Literature Review

The character of many types of clustering methods depends on the applications involved. Document analysis differ from other applications like spatial recognition and robotic vision; sample spaces in which large document sets rely, do not necessarily possess the regularity of extended space [4]. From its human-generated origins, text typically contains irregularity, where the addition of new text evolves in unpredictable ways [3]. Historically, probabilistic methods have been used to anticipate this irregularity. Ultimately this ignores vast amounts of text, and greatly affects the accuracy of the analysis [2]. Contrasting with probabilistic means, the popular k-means algorithm is deterministic in its implementation, often with strategic use of random partitioning [2].

The k-means algorithm has long been an industry standard in partition-based clustering of texts. As such, much of the developed literature starts with the procedure's shortcomings when performing highly specialized tasks. Yet the algorithm has also been used as a benchmark in the literature for measuring the performance of many other boutique procedures. Here, a short description of the k-means algorithm will provide a wider coin of vantage to the survey of document clustering.

In the field document analysis, mathematical and information theoretic tools are used to prepare text for an objective treatment. First, document sets are grouped together and relevant features, in the form of "content-bearing words" [3], are extracted and used to create a high-dimensional vector space. Individual documents would then exist as very sparse vectors (i.e. most words would be assigned to zero, as any given document uses only a small fraction of the lexical set) in the vector space. Then, the k-means algorithm would provide a similarity-measure for clusters of documents having affinity within a group, compared to clusters outside of the group.

One computational consideration that arises when running k-means, originates from the random selection of centers used to initialize the algorithm. With no unique starting point, many local minima may be found. The algorithm converges very fast, and even on large data sets this is not a problem. However, there is no guarantee that k-means will converge to the global minimum. Convergence of the algorithm to the global minimum is NP-hard. This exceeds realistic computational time constraints, and requires some additional heuristics to reduce computational effort [4].

As shown in Dhillon, et al. [4], one such heuristic, the spherical k-means technique, was developed to work with the limitations in the k-means random partitioning. After multiple iterations, k-means tends to become stuck in local minima, and not fully realize the optimal clustering scheme. To address optimization, Dhillon, et al. [4] suggested a “ping-pong” strategy that increases the computational efficiency of k-means by forming a sequence of separate clusters and moving certain documents from one cluster to another. The objective function could be studied to find a better optimization, thereby circumventing the local minimum problem. As noted by the authors of that study, the approach was limited to static amounts of documents clustered. Further work would be required for large variances in sizes of the initial document set.

Optimization was extended further in Kogan, et al. [7] by developing the distance-like function from combining the Squared-Euclidean distance with an information theoretic quantity. It was suggested that the resulting similarity measures could be tailored to specific data sets. Whole classes of variations on the k-means theme could then be studied.

Limitations to the standard k-means algorithm have also been treated with various hybrid approaches [5], using an algorithm like PDDP to guide the k-means algorithm into trajectories with higher quality. The principal direction divisive partitioning (PDDP) algorithm under consideration in this proposed study was developed by Boley who used a "divisive" method, where initially large document sets were divided into smaller partitions [2]. Boley defined "principal direction" as a process of "directing" each iteration of division by a new computation of the document space. This process further developed the distance function and similarity measure, thereby increasing computational efficiency [2].

The mathematical treatment implemented in the PDDP algorithm, can be summarized as follows. The projection of a set of vectors onto the nearest line starts the principal direction. From there, the problem reverts to maximizing the eigenvalue of the covariance matrix, by using the power method [6]. This result can also be carried over to the calculation of the Frobenius norm, in conjunction with the Lanczos algorithm. Through Lanczos, a sequence of diagonal matrices are constructed. Then, eigenvalues of the matrices are computed by finding the convergence of the largest eigenvalue. Applying these linear algebra techniques significantly reduces computational complexity and run-time [1], thereby rendering PDDP more practical for use in high-dimensional cluster analysis.

Research Objective

The ultimate objective of this study is to investigate a hybrid application of the PDDP algorithm to initialize the k-means clustering algorithm, as a way to reduce the computational effort exerted by the stand-alone k-means algorithm. The reduction in computational cost would compensate for the irregularity of text data, and enable more accuracy for sorting and querying from large document sets. A comprehensive review of the optimization performed on the PDDP algorithm, as well as cluster validation measures, will inform further investigations on the quality of the resulting clusters, and suggest further work on the strengths and limitations of hybrid approaches.

Analysis of the PDDP Algorithm

High-dimensional vector spaces, equipped with some distance-like function, are the standard environment for large document sets. This study begins with an overview of the linear algebra and optimization techniques, which underwrite these spaces. With deference to the notation of the prior works, vectors representing documents with m features will be represented as boldfaced, lowercase letters, namely, $\mathbf{b} = [b_1, b_2, ..., b_m]^{\mathrm{T}}$, and a collection of n documents will be denoted by the matrix $B = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n]$.

We start by finding the projection, $\mathbf{p}_{\mathrm{a}}$ on a line L in $\mathbf{R}^n$, parameterized by $\mathbf{y} + t\,\mathbf{x}$, on which is projected the document vector $\mathbf{a}$. Note that the vector $(\mathbf{y} - t_0\,\mathbf{x}) - \mathbf{a}$ is orthogonal to L, and also the vector $\mathbf{x}$ at the point of projection, for some t_0 , namely

$$(\mathbf{y} + t_0\,\mathbf{x} - \mathbf{a})^{\mathrm{T}}\,\mathbf{x} = 0 \;\Rightarrow\; \mathbf{y}^{\mathbf{T}}\mathbf{x} + t_0\,||\mathbf{x}||^2 - \mathbf{a}^{\mathbf{T}}\mathbf{x} = 0$$

$$\Rightarrow\; t_0 = (\mathbf{a}^{\mathbf{T}}\mathbf{x} - \mathbf{y}^{\mathbf{T}}\mathbf{x})/\,||\mathbf{x}||^2.$$

So, $\mathbf{p}_{\mathrm{a}} = \mathbf{y} + t_0\,\mathbf{x}$, where $t_0 = (\mathbf{a}^{\mathbf{T}}\mathbf{x} - \mathbf{y}^{\mathbf{T}}\mathbf{x})/\,||\mathrm{x}||^2$, which is the point that $\mathbf{a}$ projects on L. With this, then $(\mathbf{a} - \mathbf{p}_{\mathrm{a}})^{\mathrm{T}}\mathbf{x} = 0$. If we take $\mathbf{x}$ to lie on the unit sphere, and $\mathbf{y}$ orthogonal to $\mathbf{x}$, such that $\mathbf{x}^{\mathbf{T}}\mathbf{x} = 1 \Rightarrow ||\mathbf{x}||^2 = 1$, and $\mathbf{y}^{\mathbf{T}}\mathbf{x} = 0$, then

$$\mathbf{p}_{\mathrm{a}} = \mathbf{y} + [(\mathbf{a}^{\mathbf{T}}\mathbf{x} - \mathbf{y}^{\mathbf{T}}\mathbf{x}) / \|\mathbf{x}\|^2]\, \mathbf{x}$$

$$= \mathbf{y} + (\mathbf{a}^{\mathrm{T}}\mathbf{x})\, \mathbf{x}.$$

In this form, we hope to find the least squares approximation of a set of document vectors. To do this, we will minimize the sum of the distances of each projection $\mathbf{p}_i$ on L, with $\mathbf{a}_i$. In terms of $\mathbf{x}^{\mathbf{T}}\mathbf{x} = 1$,and $\mathbf{y}^{\mathbf{T}}\mathbf{x} = 0$**,** the sum, then is

$$\Sigma_{1\le i\le n}\ |\mathbf{a}_i - (\mathbf{y} + (a_i^T x)\,\mathbf{x})|^2.$$

Regrouping and simplifying each term $|\mathbf{a} - \mathbf{x}\mathbf{a}^{\mathrm{T}}\mathbf{x} - \mathbf{y}|^2$ of the sum, gives

$$|(\mathbf{a} - \mathbf{x}\mathbf{a}^{\mathrm{T}}\mathbf{x}) - \mathbf{y}|^2 = (\mathbf{a} - \mathbf{x}\mathbf{a}^{\mathrm{T}}\mathbf{x})^2 - 2\mathbf{y}(\mathbf{a} - \mathbf{x}\mathbf{a}^{\mathrm{T}}\mathbf{x}) + \mathbf{y}^2$$

$$= \mathbf{a}\cdot\mathbf{a} - 2\mathbf{a}(\mathbf{x}\mathbf{a}^{\mathrm{T}}\mathbf{x}) + (\mathbf{x}\mathbf{a}^{\mathrm{T}}\mathbf{x})^2 - 2\mathbf{a}\cdot\mathbf{y} + 2\mathbf{y}\cdot\mathbf{x}\mathbf{a}^{\mathrm{T}}\mathbf{x} + \mathbf{y}\cdot\mathbf{y}$$

$$= \mathbf{a}\cdot\mathbf{a} - 2(\mathbf{a}\cdot\mathbf{x})(\mathbf{a}\cdot\mathbf{x}) + (\mathbf{x}\cdot\mathbf{x})(\mathbf{a}^{\mathrm{T}}\mathbf{x})^2 - 2\mathbf{a}\cdot\mathbf{y} + 2(\mathbf{y}\cdot\mathbf{x})\mathbf{a}^{\mathrm{T}}\mathbf{x} + \mathbf{y}\cdot\mathbf{y}$$

$$= \mathbf{a}\cdot\mathbf{a} - 2(\mathbf{a}^{\mathrm{T}}\mathbf{x})^2 + (1)(\mathbf{a}^{\mathrm{T}}\mathbf{x})^2 - 2\mathbf{a}\cdot\mathbf{y} + 2(\mathbf{0})\mathbf{a}^{\mathrm{T}}\mathbf{x} + \mathbf{y}\cdot\mathbf{y}$$

$$= \mathbf{a}\cdot\mathbf{a} - (\mathbf{a}^{\mathrm{T}}\mathbf{x})^2 - 2\mathbf{a}^{\mathrm{T}}\mathbf{y} + \mathbf{y}\cdot\mathbf{y}$$

$$= (\mathbf{a} - \mathbf{y})^2 - (\mathbf{a}^{\mathrm{T}}\mathbf{x})^2.$$

So,

$$\Sigma_{1\le i\le n}\ |\mathbf{a}_i - (\mathbf{y} + (a_i^T x)\,\mathbf{x})|^2 = \Sigma_{1\le i\le n}\ |\mathbf{a}_i - \mathbf{y}|^2 - |\mathbf{a}_i^{\mathrm{T}}\mathbf{x}|^2.$$

Assuming a value for $\mathbf{x}$, we now use the method of Lagrange multipliers to find $\mathbf{y}$, such that

$$\min_{\mathbf{y}} \left\{\Sigma_{1\le i\le n}\ |\mathbf{a}_i - \mathbf{y}|^2, \text{ subject to } \mathbf{y}^{\mathrm{T}}\mathbf{x} = 0\right\}. \qquad (2)$$

Let $f(\mathbf{y})$ be the vector-valued objective function $f(y_1, \ldots, y_n) = \Sigma_{1\le i\le n}\ |\mathbf{a}_i - \mathbf{y}|^2$, and $g(\mathbf{y}) = \mathbf{y}^{\mathrm{T}}\mathbf{x} = 0$ be the constraint function. Taking the gradient of f yields

$$\nabla f(\mathbf{y}) = \Sigma_{1\le j\le n}\ f_j(y_1, \ldots, y_n)\, \mathbf{e}_j,$$

where f_j is the partial derivative of f with respect to the j-th component, and $\mathbf{e}_j$ is the unit basis vector of the j-th component, namely $\mathbf{e}_j = [0,\ldots, \mathbf{e}_j = 1,\ldots,0]^{\mathrm{T}}$.

With $|\mathbf{a}_i - \mathbf{y}|^2 = \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i - 2\mathbf{a}_i^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}$ for each i, then

$$\nabla f(\mathbf{y}) = \Sigma_{1\le j\le n}\ \partial/\partial y_j\ \left[\Sigma_{1\le i\le n}\ |\mathbf{a}_i - \mathbf{y}|^2\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\ \partial/\partial y_j\ \left[\Sigma_{1\le i\le n}\ \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i - 2\mathbf{a}_i^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}\right]\mathbf{e}_j$$

Since we are in $\mathbf{R}^n$, and the 2-norm is an assignment from $\mathbf{R}^n$ to $\mathbf{R}$ ($|\mathbf{a}_i - \mathbf{y}|^2$ is the Euclidean norm assigning the n-vector $\mathbf{y}$ to a real number), then from a theorem in real analysis, f is continuous. So, from another property of analysis, we can interchange the summand with partial derivative, yielding

$$\nabla f(\mathbf{y}) = \Sigma_{1\le j\le n}\ \Sigma_{1\le i\le n}\ \partial/\partial y_j\ \left[\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i - 2\mathbf{a}_i^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\ \Sigma_{1\le i\le n}\ \partial/\partial y_j\ \left[\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i - 2(a_{ji}y_j) + y_j^2\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\left[\Sigma_{1\le i\le n}\ (-2a_{ji} + 2y_j)\ \mathbf{e}_j\right]$$

$$= \Sigma_{1\le j\le n}\left[\Sigma_{1\le i\le n}\ (-a_{ji})\ \mathbf{e}_j + \Sigma_{1\le i\le n}\ y_j\ \mathbf{e}_j\right]$$

$$= 2\left[\Sigma_{1\le i\le n}\ (-\mathbf{a}_i) + \Sigma_{1\le i\le n}\ \mathbf{y}\right]$$

$$= 2n\mathbf{y} - 2\ \Sigma_{1\le i\le n}\ \mathbf{a}_i.$$

Now, taking the gradient of g, gives

$$\nabla g(\mathbf{y}) = \Sigma_{1\le j\le n}\ \partial/\partial y_j\ [y_j x_j]\ \mathbf{e}_j$$

$$\nabla g(\mathbf{y}) = \partial/\partial y_j\ [y_1x_1 + \ldots + y_nx_n]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\ (x_j)\mathbf{e}_j$$

$$= \mathbf{x}.$$

Employing the Lagrange multiplier, and solving the system for lambda yields

$$\nabla f(\mathbf{y}) = \lambda\nabla g(\mathbf{y}) \quad \Rightarrow \quad 2n\mathbf{y} - 2\,\Sigma_{1\le i\le n}\,\mathbf{a}_i = \lambda\,\mathbf{x}$$

Now, left-multiplying by $\mathbf{x}^{\mathrm{T}}$, and using the conditions $\mathbf{y}^{\mathrm{T}}\mathbf{x} = 0 \Rightarrow \mathbf{x}^{\mathrm{T}}\mathbf{y} = 0$, and $\mathbf{x}^{\mathrm{T}}\mathbf{x} = 1$,

$$\mathbf{x}^{\mathrm{T}}\,(2n\mathbf{y} - 2\,\Sigma_{1\le i\le n}\,\mathbf{a}_i\,) = \mathbf{x}^{\mathrm{T}}\,\lambda\,\mathbf{x}$$

$$\Rightarrow \qquad \mathbf{x}^{\mathrm{T}}\,(2n)\mathbf{y} - \mathbf{x}^{\mathrm{T}}\,(2)\,\Sigma_{1\le i\le n}\,\mathbf{a}_i = \lambda\,(\mathbf{x}^{\mathrm{T}}\,\mathbf{x})$$

$$\Rightarrow \qquad (2n)\mathbf{x}^{\mathrm{T}}\mathbf{y} - 2\mathbf{x}^{\mathrm{T}}\,\Sigma_{1\le i\le n}\,\mathbf{a}_i = \lambda\,(\mathbf{x}^{\mathrm{T}}\,\mathbf{x})$$

$$\Rightarrow \qquad \lambda = -2\mathbf{x}^{\mathrm{T}}\,\Sigma_{1\le i\le n}\,\mathbf{a}_i.$$

Substituting for lambda gives

$$2n\mathbf{y} - 2\,\Sigma_{1\le i\le n}\,\mathbf{a}_i = -2\mathbf{x}^{\mathrm{T}}\,\big(\Sigma_{1\le i\le n}\,\mathbf{a}_i\,\big)\mathbf{x}$$

$$\Rightarrow \qquad \mathbf{y} = 1/n\,\big[\Sigma_{1\le i\le n}\,\mathbf{a}_i - \mathbf{x}^{\mathrm{T}}\,\big(\Sigma_{1\le i\le n}\,\mathbf{a}_i\,\big)\mathbf{x}\big],$$

with **x** and its transpose as known entities. So, to minimize (2),

$$\mathbf{y} = 1/n\,\big[\Sigma_{1\le i\le n}\,\mathbf{a}_i - \mathbf{x}^{\mathrm{T}}\,\big(\Sigma_{1\le i\le n}\,\mathbf{a}_i\,\big)\mathbf{x}\big].$$

Now, we turn to the collection of column vectors representing the document set. First, we note the following property:

Property 1:

For $B = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n]$, with $A = \mathbf{b}_1\mathbf{b}_1^{\mathrm{T}} + \mathbf{b}_2\mathbf{b}_2^{\mathrm{T}} + \ldots + \mathbf{b}_{\mathrm{n}}\mathbf{b}_{\mathrm{n}}^{\mathrm{T}}$, then $BB^{\mathrm{T}} = A$.

Proof:

Since matrix B is an m x n array and B^{T} is an n x m array, then BB^{T} is an m x m array. Similarly for A, each $\mathbf{b_k b_k}^{\mathbf{T}}$ term, where

$$\mathbf{b_k b_k}^{\mathrm{T}} = [b_{\mathrm{k1}}\ b_{\mathrm{k2}}\ \ldots\ b_{\mathrm{km}}\ ]^{\mathrm{T}}\ [b_{\mathrm{k1}}\ b_{\mathrm{k2}}\ \ldots\ b_{\mathrm{km}}\ ],$$

is an m x m array, whose sum for $1 \leq k \leq n$ is also an m x m array. Let the ij^{th} entry of BB^{T} be denoted by c_{ij}. Then, performing matrix multiplication yields,

$$c_{ij} = b_{i1}b_{1j} + b_{i2}b_{2j} + \ldots + b_{im}b_{mj}.$$

Let the ij^{th} entry of $\mathbf{b_k b_k}^{\mathrm{T}} = d^{(k)}{}_{ij}$. Carrying out matrix multiplication gives $d^{(k)}{}_{ij} = b_{ik}b_{kj}$. Performing addition for $1 \leq k \leq n$ yields,

$$\Sigma_{1\leq k\leq n}\, d^{(k)}{}_{ij} = b_{i1}b_{1j} + b_{i2}b_{2j} + \ldots + b_{im}b_{mj}.$$

This sum is the ij^{th} entry of matrix A. Since $c_{ij} = \Sigma_{1\leq k\leq n}\, d^{(k)}{}_{ij}$ for the ij^{th} entries of BB^{T} and A respectively, therefore,

$$BB^{\mathrm{T}} = A.$$

Now, we want to exploit the symmetric and positive semi-definite properties of the matrix A. For

Property 2:

A is positive semi-definite.

Proof:

From the associative properties of matrix multiplication,

$$\mathbf{x}^{\mathrm{T}}\mathbf{B}\ \mathbf{B}^{\mathrm{T}}\mathbf{x} = (\mathbf{x}^{\mathrm{T}}\mathbf{B})(\mathbf{B}^{\mathrm{T}}\mathbf{x}).$$

Since $\mathbf{x}^{\mathrm{T}}\mathbf{B} = (\mathbf{B}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}$, and $\mathbf{x}^{\mathrm{T}}\mathbf{B}$ is the transpose of a column vector $\mathbf{B}^{\mathrm{T}}\mathbf{x}$, then $(\mathbf{x}^{\mathrm{T}}\mathbf{B})(\mathbf{B}^{\mathrm{T}}\mathbf{x})$ is a dot product of $\mathbf{B}^{\mathrm{T}}\mathbf{x}$ with itself. From the positivity property of inner product spaces (i.e. $\langle v,v\rangle \geq 0$, for all $v \in \mathbf{R}^n$), then

$$\mathbf{x}^{\mathrm{T}}\mathbf{B}\ \mathbf{B}^{\mathrm{T}}\mathbf{x} \geq 0.$$

So, A is positive semi-definite.

With the fact that *A* is positive semi-definite, then we know also that all of the eigenvalues of *A* are real. We know this from the real spectral theorem, and since **A** can be decomposed into the product $\mathbf{UDU}^{\mathrm{T}}$, where **U** is an orthogonal matrix and **D** is diagonal, with real entries on the main diagonal equal to the eigenvalues of **A**. Hence, the eigenvalues are real. *(From Elementary Linear Algebra, Ron Larson, 7th Ed., p.362; also http://web.mit.edu/jorloff/www/18.03-esg/notes/symmetricMatrices.pdf)*

We state the next property:

Property 3:

Every eigenvalue of A is non-negative.

Proof:

Since **A** is symmetric, then for any arbitrary eigenvalue λ of **A** satisfying $\mathbf{Ax} = \lambda\mathbf{x}$, $\lambda \in \mathbf{R}$, left multiplying $\mathbf{x}^{\mathrm{T}}$ to both sides of $\mathbf{Ax} = \lambda\mathbf{x}$, yields

$$\mathbf{x}^{\mathrm{T}}\mathbf{Ax} = \mathbf{x}^{\mathrm{T}}\lambda\mathbf{x} = \lambda\mathbf{x}^{\mathrm{T}}\mathbf{x}.$$

Since **A** is positive semi-definite,

$$\mathbf{x}^{\mathrm{T}}\mathbf{Ax} \geq 0 \;\Rightarrow\; \lambda\mathbf{x}^{\mathrm{T}}\mathbf{x} \geq 0,$$

and by positivity of inner product spaces,

$$\mathbf{x}^{\mathrm{T}}\mathbf{x} \geq 0 \;\;\Rightarrow\;\; \lambda \geq 0.$$

Therefore, the eigenvalues of A are non-negative.