International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-1 Issue-9 March-2013

Credential Clustering in Parallel Comparability Frequency Amplitude

Ameela.T[1], Kaleeswaran.D[2]

Abstract

Cluster analysis or clustering is the process of assigning a set of objects into groups such that the objects in the same cluster are more similar to each other and objects in other clusters are dissimilar. In the existing system, Locality preserving indexing method is used. Euclidean distance is used as the method to find the distance between the documents. This method also takes use of weighted function to find the distance between documents. Major difficulty is found in analyzing the weighted function. Since the documents can take many forms, the structure of the document space depends on the similarities between the documents. A new approach based on Correlation as a similarity measure is best proposed for analyzing the structure of documents arranged in the high-dimensional document space.

Keywords

Correlation similarity, Credential Clustering, Similarity measure, Semantic subspace.

1. Introduction

Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. In high-dimensional document space, the semantic structure is usually implicit. It is desirable to find a low-dimensional semantic subspace in which the semantic structure can become clear. Hence, discovering the intrinsic structure of the document space is often a primary concern of document clustering. Since the manifold structure is often embedded in the similarities between the documents, correlation as a similarity measure is suitable for capturing. The Survey System's optional Statistics Module includes the most common type, called the Pearson or product-moment correlation. The larger the value of *Corr(u,v)* stronger the association between the two vectors *u* and *v*. Online document clustering aims to group documents into clusters, which belongs to unsupervised learning. However, it can be transformed into semi-supervised learning by using the following side information: If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster and if two documents are far away from each other in the original document space, they tend to be grouped into different clusters. Mathematically the correlation between two column vectors *u* and *v* can be calculated as,

$$Corr(u,v) = \frac{u^T v}{\sqrt{u^T u}\sqrt{v^T v}} = \langle \frac{u}{\|u\|}, \frac{v}{\|v\|} \rangle \qquad (1)$$

In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the Euclidean norm**.** Older literature refers to the metric as Pythagorean metric. The Euclidean distance between point's **p** and **q** is the length of the line segment connecting them (**PQ**).Correlation can be used as a technique for clustering the documents. The main result of a correlation is called the correlation coefficient (or "r"). It ranges from -1.0 to +1.0. The results are summarized as follows:

1. The closer r is to +1 or -1, the more closely the two variables are related.
2. If r is close to 0, it means there is no relationship between the variables.
3. If r is positive, it means that as one variable gets larger, the other gets larger.
4. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation).

While correlation coefficients are normally reported as r = (a value between -1 and +1), squaring them makes then easier to understand. The square of the

Ameela.T, Department of Computer Science and Engineering, St. Michael College of Engineering and Technology, Sivagangai, India.
Kaleeswaran.D, Department of Computer Science and Engineering, St. Michael College of Engineering and Technology, Sivagangai, India.

coefficient (or r square) is equal to the percent of the variation in one variable that is related to the variation in the other. After squaring r, ignore the decimal point. An r of .5 means 25% of the variation is related (.5 squared =.25). An r value of .7 means 49% of the variance is related (.7 squared = .49). Correlations reports sometimes may bring random sampling error. If we are working with small sample sizes, choose a report format that includes the significance level.

2. Related Work

The **k-means method** is one of the methods that use the euclidean distance, which minimizes the sum of the squared euclidean distance between the data points and their corresponding cluster centers. In K-Means method Computation complexity is high. **Latent semantic indexing (LSI)** is one of the effective spectral clustering methods, aimed at finding the best subspace approximation to the original document space by minimizing the global reconstruction error (Euclidean distance). In LSI the high dimensionality of the document space, a certain representation of documents usually resides on a nonlinear manifold embedded in the similarities between the data points. Unfortunately, the euclidean distance is a dissimilarity measure which describes the dissimilarities rather than similarities between the documents. Thus, it is not able to effectively capture the nonlinear manifold structure embedded in the similarities between them. **Locality preserving indexing (LPI)** method is a different spectral clustering method based on graph partitioning theory. The LPI method applies a weighted function to each pair wise distance attempting to focus on capturing the similarity structure, rather than the dissimilarity structure, of the documents. LPI method does not overcome the essential limitation of euclidean distance. Furthermore, the selection of the weighted functions is often a difficult task. The usage of correlation as a similarity measure can be found in the **canonical correlation analysis (CCA) method**. The CCA method is to find projections for paired data sets such that the correlations between their low-dimensional representatives in the projected spaces are mutually maximized.

A. The K- means method works with numeric data only. Pick a number (K) of cluster centers (at random).Assign every item to its nearest cluster center (e.g. using Euclidean distance).Move each cluster center to the mean of its assigned items. Repeat steps 2, 3 until convergence (change in cluster assignments less than a threshold)

```
Algorithm:
Initialize k cluster centers
Do
Assignment step: Assign each data point to its closest cluster center
Re-estimation step: Re-compute cluster centers
While (there are still changes in the cluster centers)
```

B. LPI aims to cluster the documents into different semantic classes. The document space is generally of high dimensionality and clustering in such a high dimensional space is often infeasible due to the curse of dimensionality. By using locality preserving indexing (LPI), the documents can be projected into a lower-dimensional semantic space in which the documents related to the same semantics are close to each other. Different from previous document clustering methods based on latent semantic indexing (LSI) or nonnegative matrix factorization (NMF), this method tries to discover both the geometric and discriminating structures of the document space. Theoretical analysis of this method shows that LPI is an unsupervised approximation of the supervised linear discriminate analysis (LDA) method, which gives the intuitive motivation of LPI method.

C. Kernel Canonical Correlation Analysis is used to learn a semantic representation to web images and their associated text. The semantic space provides a common representation and enables a comparison between the text and images. In the experiments we look at two approaches of retrieving images based only on their content from a text query. We compare the approaches against a standard cross-representation retrieval technique known as the Generalized Vector Space Model.

D. Non-negative matrix factorization is a document clustering method based on the term document matrix of the given document corpus. In the latent semantic space derived by the non-negative matrix factorization (NMF), each axis captures the base topic of a particular document cluster, and each of the documents is represented as an additive combination of the base topics. The cluster membership of each document can be easily determined by finding the base topic (the axis) with which the document has the largest projection value. A good clustering method must have high inter-cluster distances and low intra-cluster distances.

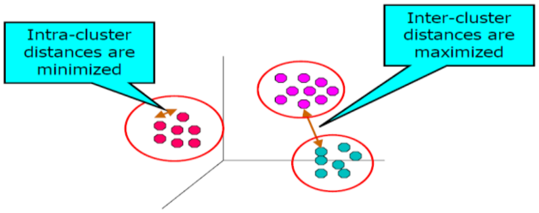
Intra-cluster distances are minimized
Inter-cluster distances are maximized

Figure 2.1: Example of good clustering

3. Proposed Approach

Credential Clustering based on correlation preserving Indexing:

All documents are projected onto the unit hypersphere (circle for 2D). The global angles between the points in the local neighbors, β_i, are minimized and the global angles between the points outside the local neighbors, α_j are maximized simultaneously, as illustrated in Fig. 3.1. On the unit hypersphere global angle can be measured by spherical arc, that is, the geodesic distance. The geodesic distance between z and $z^{'}$ on the unit hypersphere can be expressed as,

Geodesic distance,

$$d_G = \arccos(z^T z^{'}) = \arccos(Corr(z, z^{'})) \qquad (2)$$

Since a strong correlation between z and $z^{'}$ means a small geodesic distance between z and $z^{'}$, then CPI is equivalent to simultaneously minimizing the geodesic distances between the points in the local patches and maximizing the geodesic distances between the points outside these patches. The geodesic distance is superior to traditional euclidean distance in capturing Latent manifold. CPI can effectively capture the intrinsic structures embedded in the high-dimensional document space. Semi-supervised learning using the nearest neighbors graph approach in the euclidean distance space was used by LPI. CPI is a semi-supervised learning using nearest neighbors graph approach in the correlation measure space. Euclidean distance is not appropriate for clustering high dimensional normalized data such as text and a better metric for text clustering is the cosine similarity. Correlation might be a suitable distance measure for capturing the intrinsic structure embedded in document space. That is why the proposed CPI method is expected to outperform the LPI method. In parallel comparability, there is no need of algorithms for clustering. But the technique used is the correlation. The clustering based on correlation is more accurate than other algorithms.

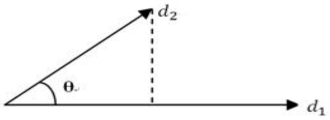

Figure 3.1: Angle between documents

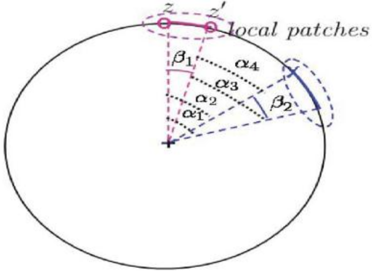

Figure 3.2: 2D Projections of CPI

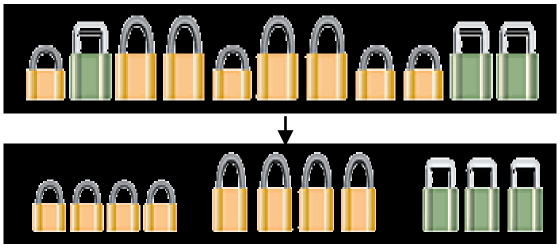

Figure 3.3: Accuracy of clusters obtained from correlation

Algorithm Derivation:

Clustering is one of the fundamental operations in data mining for grouping objects into classes. This can be done for a set of documents by:

1. A set of documents $x_{1,}$ x_2 , . . . , *xn € IRn, X* denote the document matrix.
2. Construct the local neighbor patch, and compute the matrices *MS* and *MT*.
3. The singular value decomposition of *X* can be written as *X = U∑V t.*
4. Thus the document vectors in the SVD subspace can be obtained by *X = UTX* and Compute CPI Projection.
5. Let *WCPI* be the solution of the generalized eigenvalue problem *MSW = MW.* Cluster the documents in the CPI semantic subspace.

Document Representation:

In all experiments, each document is represented as a term frequency vector. The term frequency vector can be computed as follows:

1. Transform the documents to a list of terms after words stemming operations.
2. Remove stop words. Stop words are common words that contain no semantic content.
3. Compute the term frequency vector using the TF/IDF weighting scheme. The TF/IDF weighting scheme assigned to the term t_i in document d_j is given by

$$(tf\ /\ idf\,)_{i,j} = tf_{i,j} \times idf_i \qquad (3)$$

Here, $$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (4)$$

is the term frequency of the term ti in document d_j, where $n_{i,j}$ is the number of occurrences of the considered term t_i in document d_j.

$$idf_i = \log(\frac{|D|}{|d : t_i \in d|} \qquad (5)$$

is the inverse document frequency which is a measure of the general importance of the term t_i, where

$|D|$ - The total number of documents in the test set and

$|d : t_i \in d|$ - The number of documents in which the term t_i appears.

Let $v = \{t_1, t_2 \ldots t_m\}$ be the list of terms after the stop words removal and words stemming operations. The term frequency vector x_j of document d_j is defined as

$$x_{i,j} = (\frac{tf}{idf})_{i,j} \qquad (6)$$

Using n documents from the corpus, we construct an $m \times n$ Term-document matrix X. The above process can be completed by using the text to matrix generator (TMG) code. The Credential clustering can be carried out in five steps:

1. Pre-processing Documents.
2. Frequency Measure.
3. Measuring Similarity and Dissimilarity.
4. Classification of Documents into Clusters.

The following system architecture explains the overall steps and operations performed in the proposed approach.

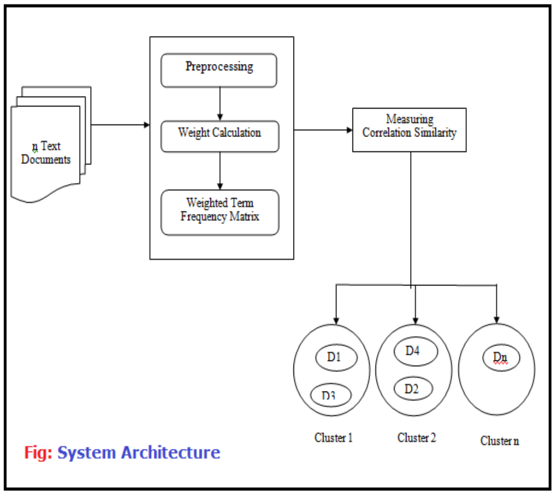

Figure 3.4: System Architecture

Pre-processing Documents:

A. Removing Stopwords: These are function words and connectives. Appear in a large number of documents and have little use in describing the characteristics of documents. Example: “of”, “a”, “by”, “and”, “the”, “instead”.

B. Stemming: Stem is the portion of a word which is left after the removal of its affixes (i.e., prefixes and suffixes).Stemming reduces variants of the same root word to a common concept. Reduce the size of the indexing structure. Techniques for stemming include: Morphological analysis (e.g., Porter’s algorithm) and Dictionary lookup (e.g., WorldNet).

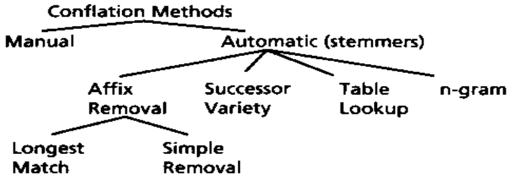

Figure 3.5: Stemming Strategies

Table Lookup: Simply looking for the stem of a word. Dependent on data on stems for the whole language and requiring considerable storage space

Successor variety: Determining of morpheme boundaries using the knowledge of linguistics

N-grams: Identification of digrams and trigrams. More a term clustering procedure than a stemming one

Affix Removal: Uses Porter algorithm which uses a suffix list for suffix striping. Apply a series of rules to the suffixes of the words.

Table 3.1: Stemming Results

Before Stemming	After Stemming
Extraction	Extract
Accounting	Account
partitioning	Partition
Possibleness	Possible
categorisation	Categorization
Adjustment	Adjust
Opening	Open
Segmentation	Segment
Prediction	Predict
Awareness	Aware

Frequency Measure:
After converting the documents in to list of terms we need to find the weight for each term in the documents to compute term frequency vector. The term frequency vector is computed by using TF/IDF weighting scheme. The TF/IDF is calculated by using the formulae,

$$(tf\ /\ idf\ )_{i,j} = tf_{i,j} \times idf_i \qquad (7)$$

To calculate TF/IDF first we find the frequency of term t_i in document d_j and divide the frequency by total no of terms in that document. Then we find inverse document frequency (idf) by finding the documents which contains the term t_i and divide the idf by total no of documents. By multiplying both *tf * idf* we get the weight for term t_i. By repeating the above process we get weight for each term in all the *n* documents. By using the weight we compute weighted term frequency vector matrix. After computing the weighted term frequency vector matrix we measure the similarity between documents by using *corr (u, v) = (u/|u|, v/|v|).*

Measuring Similarity and Dissimilarity:
As a first step, we introduce the overlap score measure: the score of a document d is the sum, over all query terms, of the number of times each of the query terms occurs in d.

$$Score(q,d) = \sum_{t \in d} tf - idf_{t,d} \qquad (8)$$

After computing the weighted term frequency vector matrix we measure the similarity between documents. By comparing the document d with all the other documents we get the similar terms of document d. By repeating the similarity measure we get the similarities between all the documents. Then we calculate the dissimilarity between n documents by using Euclidean distance. By repeating the same process we find the dissimilar documents in order to group the documents together.

Classification of Documents into Clusters:
By using the correlation similarity we cluster the similar documents together in order to find the intrinsic structure of the document space. First we take distinct similarity value from all the values and create a cluster for each distinct value. First we take the maximum value and compare with the other values and group the documents and form cluster 1 with documents with maximum value into one and then take the second maximum value and compare with the other documents and form cluster 2 with documents having the same value and so on. In the same way we cluster the documents by using the Euclidean distance in order to detect intrinsic structure of the document space.

4. Experimental Result

The performance of the proposed method is found by experiments and comparing with the existing method. A set of text documents are selected and the cluster analysis is performed. Initially the preprocessing and the frequency measure steps are completed and then result of the proposed method is found. It was then compared with the results obtained from the existing method with the same set of text documents. Then accuracy between the results from two methods is found. It was found that the proposed method serves better than the existing methods. Parameter selection plays an important role in finding the results. The generalization error of the proposed method is also very low from the existing method.

Table 4.1: Accuracy of Clusters

Cluster number	Accuracy (%)			
	k-mean	LSI	LPI	CPI
2	64.25± 10.2	67.87± 12.2	71.12± 12.7	75.06± 13.3
3	52.81± 8.59	54.53± 10.1	58.65± 9.92	60.84± 9.52
4	46.61± 8.95	47.96± 8.32	52.57± 7.86	56.20± 8.88
5	41.08± 7.51	42.72± 6.93	44.45± 7.62	47.45± 7.02
Average	44.06± 7.62	46.32± 7.54	49.29± 7.81	52.21± 7.90

5. Conclusion

In this paper we have discussed the problem of clustering documents in Correlation similarity measure space. In this we maximize the correlation between the documents in local patches and simultaneously minimize the correlation between the documents outside these patches. In this we have projected the document in low- dimensional semantic subspace where the manifold structure of the original document space is retained. During retrieval, add other documents in the same cluster as the initial retrieved documents to improve recall. Automated production of hierarchical taxonomies of documents for browsing purposes .The major challenges of clustering are dealing with outliers and working with different type of attributes.

6. Future Work

A divide-and-merge method is used clustering the documents. It uses top-down “divide” phase with a bottom-up “merge “ phase for clustering. A tree is built by the elements of the set in divide phase. The Optimal partition that respects the tree can be found quickly in merge phase. Meta search engine is used to cluster the result of web searches.

References

[1] D.Cai, X.He, and J.Han, “Document Clustering Using Locality Preserving Indexing”, IEEE Trans.Knowledge and Data Engg. vol.17, no.12, 1624-1637, Dec-2005.

[2] D.Cheng, R.Kannan, S.Vempala, and G.Wang, “A-Divide-and-Merge Methodology for Clustering,” ACM trans.Database Systems, vol.31, no.4, pp.1499-1525, 2006**.**

[3] Y.Fu, S.Yan, and T.S.Huang, “Correlation Metric for Generalized Feature Extraction Pattern Analysis and Machine Intelligence,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.30, no.12, pp.2229-2235, Dec-2008.

[4] D.R.Hardoon, S.R.Szedmak, and J.R.Shawe-taylor, “Canonical correlation analysis; An overview with application to learning methods,” J.Neural Computation, vol.16, no.12, pp.2639-2664, 2004.

[5] G.Lebanon, “Metric Learning for Text Documents,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.28, no.4, pp.497-507, Apr-2006.

[6] Y.Ma, S.Lao, E.Takikawa, and M.Kawade, “Discriminant Analysis in Correlation Similarity Measure Space,” Proc.24th Int’l Conf. Machine Learning (ICML ’07) pp.577-584, 2007.

[7] W.Xu, X.Liu, and Y.Gong, “Document Clustering Based on Non-negative Matrix Factorization,”Proc.26th Ann.Int’l ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR-’03), pp.267-273, 2003.

[8] D.Zeimpekis and E.Gallopoulos, “Design of a Mat lab toolbox for term-document matrix generation,”proc. Workshop Clustering High Dimensional Data and its Applications at the Fifth SIAM Int’l Conf.data Mining (SDM ’05) pp 38-48, 2005**.**

[9] S.Zhong and J.Ghosh, “Generative Model-based Document Clustering: A Comparative Study,” Knowledge of Information System, vol .8, no.3, pp.374-384, 2005.

[10] X.Zhu “Semi Supervised learning using Gaussian Fields and Harmonic Functions,” Technical report Computer Sciences Univ. of Wisconsin-Madison, 2005.

T.Ameela is currently a PG scholar in Department of Computer Science and Engineering at St.Michael College of Engineering and Technology, Kalayarkoil, Sivagangai. She received her Bachelor Degree in Information Technology from Sakthi Engineering College, Chennai, in 2011. Her research interests include data mining and data warehousing, Semantic web and Network Security.

D.Kaleeswaran is currently working as Assistant Professor in the Department of Computer Science and Engineering, St.Michael College of Engineering and Technology, Kalayarkoil, Sivagangai. He received his Bachelor degree in Information Technology from Sowdambiga Engineering College in 2010 and completed his Post Graduate in Computer Science and Engineering from St.Michael College of Engineering and Technology, Kalayarkoil in 2012. His Research areas include data mining and data warehousing, grid computing, cloud computing and wireless sensor network security.