Fig. 2, a significant change in group of others comparison among three datasets D1, D2, D3 is shown.

(i)	D⟵Input Documents (or Snippets)	
(ii)	{Step-1: Preprocessing}	
(iii)	{Step-2: Frequent Phrase Extraction}	
(iv)	{Step-3: Cluster Label Induction}	
(v)	{Step-4: Cluster Content Discovery}	
	(a)	A⟵Term-Document matrix of terms marked as stop-words and with frequency higher than the Term Frequency Threshold
	(b)	Σ, U, V⟵SVD(A); {Product of SVD Decomposition of A}
	(c)	U_k⟵reduce U matrix obtained from SVD of (A)
	(d)	Σ_k⟵sigma matrix of singular values
	(e)	V^T_k⟵reduce matrix of left singular vectors
	(f)	$A_k = U_k * \Sigma_k * V^T_k$⟵reduce txd matrix upt k terms
	(g)	Fall all L ∈ Cluster Label Candidates do
	(h)	Create Cluster $C = Q^T * A_k$ described with L
	(i)	Add to C all documents whose similarity to C exceeds the Snippet Assignment Thresold
	(j)	End for
(vi)	{Step-5: Final Cluster Formation}	
	(a)	Fall all clusters do
	(b)	Cluster Score Label Scores * ||C||;
	(c)	End for

FIG. 1. PSEUDO-CODE OF LABEL INDUCTION GROUPING ALGORITHM WHILE USING LSA IN CLUSTER CONTENT DISCOVERY

The quality of cluster is determined by its score. Fig. 3 shows a graph that clarifies comparison between proposed and existing methodology. A significant change in the cluster quality on D1 can be observed from the graph.

Fig. 4 depicts a comparison of another dataset D2. It confirms the fact that the quality of clusters in new methodology is much better in results than that of the existing technique.

Fig. 5 shows third tier of comparison made on dataset D3. The third version again identifies a major difference in the quality of clusters in new methodology.

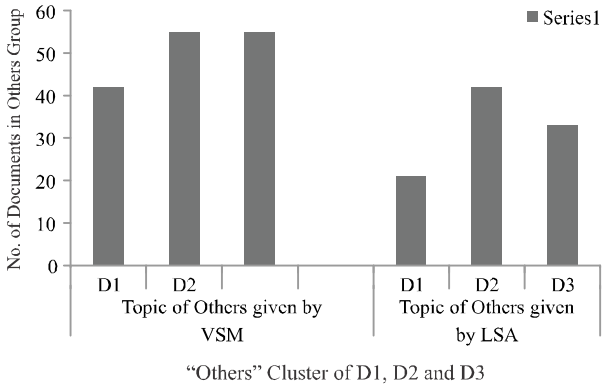

FIG. 2. TOPIC OF OTHERS COMPARISONBETWEEN VSM AND LSI GRAPHICAL REPRESENTATION

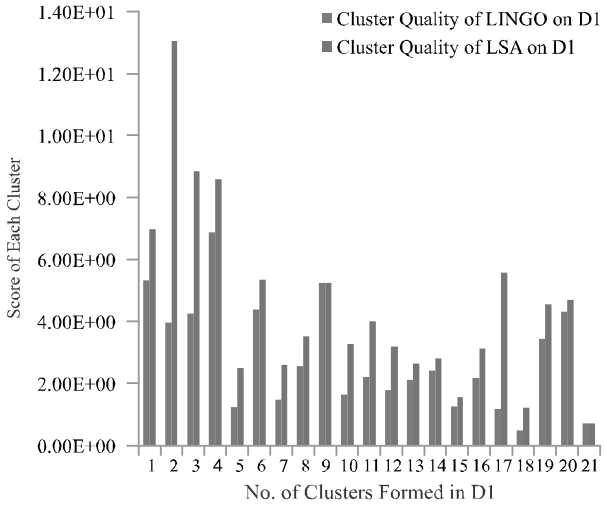

FIG. 3. QUALITY OF CLUSTERS GIVEN BY VSM AND LSA ON D1

5. CONCLUSION

In this work we have presented that how different are the results of LINGO when it adopts the method of assigning the documents to induced labels on the basis of statistically driven concepts instead of on the basis of lexical terms. The aspiration of this work is a new clustering algorithm named as LINGO. This algorithm is used for the automatic grouping of results incurred from the search engine against a query. This algorithm focuses on the label quality and finds the contents of induced labels in the traditional way using VSM. Particularly our share in

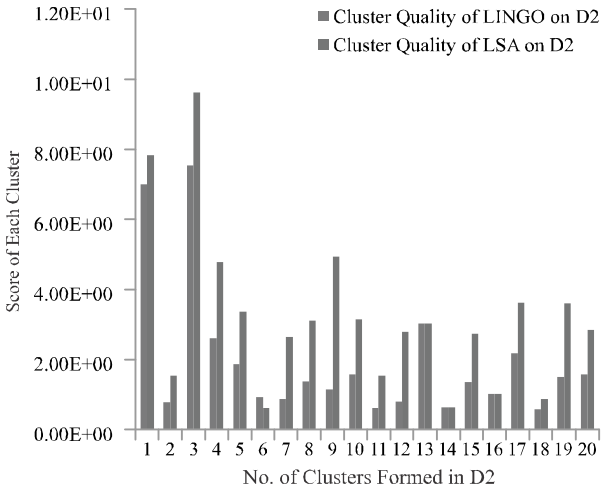

FIG. 4. QUALITY OF CLUSTERS GIVEN BY VSM AND LSA ON D1

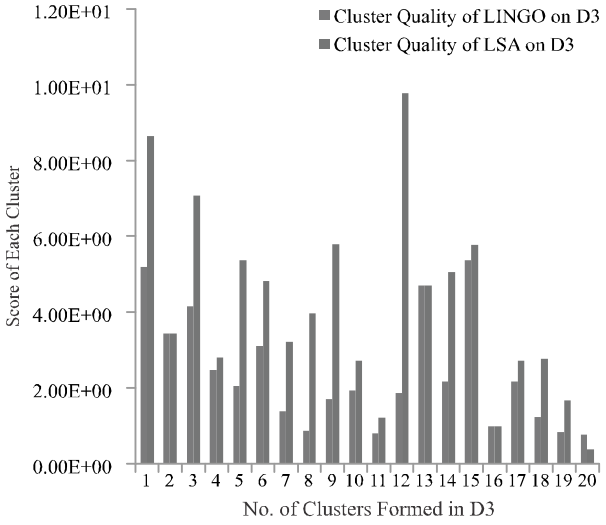

FIG. 5. QUALITY OF CLUSTERS GIVEN BY VSM AND LSA ON D1

this work is the assignment of documents to induced labels on the basis concepts rather than on the basis of lexical terms. From the graphical representation it is clear that our method of assigning the documents to labels gives better results. It reduces the topic of others and improves the cluster quality. The work on the algorithm is not finished completely. More advanced methods can be brought in to induce hierarchical relationships among the topics. Finally a detailed valuation technique will be necessary to constitute weak points of the algorithm.

ACKNOWLEDGEMENT

Authors are very thankful to University of Engineering & Technology, Lahore, Pakistan, for providing them the conducive environment for studies. First author also want to pay thanks to his parents, friends and colleagues for their motivation and support throughout in his studies.

REFERENCES

[1] Hearst, M.A., and Pedersen, J.O., "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results", Proceedings of 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996.

[2] Zamir, O., and Etzioni, O., "Grouper: A Dynamic Clustering Interface to Web Search Results", Computer Networks, Volume 31, No. 11, pp. 1361-1374, 1999.

[3] OsiDski, S., Stefanowski, J., and Weiss, D.L., "Search Results Clustering Algorithm Based on Singular Value Decomposition", Intelligent Information Processing and Web Mining, pp. 359-368, Springer, 2004.

[4] Zhang, D., "Towards Web Information Clustering", Ph.D. Dissertation, Southeast University, Nanjing, China, 2002.

[5] Li, P., Wang, H., Zhu, K., Wang, Z., Hu, X.-G., and Wu, X., "A Large Probabilistic Semantic Network Based Approach to Compute Term Similarity", IEEE Transactions on Kowledge and Data Engineering, Volume 27, No. 10, pp. 2604-2617, October, 2015.

[6] Hannappel, P., Klapsing, R., and Neumann, G., "Mseec-A Multi Search Engine with Multiple Clustering", Proceedings of 99th Conference on Information Resources Management Association, 1999.

[7] MasBowska, I., "Phrase-Based Hierarchical Clustering of Web Search Results", Springer, 2003.

[8] Skoutas, D., Sacharidis, D., Simitsis, A., and Sellis, T., "Ranking and Clustering Web Services Using Multicriteria Dominance Relationships", IEEE Transactions on Services Computing, Volume 3, No. 3, pp. 163-177, 2010.

[9] Farhan, M., "A Methodology to Enrich Student-Teacher Interaction in Elearning", 30th ACM/SIGAPP Symposium on Applied Computing. 2015.

[10] Baker, K., "Singular Value Decomposition Tutorial", The Ohio State University, No. pp. 1-24, 2005.

[11] Farhan, M., Ahmed, A., Ramzan, M., Bashir, S.R., Iqbal, M.M., and Ahmed, F., "Reusable Open Source Software Component's Life Cycle Management", International Journal of Multidisciplinary Science and Engineering, Volume 3, No. 4, 2012.

[12] Farhan, M., Munwar, I., Aslam, M., Enriquez, A.M., Farooq, A., Tanveer, S., and Mejia, A.P., "Automated Reply to Students' Queries in e-Learning Environment using Web-Bot", IEEE 11th International Conference on Artificial Intelligence, Mexican, 2012.

[13] Farhan, M., Zahra, R., Iqbal, M.M., and Aslam, M., "Extracting Parameters from e-Feedback Text Data Based on Keyword Based Parsing in Elearning Environment", Science International, Volume 26, No. 3, 2014.

[14] Farhan, M., Muhammad, S., Anwar, M., and Mohsin, A., "Cpecaee: Collaboration Platform for Extra Curricular Activities in the e-Learning Environment", International Journal of Multidisciplinary Sciences & Engineering, Volume 2, No. 6, pp. 1-4, 2011.

[15] Iqbal, M.M., Farhan, M., Saleem, Y., and Aslam, M., "Automated Web-Bot Implementation using Machine Learning Techniques in Elearning Paradigm", Journal of Applied Environmental and Biological Sciences, Volume 4, No. pp. 90-98, 2014.

[16] Salton, G., "Automatic Text Processing: The Transformation, Analysis, and Retrieval of", Reading: Addison-Wesley, 1989

[17] Berry, M.W., Dumais, S.T., and O'Brien, G.W., "Using Linear Algebra for Intelligent Information Retrieval", SIAM Review, Volume 37, No. 4, pp. 573-595, 1995.

[18] OsiDski, S., "An Algorithm for Clustering of Web Search Results", PoznaD University of Technology, Poland, 2003.

[19] Osiñski, S., and Weiss, D., "Conceptual Clustering using LINGO: Evaluation on Open Directory Project Data", Intelligent Information Processing and Web Mining, pp. 369-377, Springer, 2004.