DOCUMENT CLUSTERING USING ARTIFICIAL NEURAL NETWORKS

K.Selvi[1] R.M.Suresh[2]

[1]Research Scholar, Sathyabama University, Chennai, India.
[2]Sri Muthukumaran Institute of Technology Chennai, India.
ssi.cse@rmkec.ac.in, rmsuresh@hotmail.com

Abstract

Finding the required information in enormous amount of data is one of the challenging tasks of today. Document clustering is a fundamental task in Text mining that has been using statistical methods. Due to a large number of documents available in the web, intelligent methods have to be developed to retrieve the documents effectively. In this paper, the method using artificial neural networks (ANN) in document clustering is discussed. By fine tuning, the various parameters of the ANN algorithms, effective document clustering can be achieved.

Key words: Information Retrieval , Text Mining, Document Clustering, Artificial Neural Networks, Similarity Measure

I. INTRODUCTION

Document clustering is important for quick accessing of relevant documents in the web for given pair of words. The time taken for retrieving the documents should be less than a second so that the user can verify whether the documents what they are looking for are the right documents. Web search engines require algorithms and software that can satisfy the query from the users and return relevant documents correctly in time. Therefore, there is a high need for a new document clustering algorithms, which are more efficient than conventional clustering algorithms (13).

The increasing nature of World Wide Web has imposed high challenges for researchers to cluster the similar documents over the internet thereby improving the efficiency of search. Search engine is getting more confused in selecting the relevant documents among huge volumes of search results returned to a simple query. A potential solution to this problem is to cluster the similar web documents, which helps the user in identifying the relevant data easily and effectively (7).

Most internet web documents are publicly available for providing services required by the user. In such documents, there are no confidential or sensitive data (open to all). Then how can we provide privacy of such documents? Now days, same information exists in more than one document in duplicate forms. One way of providing privacy of those documents is to avoid document duplication, thereby protecting the privacy of individual copyrights of documents. Many duplicate document detection techniques are available such as syntactic, URL based, semantic approaches (12).

Two documents are considered to be similar if their similarity measure is 1 and they are considered as different if the similarity measure is partially or not fully 1(17). Document clustering tasks can generally be divided into two categories: offline techniques that seek to cluster a static, previously compiled collection of documents, and online techniques that operate on an incrementally compiled set of documents.

In order to achieve good document clustering, best features have to be selected. The selection of best similarity measure and appropriate algorithms for comparison of documents are considered in this paper for effective document clustering.

This paper is organized as follows: A detailed introduction to Document Clustering and Similarity measure techniques to overcome document duplication are discussed. In Section 2, a comprehensive survey on the study of research methods and the problems faced by the same are highlighted. Section 3, briefs the network algorithms used in the proposed method. In Section 4 the proposed method is summed up.

II. RELATED WORK

Ling Zhuang Honghua Dai 2004 introduced the initial points as centers for k-means algorithm. However, k-means clustering is a completely unstructured approach,

sensitive to noise that produces an unorganized collection of clusters not favorable for interpretation [Michael Steinbach, et al, 2000].

To minimize the overlapping of documents, Beil, et al, 2002 introduced a method called HFTC (Hierarchical Frequent Text Clustering) which is another frequent itemset based approach to choose the next frequent item sets. The clustering result depends on the order of choosing next frequent itemsets. The resulting hierarchy in HFTC contains many clusters at first level. As a result, the documents in the same class are distributed into different branches of hierarchy, which decreases the overall clustering accuracy.

BenjaminFung, et al, 2003, has introduced FIHC (Frequent Item set based Hierarchical Clustering) method for document clustering, which employed a cluster topic tree , constructed based on the similarity among clusters. Wordsets based document clustering algorithm for large datasets was introduced by Sharma et al., 2009.

Web document clustering using document index graph is put forth by Momin et al., 2006. Document clustering methods are generally based on a single term examination of document data set. To attain more precise document clustering, more informative feature like phrases are essential in this scenario. Document Index Graph (DIG) that permits incremental phrase-based encoding of documents is capable of phrase matching. It stress on efficiency of phrase-based similarity measure over conventional single term based similarities. In the second part, a Document Index Graph based Clustering (DIGBC) algorithm is provided to improve the DIG model for incremental and soft clustering. This technique incrementally clusters documents based on presented cluster-document similarity measure. It permits assignment of a document to more than single cluster.

Muflikhah et al. 2009, introduced space and cosine similarity measurement. The authors used Latent Semantic Index (LSI) approach with Singular Vector Decomposition (SVD) or Principle Component Analysis (PCA). The method decrease the matrix dimension by identifying the pattern in the document collection which refers to simultaneous terms. Every technique is employed to weight of term-document in vector space model (VSM) for document clustering with the help of fuzzy c-means technique. Affinity-based similarity measure for Web document clustering is presented by Shyu et al., 2004.

The concept of document clustering is extended into Web document clustering by establishing the approach of affinity based similarity measure, which makes use of the user access patterns in finding the similarities among Web documents through a probabilistic model. Various experiments are conducted for evaluation with the help of real data set, and the experimental results illustrated that the presented similarity measure outperforms the cosine coefficient and the Euclidean distance technique under various document clustering techniques.

Eldesoky et al., 2009, given a novel similarity measure for document clustering based on topic phrases. In the conventional vector space model (VSM) researchers have used the unique word that is contained in the document set as the candidate feature. Currently the latest trend which uses the phrase to be a more informative feature has considered the issue that contributes in enhancing the document clustering accuracy and effectiveness. This paper presented a new technique for evaluating the similarity measure of the traditional VSM by considering the topic phrases of the document as the comprising terms for the VSM instead of the conventional term and applying the new technique to the Buckshot technique, which is a combination of the Hierarchical Agglomerative Clustering (HAC) technique and the K-means clustering method. Such a method may increase the effectiveness of the clustering by incrementing the evaluation metrics values.

Cobos et al., 2010 used k-means, termsets, Bayesian information for web document clustering. Haojun et al., 2008 developed hierarchical algorithms for document clustering by using cluster overlapping rate to improve efficiency. The expectation-maximization (EM) method was used in the Gaussian mixture model to count the parameters and formulate the two sub-clusters combined when their overlying is the biggest.

III. ARTIFICIAL NEURAL NETWORK

Artificial neural networks (ANN) are a mathematical model and visual representation of neural connection in a human brain. The ANN is represented by a topology of connections among different elements. The elements are nodes or neurons or perceptions that are arranged in

layers. The link among neurons of the layers is expressed by connection strength using matrices for computation purposes. The connections are otherwise the fluid in which the neurons are placed in the brain. The ANN topology is trained by algorithms developed by different researchers. The algorithms are implemented in a computer program. Data have to be presented to the program so that, a set of representative numbers called final weights are obtained and can be stored in the database for subsequent use.

IV. PROPOSED METHOD

Documents are files which contain ascii and non ascii characters. Sample categories are thesis, paper, journal, Invoice, quote, RFP, Proposal, Contract, Packing slip, Manifest, Report detailed & summary, Spread sheet, Waybill, Bill of Lading, Financial statement, Nondisclosure agreement, Mutual nondisclosure agreement, summons, certificate, license, gazette, white paper, application forms, user-guide, brief, mock-up, script. These documents can be created using templates. The documents can be computerized by using different types of editors starting from basic editors which use only ascii characters to sophisticated editors which embed graphics. The graphics in a document is represented by special characters.

Searching of similar documents involve processing of ducuments where elimination of information should be avoided taking care of word pair search. When the documents are properly preprocessed then the measures of similarity will be perfect, and the quality of search will be maximum.

The documents should first be preprocessed. In the preprocessing step, the documents should be transformed into a representation suitable for applying the learning algorithms. The most widely used method for document representation is the vector space model introduced by Salton et al, 1975,

Each document can be converted into a vector D. Each vector contains more than two dimensions. Each dimension of a vector is a unique representation of a document. When the contents of two documents are almost similar, then the numerical values of the vector also be almost same. Each feature of the vector will represent a distinct term. The term is a single word or phrase. A phrase can have more than one word. Phrases can be extracted by using statistical or Natural Language Processing (NLP) techniques. By statistical methods, Phrases can be extracted using the frequently appearing words in the document. Extraction of phrases by using NLP presented by Fuernkranz et al, 1998.

Steps for training ANN algorithms:

Phase-1: Preprocessing the document

Step 1: Pair of words is given in the search space.

Step 2: In the search space each document is preprocessed and converted into vectors of numerical values. If the vectors of numerical values are already available in the searching folder corresponding to the available documents, then preprocessing of the documents and converting into vectors need not be carried out.

Step 3: The bags of representation of words (distinct single words) is created. During this process, irrelevant data are omitted for calculation purposes.

Step 4: Indexing the bags of words is done. In this process, the following tasks are carried out:

Tokenization where the string is segmented into tokens by white space and punctuations.

Each token is **stemmed** into its root form by converting a noun into its singular form and removing grammatical words like articles, conjunctions, or pronouns.

Stop Word Elimination

The remaining words are coded into numerical values.

Step 5: Converting vectors into a matrix is formed. In this matrix presence of absence of words corresponding to a document is indicated using 1 or 0. On the other hand, frequency of words is represented in fractions in the matrix by normalization.

Phase-2: Training the ANN algorithm

In phase-2, depending upon the type of algorithm used for training the topology of ANN, training patterns are formed from the matrix of numbers formed in Step 5 of phase-1. The type of algorithm that can be used for training the ANN topology can be a) supervised algorithm, 2) unsupervised algorithm and 3) recurrent algorithm.

Supervised algorithm: In this algorithm, training patterns with input features representing a document is

used. A labelling feature that correspond to the pattern is used. Hence, the training pattern should contain, input features and target output (labelling). Examples of supervised learning algorithms are back propagation algorithm, radial basis function, counter propagation algorithm and time delayed network.

Unsupervised algorithm: In this algorithm, training patterns with input features representing a document is used. No labelling feature is used as target. Hence, the training pattern contains only input features. Examples of unsupervised algorithms are Wake-sleep algorithm, Helmholtz machine, Generalized Hebbian Algorithm, Deep neural network, adaptive resonance theory and self-organizing map.

Recurrent algorithm: In this algorithm, training patterns with input features representing a document is used. A labelling feature that correspond to the pattern is used. The algorithm is used in the ANN topology with directed cycle. Examples of the recurrent algorithms are Fully recurrent network, Hopfield network, Elman networks and Jordan networks, Echo state network, Long short term memory network, Bi-directional RNN, Continuous-time RNN.

The algorithms are developed based on different weight updating rules. In each weight updating rules, errors are calculated in the forward process of the ANN and weight updation are done during the reverse or recurrent process.

In the training process of the ANN algorithms, the connections among the nodes between layers are represented by matrices. In most of the algorithms, the matrices are initialized with random numbers. At the end of the training process, the matrices contain final weight values for mapping inputs and outputs. During the testing of the ANN or testing the retrieving of the documents corresponding to word pair, final weights are used for processing with the vector corresponding to word pair. In another method, final weights are obtained from the pattern itself without any initialization of the weight matrices.

Phase 3: Testing the ANN for document clustering

Assuming that the documents already preprocessed and the vector representation exists, the word pair is converted into vector and processed with final weights obtained from an ANN algorithm. The outputs in the output layer of the ANN is further used for similarity measures. Instead of using similarity measures, the outputs of ANN can be as well interpreted whether the documents retrieved or clustered are relevant to the words. Inorder to evaluate the quality of the implemented algorithms for document clustering,

Phase 4: Document Similarity Measure

To use a clustering or classification algorithm, a similarity measure between two documents must be defined. Some of the existing similarity measures are as follows: a) Euclidean Distance, b) Cosine Similarity, c) Jaccard Coefficient, d) Pearson Correlation Measure, and e) Multiviewpoint-Based Similarity Measure.

V. CONCLUDING REMARKS FROM SIMILARITY MATRIX

The similarity matrix refers to a word-by-word square matrix where each entry indicates a semantic similarity between two words corresponding to rows and columns. The similarity matrix has two properties: its diagonal elements are 1.0 and it is a symmetry matrix.

In similarity matrix, its columns and rows correspond to words; the i^{th} column and the i^{th} row correspond to the identical word. The entry of the similarity matrix, indicates the semantic similarity between the word which corresponds to i^{th} column or i^{th} row and the word which corresponds to j^{th} column or j^{th} row.

Therefore, in the given corpus, the more the documents include both words, the higher the semantic similarity between the words. The similarity matrix is that it is symmetric.

The semantic similarity between two words characterizes the similarity matrix. The similarity matrix is that its diagonal elements are given as 1.0s. In other words, the value of s_{ii} is 1.0. This property shows that although two identical words may have different meanings depending on their context.

VI. CONCLUSION

This paper emphasizes the applicability of artificial neural networks in document clustering. The different phases that have to be considered for document clustering is given in four phases. The method using

neural network algorithms for document clustering is described. As part of future work, ANN algorithms can be implemented in document clustering.

REFERENCES

[1]. Beil F., Ester M., Xu, X., 2002, Frequent Term-Based Text Clustering, In Proceedings of 8th International Conference on Knowledge Discovery and Data mining 2002 .

[2]. BenjaminFung, C.M., Wang Ke., Ester Martin, 2003, Hierarchical Document Clustering using Frequent Item Sets. In Proceedings SIAM International Conference on Data Mining, pp.59-70.

[3]. Cobos C., Andrade J., Constain W., Mendoza M., and Leon E., 2010, Web document clustering based on Global-Best Harmony Search, K-means, Frequent Term Sets and Bayesian Information Criterion, IEEE Congress on Evolutionary Computation, pp.1-8.

[4]. Eldesoky A.E., Saleh M., and Sakr N.A., 2009, Novel Similarity Measure for Document Clustering based on Topic Phrases, International Conference on Networking and Media Convergence, pp.92-96.

[5]. Fuernkranz J., Mitchell T., and Riloff E., 1998, A Case Study in Using Linguistic Phrases for Text Categorization on the WWW, Sahami M., editor, In Learning for Text Categorization:

[6]. Haojun Sun., Zhihui Liu, and Lingjun Kong, 2008, A Document Clustering Method Based on Hierarchical Algorithm with Model Clustering, 22nd International Conference on Advanced Information Networking and Applications, pp.1229-1233.

[7]. Jain A.K., Murty M.N., Flynn P.J., 1999, Data Clustering: A Review, In the Proceedings of ACM Computing Surveys, Vol.31, No.3, pp.264-323.

[8]. Ling Zhuang, Honghua Dai., 2004, A Maximal Frequent Item Set Approach for Web Document Clustering In Proceedings of the IEEE Fourth International Conference on Computer and Information Technology.

.

[9]. Michael Steinbach, George karypis, and Vipinkumar 2000, A Comparison of Document Clustering Techniques, In Proceedings of the Workshop on Text Mining, 2000 (KDD-2000), Boston, pp.109-111.

[10]. Momin, B.F., Kulkarni, P.J., and Chaudhari, A., 2006, Web Document Clustering Using Document Index Graph, International Conference on Advanced Computing and Communications, Pp. 32 - 37.

[11]. Muflikhah L., and Baharudin B., 2009, Document Clustering Using Concept Space and Cosine Similarity Measurement, International Conference on Computer Technology and Development, Vol.**1**, pp.58-62.

[12]. Prasannakumar J., and Govindarajulu P., 2009, Duplicate and Near Duplicate Documents Detection: A Review. European Journal of Scientific Research ISSN 1450-216X Vol.32 No.4, pp.514-527.

[13]. Ruxixu and Donald Wunsch, 2005, A Survey of Clustering Algorithms. In the Proceedings of IEEE Transactions on Neural Networks, Vol.16, No.3, pp.645-678.

[14]. Salton G., Yang C., and Wong A., 1975, A Vector-Space Model for Automatic Indexing, Communications of the ACM, Vol.18, No.11, pp.613–620.

[15]. Sharma, A., and Dhir R., 2009, A Wordsets based Document Clustering Algorithm for Large datasets, Proceeding of International Conference on Methods and Models in Computer Science.

[16]. Shyu M.L., Chen S.C., Chen M., and Rubin S.H., 2004, Affinity-based similarity measure for Web document clustering, IEEE International Conference on Information Reuse and Integration, pp.247-252.

[17]. Syed Mudhasir Y., and Deepika J., 2011, Near Duplicate Detection and Elimination Based on Web Provenance for Efficient Web Search. In the Proceedings of International Journal on Internet and Distributed Computing Systems, Vol.1, No.1, pp.22-32.