Document Clustering using Correlation Preserving Indexing with Concept Analysis

M. Mohanasundari, P. Malathi
Dept. of CSE, Velalar College of Engineering and Technology

Abstract
Clustering is one of the most important techniques in which the machine learning and data mining tasks. Similar data grouping is performed using clustering techniques. Hierarchical clustering model produces tree structured results. Partitioned clustering produces results in grid format. The documents are projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters. The Euclidean distance is a dissimilarity measure describes the dissimilarities between the documents. Correlation indicates the strength and direction of a linear relationship between two random variables. A scale-invariant association measure is used to calculate the similarity between two vectors. Correlation preserving index (CPI) based clustering is used for document clustering process. The similarity-measure-based CPI method is used for detecting the intrinsic structure between nearby documents. In CPI method the documents are projected into a low-dimensional semantic space. Correlations between the documents in the local patches are maximized. Correlations between the documents outside these patches are minimized simultaneously. The spectral clustering is applied on the correlation similarity model with nearest neighbor learning process. The Ontology repository is used to manage the term concept relations. Local patch extraction is carried out with Ontology support. Term frequency based weight is replaced with concept weight based model. The document preprocess operations are carried out to extract term information. Stop word elimination and stemming process are applied on the term collection. Porter stemming algorithm is used for suffix analysis. Ontology is used to extract term relationships.

Keywords
Correlation Preserving Index, Ontology, Neighbor Learning Process, Correlation Similarity

I. Introduction
The structured representation of a text document presents its contents as a hierarchy of text blocks, tables, frames, and other related objects. It provides a logical structure to the document and describe how their contents will be displayed. Generally, tables and frames are used to group other structures while text blocks contain the actual textual information. New elements are created and inserted into the document programmatically with a QTextCursor or QTextEdit. Elements can be given a particular format when they are created; otherwise they take the cursor’s current format for the element. The structural elements used in a rich text document, highlights their uses and features, and tells how to examine their contents. Document is explained in the QTextCursor Interface editing.

Document clustering or Text clustering is an automatic document organization, extraction of topic and immediate information retrieval or filtering. It is similar to data clustering. A web search engine often returns thousands of pages in response to a big query, making it hard for users to browse or to identify similar information. Clustering methods may be used to group the retrieved documents automatically into a list of meaningful categories, and it is achieved by Enterprise Search engines such as Vivisimo and Northern Light, consumer search engines such as Heliod and PolyMeta, or open source software such as Carrot2. FirstGov.gov, the official Web portal for the government of U.S, utilize document clustering to automatically organize its search results into categories. For example, if a user submits “immigration”, next to their set of results they will see categories for “Immigration Reform”, “Employment” and more. Probabilistic Latent Semantic Analysis (PLSA) can also be conducted to perform document clustering.

Document clustering involves the use of descriptors and their extraction. Descriptor is the sets of words that describe the content in the document within the cluster. It is generally considered to be a centralized process. Examples includes web document clustering for search users. Document clustering can be categorized to two types, offline and online. Online applications are usually constrained by efficiency problems when compared offline applications.

Generally, there are some common algorithms, it is hierarchical based algorithm, which includes complete linkage, group average, single link and Ward’s method. By dividing, documents can be clustered into hierarchical structure, which is suitable for browse. However, some algorithm usually suffers from efficiency problems. Usually, it provide greater efficiency but less accurate than the hierarchical algorithm. Other algorithm involves ontology supported clustering, graph based clustering and order sensitive clustering.

II. Problem Statement
Document clustering aims to which automatically group related documents into clusters. It is one of the important tasks in machine learning and artificial intelligence and has received much attention in recent years [3]. Based on various distance measures, a number of methods for document clustering [6, 9-10]. Low computation cost is achieved in spectral clustering methods, in which the documents are projected into a low-dimensional semantic space and then a traditional clustering algorithm is used to find the clustered documents. Latent semantic Indexing (LSI) is one of the spectral clustering methods, aims to finds the best subspace approximation to the original document space by minimizes the global reconstruction error.

However, because of 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 [1]. Unfortunately, the Euclidean distance is a dissimilarity measure which describes the dissimilarities rather than similarities between the documents. It is not effectively capture the nonlinear manifold structure embedded in the similarities between them [16]. An effective document clustering method must be able to find a low-dimensional representation of the documents that can best preserve the similarities between the data points. Locality preserving indexing (LPI) method is a different spectral clustering method based on graph partitioning theory. This method applies a weighted function to each pairwise distance attempting to focus on capturing the similarity structure, other than the dissimilarity.
structure, of the documents. It does not overcome the essential limitation of euclidean distance. The selection of the weighted functions is often a difficult task.

In recent years, some studies [13-4] suggest that correlation as a similarity measure can capture the intrinsic structure embedded in high-dimensional data, especially when the input data is sparse. In statistics and probability theory, correlation indicates the strength and direction of a linear relationship between two random variables which reveals the nature of data represented by the classical geometric concept of an “angle”. It is a scale-invariant association measure usually used to calculate the similarity between two vectors. In many cases, correlation can effectively represent the distributional structure of the input data which conventional euclidean distance cannot explain.

In this paper, we propose a new document clustering method based on correlation preserving indexing (CPI), which explicitly manifold structure embedded in the similarities between the documents. It aims to which an optimal semantic subspace by simultaneously minimizing the correlations between the documents outside the patches maximizing the correlations between the documents in the local patches is performed. It differs from LSI and LPI, which are based on a dissimilarity measure and are focused on detecting the intrinsic structure across widely separated documents rather than on detecting the intrinsic structure between nearer documents. The similarity-measure-based CPI method is used to detect the intrinsic structure between nearby documents rather than on detecting the intrinsic structure between the documents. The intrinsic semantic structure of the document space is often embedded in the similarity between the documents, CPI can detect the intrinsic semantic structure of the high-dimensional document space. It is similar to Latent Dirichlet Allocation (LDA) [7] which attempts to capture significant intradocument statistical structure via the mixture distribution model.

III. Ontologies

An ontology is a “a specification of a conceptualization” whereby a conceptualization is a collection of objects, concepts and other entities that are presumed to exist in some domain and that are tied together with some relationships. A conceptualization is a simplified view of the world, the way of thinking about some specified domains.

Ontologies belong to the knowledge representation approaches that have been discussed above and they aim to provide a shared understanding of a domain both for the computers and for the humans. Thereby, ontology describes a domain of interest in such a formal way that computers can process it. The outcome is that the computer system knows about this domain. Ontology is a formal classification schema, which has a hierarchical order and which is related to some domain. An Ontology comprises the logical component of a “Knowledge Base”. The knowledge base consists of an Ontology, some data and also an inference mechanism. Ontology, comprising the logical component of the knowledge base, defines rules that formally describe how the field of interest looks like. The data can be any data related to this field of interest that is extracted from various resources such as databases, document collections, the Web etc. The inference mechanism would deploy rules in form of axioms, restrictions, logical consequences and other various methods based on the formal definition in the ontology over the actual data to produce more information out of the existing one.

Resource Description Framework (RDF) is a specification proposed by the World Wide Web Consortium (W3C) for describing and then interchanging semantic metadata. The basic element of RDF is a statement, each consisting of a subject, an object and a predicate. A triplet of the form <subject, predicate, and object> is used to express an RDF statement. This statement can be defined as “the sub has an attribute whose value is given by the obj” or “the sub has a relation with the obj”. Based on RDF, RDF Schema (RDFS) is further proposed to define RDF vocabularies for constructing RDF statements. In this study, the system uses RDF and RDFS to encode concepts and semantic relations which are extracted from textual Web content.

IV. Correlation Similarity Analysis

The usage of correlation as a similarity measure can be found in the Canonical Correlation Analysis (CCA) method [11]. 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. Specifically, given a paired data set consisting of matrices X = \{x_1, x_2, \ldots, x_n\} and Y = \{y_1, y_2, \ldots, y_n\}, we would like to find directions wx for X and wy for Y that maximize the correlation between the projections of X on wx and the projections of Y on wy. This can be expressed as

\[ \text{Max}_{w_x, w_y} \frac{\langle X w_x, Y w_y \rangle}{\|X w_x\| \cdot \|Y w_y\|} \]  

Where \(\langle \cdot, \cdot \rangle\) and \(\| \cdot \|\) denote the operators of inner product and norm, respectively. As a result, the powerful statistical technique, the CCA method has been applied in the field of pattern recognition and machine learning [5]. Rather than finding a projection of one set of data, CCA finds projections for two sets of corresponding data X and Y into a single latent space that projects the corresponding points in the two data sets to be as nearby as possible. In the application of document clustering, while the document matrix X is available, the cluster label (Y) is not. So the CCA method cannot be directly used for clustering.

In our paper, a new document clustering method based on Correlation Preserving Indexing (CPI), that explicitly considers the manifold structure embedded in the similarities between the documents. The main aim to find an optimal semantic subspace by simultaneously maximize the correlation between the documents in the local patches and the documents outside these patches. The concept is different from LPI and LSI, which are based on dissimilarity measure and focused on detecting the intrinsic structure. The similarity based CPI method focuses on detecting the intrinsic structure between nearby documents rather than detecting the intrinsic structure between separated documents. The intrinsic semantic structure of the document space is often embedded in the similarities of the documents, it can effectively detect the intrinsic semantic structure of the high-dimensional document space. Here, it is similar to Latent Dirichlet Allocation (LDA) which attempts to capture intradocument statistical structure via the mixture distribution model.

V. Correlation Similarity based Document Clustering

In high-dimensional document space, the semantic structure is implicit usually. It is used to find a low-dimensional semantic subspace in which the semantic structure become clear. Here, the intrinsic structure of the document space can be discovered, often a primary concern of document clustering. Since the manifold structure is embedded in the similarities between the documents, correlation is a similarity measure it is mainly suitable for capturing
the manifold structure embedded in the high-dimensional document space. Mathematically, the correlation between two vectors \( u \) and \( v \) is defined as

\[
\text{Corr}(u, v) = \frac{\langle u, v \rangle}{\| u \| \| v \|} = \left( \frac{\| u \|}{\| v \|} \right) \langle u, v \rangle
\]  

(2)

Note that the correlation corresponds to an angle \( \theta \) such that \( \cos \theta = \text{Corr}(u, v) \). The larger the value of \( \text{Corr}(u, v) \), the stronger the association between the two vectors \( u \) and \( v \).

However, it can be transformed into semi-supervised learning by using the following side information:

If two documents are very close to each other in the original document space, and then grouped into the same cluster \([8]\).

A2. If two documents are far away from each other in the original document space, and then grouped into different clusters.

Based on these assumptions, we propose a spectral clustering in the correlation similarity measure space through the nearest neighbors learning graph process.

**A. Clustering Criterion Function**

Suppose \( y_i \in Y \) is the low-dimensional representation of the \( i \)th document \( x_i \in X \) in the semantic subspace, where \( i = 1, 2, \ldots, n \). Then the above assumptions (A1) and (A2) can be expressed as

\[
\begin{align*}
\text{max} \sum_{i,j} \sum_{x_i \in N(x_i)} \text{Corr}(y_i, y_j) \\
\text{min} \sum_{i,j} \sum_{x_i \in N(x_i)} \text{Corr}(y_i, y_j)
\end{align*}
\]

(3)

and

\[
\begin{align*}
\text{max} \sum_{i,j} \sum_{x_i \in N(x_i)} \text{Corr}(y_i, y_j) \\
\text{min} \sum_{i,j} \sum_{x_i \in N(x_i)} \text{Corr}(y_i, y_j)
\end{align*}
\]

(4)

respectively, where \( N(x_i) \) denotes the set of nearest neighbors of \( x_i \). The optimization of (3) and (4) is equivalent to the following metric learning:

\[
d(x, y) = \alpha \cdot \cos(x, y),
\]

where \( d(x, y) \) denotes the similarity between the documents \( x \) and \( y \), \( \alpha \) corresponds to whether \( x \) and \( y \) are the nearest neighbors of each other.

The maximization problem (3) is an attempt to ensure that if \( x_i \) and \( x_j \) are close, then \( y_i \) and \( y_j \) are also close as well. Similarly, the minimization problem (4) is an attempt to ensure that if \( x_i \) and \( x_j \) are far away, \( y_i \) and \( y_j \) are also far away. Since the following equality is always true

\[
\sum_{i,j} \sum_{x_i \in N(x_i)} \text{Corr}(y_i, y_j) = \sum_{i,j} \sum_{x_i \in N(x_i)} \text{Corr}(x_i, x_j)
\]

(5)

The simultaneous optimization of (3) and (4) can be achieved by maximizing the following objective function

\[
\sum_{i,j} \sum_{x_i \in N(x_i)} \text{Corr}(y_i, y_j)
\]

(6)

Without loss of generality, we denote the mapping between the original document space and the low-dimensional semantic subspace by \( W \), that is, \( Wx_i = y_i \). Following some algebraic manipulations, we have where \( \text{tr}(.) \) is the trace operator.

\[
\sum_{i,j} \sum_{x_i \in N(x_i)} \frac{\langle y_i, y_j \rangle}{\sqrt{\| y_i \| \| y_j \|}} = \frac{\text{tr}(W^T X Y)}{\text{tr}(W^T W)}
\]

(7)

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 the latent manifold [12]. Based on this conclusion, CPI can effectively capture the intrinsic structures embedded in the high-dimensional document space.

It is worth noting that semi-supervised learning using the nearest neighbors graph approach in the euclidean distance space was originally proposed in the literatures [14] and [15], and LPI is also based on this idea. Differently, CPI is a semi-supervised learning using nearest neighbors graph approach in the correlation measure space. Zhong and Ghosh showed that 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 [2].

**B. CPI based Clustering Algorithm**

Given a set of documents \( x_1, x_2, \ldots, x_n \in \mathbb{R}^n \). Let \( X \) denotes the document matrix. The algorithm for document clustering based on CPI can be summarized as follows:

1. Construct the local neighbor patch, and compute the matrices \( MS \) and \( MT \).

2. Project the document vectors into the SVD subspace by throwing away the zero singular values. The singular value decomposition of \( X \) can be written as \( X = U \Sigma V^T \). Here all zero singular values in \( \Sigma \) have been removed. The vectors in \( U \) and \( V \) that correspond to these zero singular values are removed as well. Thus the document vectors in the SVD subspace can be obtained by \( X = U^T X \).

3. Compute CPI Projection. Based on the multipliers \( \lambda_1, \lambda_2, \ldots, \lambda_n \) one can compute the matrix \( W = \lambda_1 x_1 x_1^T + \ldots + \lambda_n x_n x_n^T \). Let \( W CPI \) be the solution of the generalized eigenvalue problem \( M_W = \lambda M_W \). Then, the low dimensional representation of the document can be computed by

\[
Y = W^T_{CPI} X = W^T X
\]

where \( W = UW CPI \) is the transformation matrix.

4. Cluster the documents in the CPI semantic subspace. Since the documents were projected on the unit hypersphere, the inner product is a natural measure of similarity. We seek a partitioning

\[
\{ \pi_i \}_{j=1}^k
\]

of the document using the maximization of the following objection function:

\[
Q(\{ \pi_j \}_{j=1}^k) = \sum_{j=1}^k \sum_{x \in \pi_j} x^T c_j
\]

with \( c_j = \frac{m_j}{|m_j|} \), where \( m_j \) is the mean of the document vectors contained in the cluster \( \pi_j \).
VI. Document Clustering with Ontology Support
The spectral clustering is applied on the correlation similarity model with nearest neighbor learning process. The Ontology repository is used to manage the term concept relations. Local patch extraction is carried out with Ontology support. Term frequency based weight is replaced with concept weight based model. The document preprocess operations are carried out to extract term information. Stopword elimination and stemming process are applied on the term collection. Porter stemming algorithm is used for suffix analysis. Ontology is used to extract term relationships. Statistical analysis model is used to estimate the term weight. Term relationships are not reflected in the term weight based model. Terms and their conceptual relationship are used in the weight estimation process. Related terms are grouped into the same concept level.
Subspace selection is performed with intrinsic structure of the document. Term and its geometric property extraction are required for the subspace filtering process. Concept identification process is integrated with the subspace selection task. Concepts and their associated elements are grouped into the same subspace. The document attribute dimension is increased in the term analysis model. Each term is represented as a separate entity in the vector. Irrelevant and infrequent terms are removed from the document vectors. Document vector dimension is reduced with concept groups. Ontology is constructed with concept and associated terms. The document contents are analyzed with the Ontology information. Semantic weight estimation is carried out for the concepts and its elements. Similarity analysis and clustering processes are done with semantic weight values.

VII. Concept Relationship Analysis for Document Clustering
The system is designed to group up relevant documents. Subspace based document clustering mechanism is used in the system. Correlation Preserving Index (CPI) is used in the document clustering process. The system is divided into six major modules. They are Document preprocess, Term analysis, Semantic analysis, Dimensionality reduction process, CPI estimation and Clustering process.
Document parsing, stopword elimination and stemming process are carried out under document preprocess module. Term weight estimated under term analysis module. Semantic analysis is performed to identify concept relationships. Dimensionality reduction process module is designed to reduce document vector size. CPI estimation module is designed to perform correlation similarity and index estimation process. Clustering process module is designed to partition the documents.

A. Document Preprocess
Document preprocess is performed to parse the text documents into words. Document cleaning is applied to remove stop words. Stemming process is applied to detect the base term. Terms are updated with their frequency values.

B. Term Analysis
The term analysis is performed to estimate the term weight values. Statistical method is used for the term weight estimation process. Term Frequency (TF) and Inverse Document Frequency (IDF) are used for the term weight estimation process. Terms and associated weight values are updated into the database.

C. Semantic Analysis
The semantic analysis is performed to identify the concept relationships. Ontology is constructed for the selected domains. Terms and associated concept relationships are identified using the Ontology. Semantic weights are assigned with reference to the concept relationship type.

D. Dimensionality Reduction Process
The document vector is constructed with term and weight values. The term weight based document vector is build with high dimensionality. Infrequent term elimination is performed with threshold values. The document vector is updated with reduced term collections.

E. CPI Estimation
The terms and its structure information are used to identify the subspaces. The correlation similarity is estimated with the subspace information. The correlation similarity is estimated between the documents. Correlation Preserving Index (CPI) is prepared with similarity values.

F. Clustering Process
The clustering process is applied to group up the documents using the similarity values. The CPI intervals are analyzed with threshold values. The cluster count is collected from the user. The clustering process is performed with term and semantic weight models.

VIII. Conclusion
The text documents are high dimensional data elements. Term and geometric patch information are used in the similarity analysis. Correlation Preserving Index (CPI) based clustering algorithm is used for the clustering process. Concept relationships are extracted and weight values are used in the clustering process. The dimensionality is reduced with concept relationships. Label based patch extraction model. Patches are represented with patch weight and term weight values. The system produces the clustering results in a hierarchical manner.

References


M. Mohanasundari received her B.E. degree in Computer science and Engineering from Anna University, Chennai, in 2006, the M.E. degree in Computer Science and Engineering from Anna University, Coimbatore, in 2010. She was worked as Lecturer for 2 years from 2006 to 2008 at Velalar College of Engineering and Technology, Erode. She has presented paper in various National and International Conferences. Her research interests include Data Mining and Biometrics. At present she is working as an Assistant Professor in Computer science and Engineering Department at Velalar College of Engineering and Technology, Erode.

P. Malathi, received B.E (CSE) degree in 2010 from Anna University Chennai ,India. Currently pursuing M.E degree in Computer Science and Engineering in velalar college of Engineering and Technology, Erode ,Affliated to Anna University Chennai ,India. Diploma in Thiru RamakrishnaNallammai Polytechnic college, Dharapuram.