An Approach for Document Clustering using Agglomerative Clustering and Hebbian-type Neural Network

Gopal Patidar  
Student [M.Tech. (CSE)]  
BUIT, BU Bhopal

Anju Singh, Ph.D  
Asst. Proff, Dept. of IT BUIT  
Barkatullah University, Bhopal

Divakar Singh, Ph.D  
HOD (CSE), BUIT  
Barkatullah University, Bhopal

ABSTRACT
Clustering is a useful method that categorizes a large quantity of unordered text documents into a small number of meaningful and coherent collections, thereby providing a basis for instinctive and informative navigation and browsing mechanisms. Different type of distance functions and similarity measures have been used for clustering, such as squared cosine similarity, Euclidean distance and relative entropy.

This paper presents text document space dimension reduction in text document retrieval by agglomerative clustering and Hebbian-type neural network. Hebbian-type neural network reduce document space to two dimensions so each document is represented as a point in the reduced document space. Furthermore, the clusters are formed in compact document space.

KEYWORDS
Agglomerative and Oja Learning Rule, of hebbian-type neural network, F-measure.

1. INTRODUCTION
Information retrieval from text documents in natural language is widely used procedure nowadays. Due to the large number of documents, mainly on the internet, it is needed to cope with large amounts of data. There are billions of pages on the internet, so the dimension reduction becomes very important area of research in information retrieval. Typical information retrieval system consists of user subsystem and document subsystem. User sends a query on the document subsystem. Document subsystem finds relevant documents from the document collection to this query a brings them back to the user. User then can find relevant information in the document. Information retrieval systems from text documents typically uses bag-of-words approach [1, 2, 3], which is often called vector space model.

The vector space model matrix consist of lots of documents and keywords, which one needs to reduce to the acceptable size and one should also cope with the sparsely of data. Dimension reduction can be made by Latent Semantic Analysis [1, 4, 5, 8, 14, 15], which uses singular value decomposition of the vector space model matrix. Singular value decomposition (SVD) divides the vector space model matrix into 3 matrices, S, U and V, where S is diagonal matrix of positive singular values U is matrix of left eigenvectors, V is the matrix of right eigenvectors.

\[
X = USV^T
\]  
(1)

By removing the k – smallest singular values and preserving r greatest singular values, the relevance of documents is preserved as precisely as possible and the dimension is greatly reduced by

\[
T = Xr Ur S^r
\]  
(2)

So the r-dimensional representation of the document collection can be obtained by the formula

\[
B = Xr Ur * inv( S^r )
\]  
(3)

and r-dimensional representation of a query is given by the formula

\[
Qr = Qr Ur * inv( S^r )
\]  
(4)

Similarity between document representation and query representation is computed by the cosine similarity measure. SVD has a tight connection to Principal Component Analysis [2, 3, 4, 7, 9]. Principal Component Analysis is a technique that finds eigenvalues and eigenvectors from the covariance matrix. By removing the smallest eigenvalues and corresponding eigenvectors, the most precise dimension reduction will be performed. Due to the high complexity of computing the eigenvalues and eigenvectors (by characteristic equation), sometimes it is useful to replace these algorithms with the neural networks Nowadays, the document clustering methods are emerging in the text document retrieval [7, 9, 10, 11, 16, 17, 18]. Clustering enables user to search for an information in one cluster, the cluster with most similar documents to her query. For document clustering many types of neural networks are used. The most popular are Kohonen neural networks [10, 11, 12], that organize documents into a two-dimensional map, where they create clusters. There are also other types of neural networks for document clustering.

The neural networks used in this paper make the dimension reduction and clustering by principal component analysis in one solution for each neural network [2, 3, 4, 7, 9, 15, 16, 17, 18]
2. AGGLOMERATIVE AND OJA LEARNING RULE

We used Hebbian-type neural network which based on Oja learning rule [7, 9, 16, 17, 18]. This neural network is a two-layer neural network. First layer represents one document as a keyword vector and the second layer represents that document in the reduced document space. The weights of Hebbian-type neural network are at the beginning chosen as a small random variables. From the input variables and weight variables output is calculated by the formula:

\[ y_i(t) = \sum_s w_{sy}(t) \phi(x_i(t)) \]

(5)

So the activation function in the output neurons is linear function. Such a network represents linear Principal Component Analysis. On the base of input, weights and output, weight change in the next time step can be computed.

\[ w_{ij}(t+1) = \beta w_{ij}(t) + \beta(t) y_i(t) x_j(t) \]

(6)

where \( \beta(t) \) is the learning rate usually from the interval \((0, 1)\), in our case, \( \beta(t) \) is set to 0.001. The only one input to this network are documents of vector space model matrix as a keyword vectors. So this network uses unsupervised learning algorithm. After training the input document is represented in the reduced document space at the output. Output of the network has so many neurons, how many dimensions are in the reduced documents space. In this paper the number of output neurons is always two, so the documents are represented as a points in the two-dimensional document space.

3. RELATED WORK

Barry de Ville et. al. [20] proposed the data-mining classification and predictive modelling algorithms that are based on bootstrapping techniques. This describes re-use of source data repeatedly that can render a holographic view of the modelled data. This holographic application is mostly used in industrial area that involves text mining warranty claims at a major international truck, car and heavy equipment manufacturer. This paper shows, how they work, and how they perform in text mining area as supplied to warranty claims. The key goal is to obtain the performances better than human. Mine.T et al., [21] put forward that text mining system obtain the relationship between the topics of international conference. This research paper not only says about the relation between conference and subject, but also says the relationship between information entities that users are interested. arbitrary relations between concepts of a molecular biology ontology for the purpose of supporting text mining and manual ontology building. have given insights on work done on the WWW corpus for text mining based on ontological systems. Fundamentally ontology is defined as specification of a conceptualization & this also refers to the subject of presence, knowledge based framework and JAVA based ontology provides a plug-and-play environment that makes it a flexible for rapid prototyping.

Qiaozhu Mei et al., says new general probabilistic model for contextual text mining that can cover several existing models as superior cases. The extension of the Probabilistic Latent Semantic Analysis (PLSA) model the context variables models the context of a document. The propose mixture model, termed contextual probabilistic latent semantic analysis (CPLSA) model, can be useful to many interesting mining jobs, such as temporal document mining. PLSA [23] document act as a mixture of features, where each feature is represented by a multinomial distribution to evade over fitting in PLSA, Blei and co-authors proposed a propagative aspect model called Latent Dirichlet Allocation (LDA), which could group up the themes from document. Then in this paper we approach the software mining task with a combination of text mining and link analysis technique. This primarily deals with interlinks between one occurrence to another occurrence.

A.Hotoh et al., [24] suggest that in high dimensional space text clustering involves, but it difficult that it appears for all types of setting. This is one of the new approaches for applying background knowledge during pre-processing in order to improve clustering results and allow for selection between results. In order to overcome the striving, we calculate multiple clustering results using k-Means. The outcomes may be distinguished and clarified by the corresponding selection of concepts in the ontology. The problem of clustering high-dimensional data sets has been researched by Agrawal et al. [25] They present a clustering algorithm called CLIQUE that identifies dense clusters in subspace of maximum dimensionality. Hinneburg & Keim [26] show how projections improve the effectiveness and efficiency of the clustering procedure. Their paper shows that projections are very important for improving the performance of clustering algorithms.

Crescenzi and colleagues [27] present a scheme that automatically extract data from large data-intensive Web sites their “data grabber” explores a large Web site and then create a model for it, which describe it as a directed graph with nodes describing classes of structurally similar pages and arcs representing links between these sheets. After the analytic classes of interest, a library of wrappers can be produced, one per class with the help of an external wrapper producer & appropriate data can be extracted.

Pallav Roxy, and Durga Toshniwal, projected several approaches and that can be classified into two main categories, model-based approach and similarity-based approach. Model-based approaches, on the other hand, attempt to learn multiplicative models from the datasets or documents, with each model representing one particular document group. Similarity-based approach is a pair wise similarity document clustering approaches, aiming to maximize the average similarities within clusters and minimize the average similarities between the clusters. Several approaches have been so far proposed for document clustering from mid nisssneties. New technique such as self-organizing map [28], mixture of Gaussians [29], spherical k-mean[30], bi-sector k-means [31], mixture of multinomial [32, 33]. K-means is one of the simplest unsupervised learning methods that solve the well-known clustering difficulty. The key idea is to define k centroids, for every cluster. These centroids should be placed far away from each other. The next phase is to take each point belonging to a given data set and associate it to the nearest centroid, then necessity to re-calculate k new centroids, Next we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been produced. As a result of this loop it is seen that the k centroids change their location step by step until no more changes are done, lastly this algorithm aims at reducing an objective function.

Shady Shehata put forward new view called Vector Space Model (VSM) .Vector Space Model (VSM) [34] which is a
widely used for representing data for text classification and clustering. It says each document as a feature vector of the document. Each feature vector encloses term-weight. Selecting and weighting these features accurately affect the result of the clustering algorithm substantially. Including semantic features from the WordNet lexical database is one of best approaches that have been tried to improve the accuracy of text clustering process. In this paper he also introduced a new model which are based on semantic. The suggested model captures the semantic structure of each term within a sentence rather than the frequency. Each sentence in a document is labelled by a semantic role labeller, it can be either a word or phrase dependent on the semantic structure of the sentence. Based on this analysis, each term which is used in document assign some weight value. The terms that have highest weights value are extracted as top terms. Synonyms of each individual word are added to the term vector. These concepts are analysed on the sentence. Top terms and used in text document clustering. When a new text document is introduced to the system, the projected model can detect a concept match from this document to all the previously processed documents in the data set by scanning the new document and extracting the matching concepts.

Li haiying et al.,[35] spotted about real-time clustering that involves in two stages process, the first stage involves in extraction process that is partial parser and a shallow stemmer are appealed here, it can be used for both linguistic and statistical methods, in order to decrease the term variations within the returned text snippets the system introduce normalization algorithms. This avoids redundancy, for example: "games downloads" and "download games" the second stage is mainly for clustering purpose, that is combining both statistical clustering and linguistic clustering. They generate hierarchical clustering for grouping of related text documents. This is done in real-time without any predefined grouping scheme, pre-processing of all the document collections used by the search engines. The clustering algorithms allow the same document can be appears in multiple clusters. That means different persons usually will group same information differently.

4. PROPOSED APPROACH
The main work is to develop a novel hierarchal algorithm for document clustering which provides maximum efficiency and maximum performance. It is mainly dedicated in studying & making use of cluster coinciding phenomenon to design cluster merging principles. Suggesting a new way to calculate the overlap rate in order to improve time efficiency and the veracity is primarily concentrated. Based on the HCM(Hierarchical Clustering Method), the usage of Expectation-Maximization (EM) algorithm in the Gaussian Mixture Model to count the parameters and make the two sub-clusters combined when their overlap is the largest is narrated.

Experiments in both two cluster public data and document clustering data show that this approach can improve the efficiency of clustering and save computing time.

4.1 Pre-Processing
We need to convert the text documents into vectors before training since the input to SOM need to be vector-form. In this step common procedures for processing text documents, such as word segmentation, stop word elimination, and stemming, were first applied to obtain a set of keywords that can describe the contents of a document. All keywords were collected into the vocabulary of the corpse. A document is then transformed into a vector according to the keywords it contains. Let document.

\[
d_j = \{k_{ij} | 1 \leq i \leq |\tilde{V}|, 1 \leq j \leq N\},
\]

where \(N\) is the number of documents, \(n_j\) is the number of distinct keywords in \(d_j\), and \(k_{ij}\) represents the \(i\)th keyword in \(d_j\). The vocabulary, denoted by \(V\), is just the union of all \(d_j\), i.e.

\[
V = \{v_i | 1 \leq i \leq |V|\},
\]

where \(v_i\) is the \(i\)-th word in \(V\). A document is encoded into a binary vector of length \(|V|\) according to those keywords that occur in it. When a keyword \(v_i\) occurs in this document, the \(i\)th element of the vector will have value 1; otherwise, the component will have value 0. With this scheme, a document \(d_j\) will be encoded into a binary vector

\[
d_j = \{\delta_{ij} | 1 \leq i \leq |V|\}.
\]

4.2 HAHNN ALGORITHMS
For \(n\) samples, agglomerative algorithms begin with \(n\) clusters and each cluster contains a single sample or a point. Then two clusters will merge so that the similarity between them is the closest until the number of clusters becomes 1 or as specified by the user.

1. Start with \(n\) clusters, and a single sample indicates one cluster.
2. Find the most similar clusters \(C_i\) and \(C_j\) then merge them into one cluster.
3. Repeat step 2 until the number of cluster becomes one or as specified by the user.

The distances between each pair of clusters are computed to choose two clusters that have more chance to merge. There are a number of ways to calculate the distances between the clusters \(C_i\) and \(C_j\) as show in table 4.2

<table>
<thead>
<tr>
<th>Table 4.2 different methods to calculate distance between two cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>Single linkage</td>
</tr>
<tr>
<td>Complete linkage</td>
</tr>
<tr>
<td>Average linkage</td>
</tr>
</tbody>
</table>
5. EVALUATION OF CLUSTERING

Evaluation of clustering results sometimes is referred to as cluster validation. There have been several suggestions for a measure of similarity between two clustering’s. Such a measure can be used to compare how well different data clustering algorithms perform on a set of datasets or data. These measures are generally tied to the type of criterion being considered in assessing the quality of a clustering method evaluation. When a clustering result is evaluated based on the data that was clustered itself, this is called IE(internal evaluation).

These methods usually assign the best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters. External evaluation external evaluation, clustering results are calculated based on data objects that was not used for clustering, such as known external standards and class labels. Such standards consist of a set of pre-classified items, and these sets are often produced by human experts. Thus, the standard sets can be thought of as a gold standard for evaluation. These types of evaluation approaches measure how close the clustering is to the predetermined standard classes.

5.1 Purity of Clusters

Purity is a simple and transparent evaluation measure. The F measure in addition supports differential weighting of these two types of errors. To compute assigned to the class which is most frequent in the document cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N.

Formally

Purity(Ω,C)=1/N \sum_{i=1}^{q} \max_j | w_i \cap c_j |

5.2 F-Measure

The F-measure function can be used to balance the contribution of false negatives by weighting recall through a parameter \( \beta \). Let recall and precision be defined as follows:

\[
R = \frac{TP}{TP + FN} \\
P = \frac{TP}{TP + FP}
\]

Where P the precision rate is rate and R is the recall rate.

We can calculate the F following formula

\[
F_\beta = \frac{(\beta^2+1) \cdot P \cdot R}{\beta^2 \cdot P + R}
\]

Notice that when \( \beta=0 \), \( F_0 = P \). Otherwise, recall has no impact on the F-measure when \( \beta=0 \) and increasing \( \beta \) allocates an increasing amount of weight to recall in the final F-measure.

5.3 Entropy

entropy measure that how the several classes of documents are spread within each cluster. Given a particular cluster \( S_r \) of size \( n_r \), the entropy of this type of cluster is defined as

\[
E(S_r) = -\frac{1}{\log q} \sum_{i=0}^{q} \frac{n_r}{n_r} \log \frac{n_i}{n_r}
\]

where \( q \) is the number of classes in the dataset, and \( n_r \) is the number of documents of the \( i \)th class that were assigned to the \( r \)th cluster. The entropy of the complete clustering solution is then defined to be the sum of the specific cluster entropies weighted according to the cluster size. That is,

\[
\text{Entropy} = \sum_{r=1}^{k} \frac{n_r}{n} E(S_r)
\]

6. EXPERIMENTS AND RESULT

In our experiments, we used a total of three different datasets. The smallest of these datasets contained 1504 documents and the largest contained 9558 documents. To ensure diversity in the datasets, we obtained them from different sources. For all data sets, we used a stop-list to remove common words, and the words were stemmed with the help of Porter’s suffix-stripping algorithm [36]. Moreover, any term that occurs in fewer than two documents was eliminated.

Clusters in compact document space are placed approximately in the same place but the scale of the coordinates almost always differs. The distances between clusters of Hebbian neural network are smaller than distances between clusters of the auto associative neural network because of the different learning rate. In the Hebbian-type neural network the learning rate is equal 0.001, and in the auto associative neural network the learning rate is equal 0.01.

Table 6.1 show the result of different clustering methods with their entropy and purity percentages. Figure 6.1 show the graphical representation of result which is.

<table>
<thead>
<tr>
<th>Method</th>
<th>Entropy</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeated bisection method</td>
<td>0.191</td>
<td>0.831</td>
</tr>
<tr>
<td>Directed method</td>
<td>0.217</td>
<td>0.803</td>
</tr>
<tr>
<td>HAHNN</td>
<td>0.249</td>
<td>0.839</td>
</tr>
</tbody>
</table>

7. CONCLUSION

The aim of this paper was to compare HAHNN (hierarchical agglomerative and Hebbian-type neural network) and Repeated bisection and direct method of clustering from the point of dimension reduction of document space and clustering. It was shown that they give very similar results. They both divided collection into four clusters with the same documents in one cluster. The differences in the output clusters can be caused by different learning rules in the compared neural networks and by the random initial weights set on the beginning of training phase. It can be seen, that both neural networks represented dimension reduction of document space by PCA technique, what was expected from them.
8. REFERENCES:


