Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

A Classification of Arial Data Based on Data Mining Clustering Algorithm

Dr. Vuda.Sreenivasarao[1], Dr.P.Ramesh P.V.V.S.Gangadhar[2] and Dr.G.Ramaswamy[3]

[1] JNTU

Received: 13 November 2011 Accepted: 12 December 2011 Published: 23 December 2011

Abstract

The Arial data contains date periodically observed with parameters of texture (min, max), flora, and density (min, max). The proposed Arial prediction system cluster and analyze, three input features that is average texture, flora, average density according to number of days to predict Arial for Surveillance applications. The proposed system realizes the k-means clustering algorithm for grouping similar features based on user intended period, further the system analyze using PCA (Principal Component Analysis) on same data.

Index terms— Data mining, Arial data, cluster algorithm, Principal Component Analysis.

1 INTRODUCTION

he Arial data contains date periodically observed with parameters of texture (min, max), flora, and density (min, max). The amount of data kept in computer files and databases is growing at a phenomenal rate. At the same time, the users of these data are expecting more sophisticated information from them. For example a marketing manager is no longer satisfied with a simple listing of marketing contacts, but wants detailed information about customers past purchases as well as predictions of future purchases. Simple structured/query language queries are not adequate to support these increased demands for information. Data mining steps in to solve these needs. Data mining is often defined as finding hidden information in a database. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association, sequence or path analysis, classification, clustering, and forecasting. This paper deals with implementation of an automated Arial prediction system for agriculture applications using data mining tools such as clustering (k-means algorithm) and Principle Component Analysis (PCA). For making accurate decision on large observation is an important factor, but with increasing information the clustering algorithm faces various limitations/problems. Among them current clustering techniques do not address all the requirements adequately (and concurrently). Dealing with large number of dimensions and large number of data items can be problematic because of time complexity. The effectiveness of the method depends on the definition of ”distance” (for distance-based clustering). If an obvious distance measure doesn’t exist we must ”define” it, which is not always easy, especially in multi-dimensional spaces. The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways. The main objective of this paper is to develop an accurate and efficient Arial prediction system for agriculture applications using data mining tools such as clustering (k-means algorithm) and PCA.

2 II.

3 DATA CLUSTERING

Clustering is a divided number of groups of similar data objects. Each group called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing the data by fewer clusters

necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in numerical analysis, mathematics and statistics. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. In practical perspective clustering plays an outstanding performance in data mining applications such as, computational biology ,information retrieval and text mining, scientific data exploration ,marketing, medical diagnostics, spatial database applications, and Web Clustering is the subject of active research in several fields such as statistics, pattern recognition, data mining, grouping and decision making, pattern classification, bio informatics and machine learning. A very important characteristic of most of these application domains is that the size of the data involved is very large. So, clustering algorithms used in these application areas should be able to handle large data sets of sizes ranging from gigabytes to terabytes and even pica bytes. Typically, clustering algorithms paper on pattern matrices, where each row of the matrix corresponds to a distinct pattern and each column corresponds to a feature. Most of the early paper on clustering dealt with the problem of grouping small data sets, where the benchmark data sets used to demonstrate the performance of the clustering algorithms were having a few hundreds of patterns and a few tens of features. Fisher's Iris data is one of the most frequently used benchmark data sets. This data has three classes, where each class has 50 patterns and each pattern is represented using four features. However, several real-world problems of current interest are very large in terms of the pattern matrices involved. For example, in data mining and web mining the number of patterns is typically very large, whereas in clustering biological sequences, the number of features involved is very large. Cluster analysis is a way to examine similarities and dissimilarities of observations or objects. Data often fall naturally into groups, or clusters, of observations, where the characteristics of objects in the same cluster are similar and the characteristics of objects in different clusters are dissimilar. In one of the earliest books on data clustering, Underberg defines cluster analysis as a task, which aims to finding of natural groups from a data set, when little or nothing is known about the category structure. Bailey, who surveys the methodology from the sociological perspective, defines that cluster analysis seeks to divide a set of objects into a small number of relatively homogeneous groups on the basis of their similarity over N variables. N is the total number of variables in this case.

4 III.

5 SYSTEM ARCHITECTURE

The Arial prediction system architecture is shown in figure 2.The overall system design consists of Input (Arial Data), modified input, Feature selection, Clustering Using k-means on Selected Feature And PCA on Selected Feature.

Output: Clustering is a division of data into groups of similar objects. Each group called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. K-means is a typical unsupervised learning clustering algorithm. It partitions a set of data into k clusters. However, it assumes that k is known in advance. Following is the summary of the algorithm:

1. Put K points into the representation of space by the objects that are being clustered. These points represent group Centroids.

2. Allocated each object to the group that has the closest centroids 3. When all objects have been allocated, again calculation of the positions of the K centroids. 4. Repeating of Steps 2 and 3 up to the centroids no longer move. This produces a separation of the objects into number of groups from which the metric to be minimized can be calculated. The papering flow of "clustering using k-means on selected data" is explained by flowchart shown in figure 3. Start the procedure, next accept starting period, ending period data from user, and accept k value from user. Accept clustering data option, 4) Average Texture 5) Flora 6) Average Density Cluster Arial data using k-means algorithm on selected feature (between starting and ending period days). If select 4 as our option than average Texture data is cluster, if select 5 than rain, if select 6 than average density data is cluster using k-means clustering algorithm. Display final results and Stop the procedure. Apply PCA algorithm on selected feature. The papering flow of "PCA on selected data" is explained by flowchart shown in figure **??**. IV.

6 Input

7 RESULT ANALYSIS

8 CONCLUSION

The Arial data contains date periodically observed with parameters of texture (min, max), flora, and density (min, max).Farmer needs timely and accurate Arial data. In order to achieve this, data should be continuously recorded from stations that are properly identified, manned by trained staff or automated with regular maintenance, in good papering order and secure from tampering. The stations should also have a long history and not be prone to relocation. The collection and archiving of Arial data is important for the input information because it provides an economic benefit but the local/national economic needs are not as dependent on high data quality as is the Arial risk market. In this study, it was found that the data mining tools could enable experts to predict Arial with satisfying accuracy using as input the Arial parameters of previous years. The Kmeans clustering and PCA

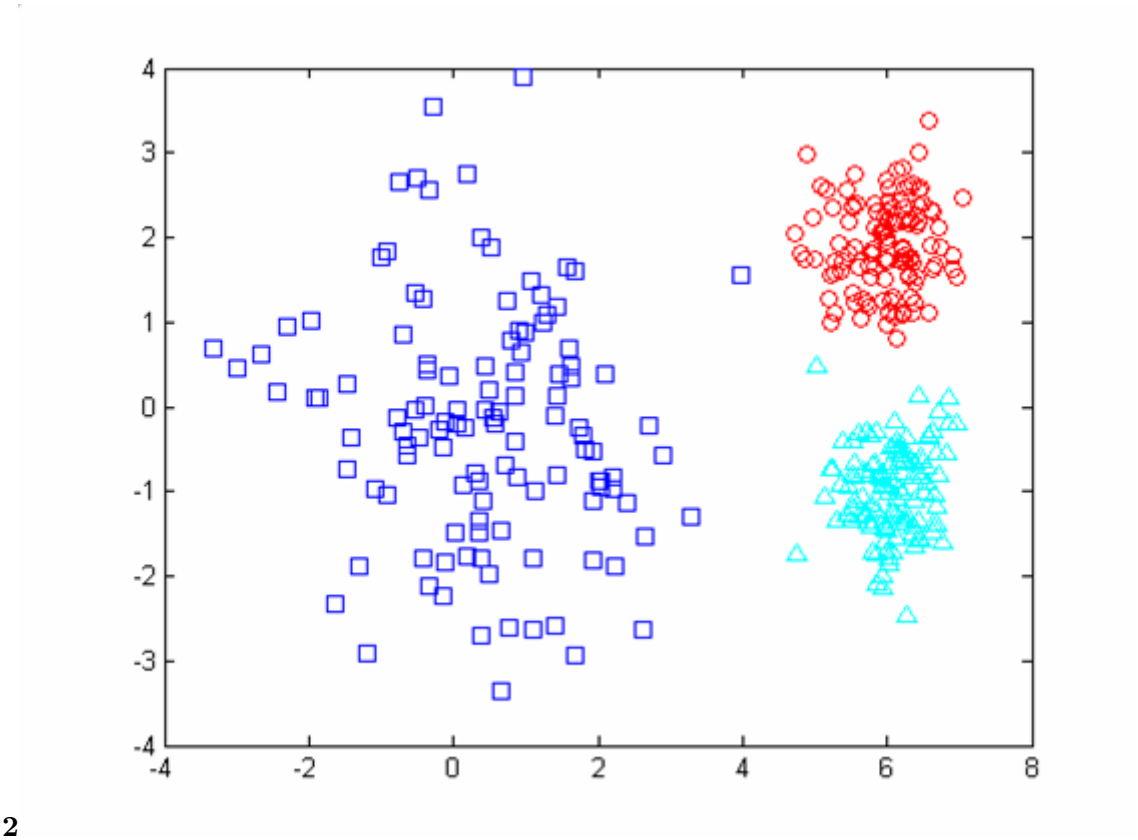
algorithms are suggested and tested for period of 11 years with multiple features to early prediction of Arial for agriculture applications. [1] [2]

2011

Figure 1: T © 2011

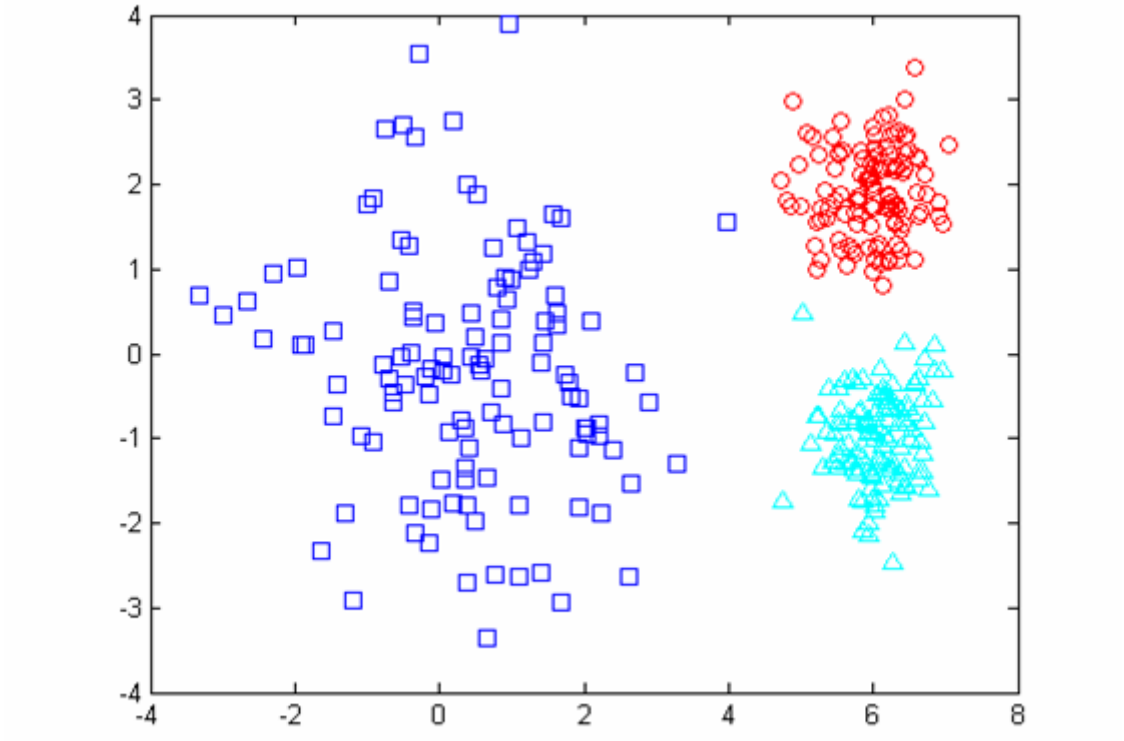
[1] © 2011 Global Journals Inc. (US) Global Journal of Computer Science and Technology Volume XI Issue XXII Version I

[2] December



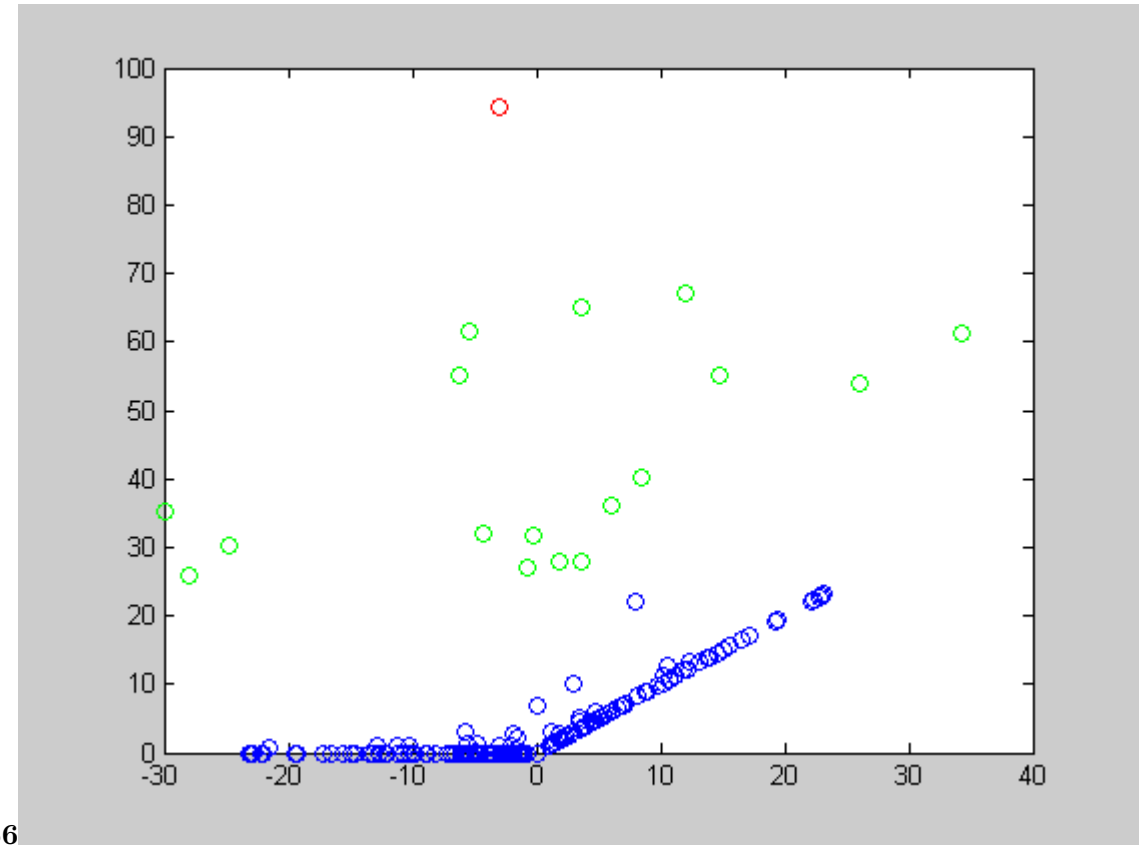
2

Figure 2: Figure 2 :



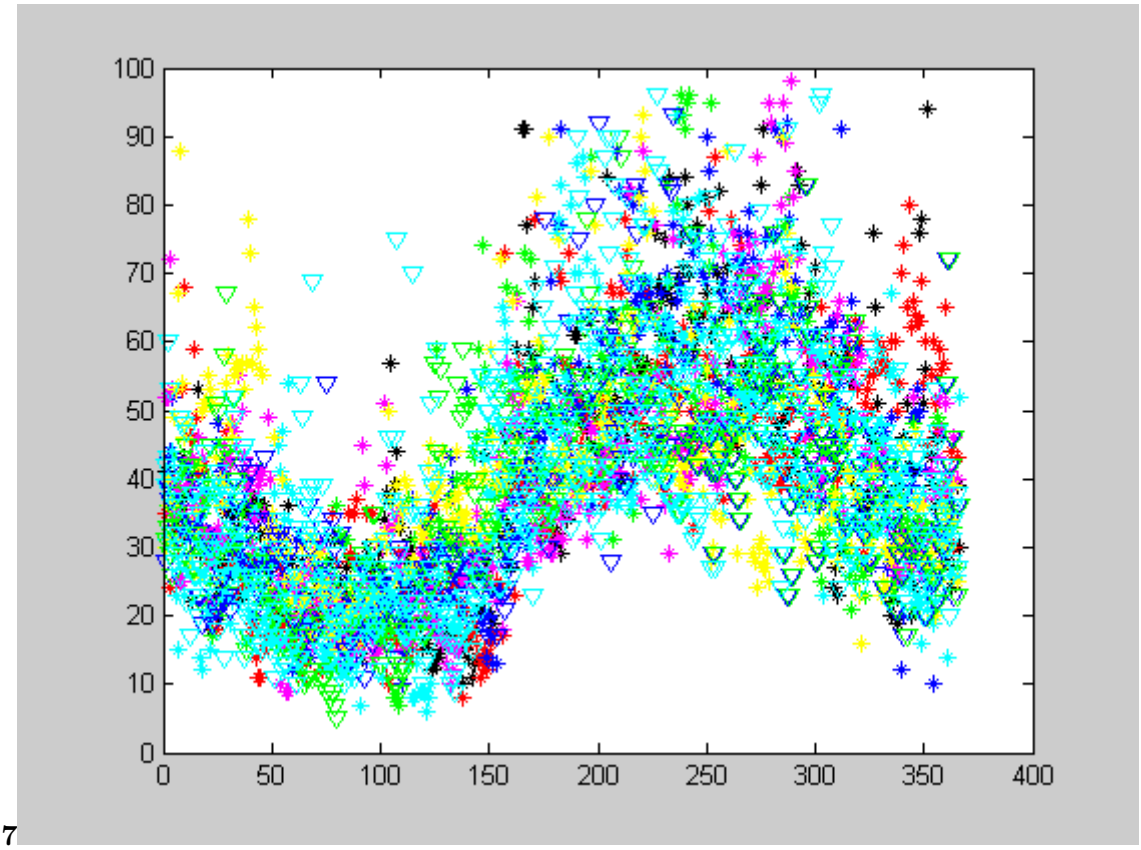
3

Figure 3: Figure 3 :



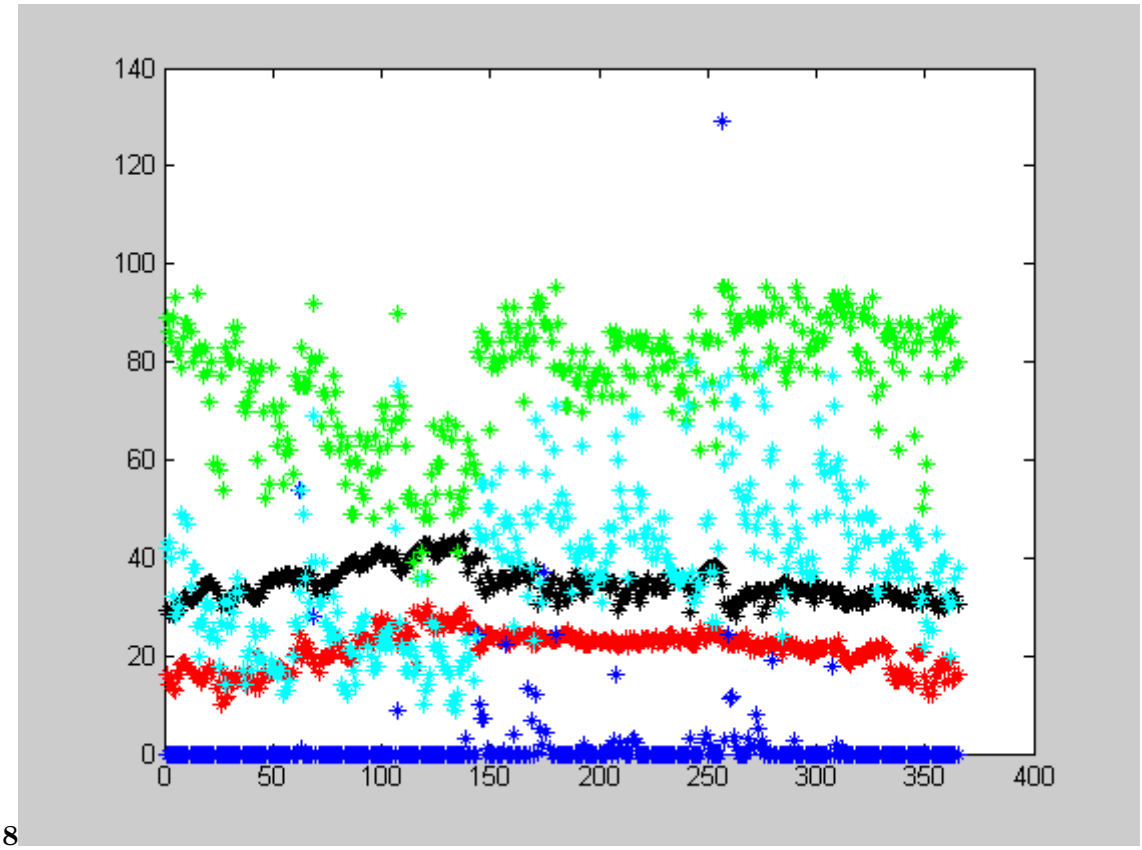
56

Figure 4: Figure 5 :Figure 6 :



7

Figure 5: Figure 7 :



20118

Figure 6: © 2011 DecemberFigure 8 :

[Todd et al. ()] 'A combined satellite infrared and passive microwave technique for estimation of small-scale flora'. M C Todd , C Kidd , D Kniveton , T J Bellerby . *J.Atmos, Oceanic technol* 2001. 18 p. .

[Grimes et al. ()] 'A neural netpaper approach to real-time flora estimation for Africa using satellite data'. D I Grimes , E Coppola , M Verdecchia , G Visconti . *J. Hydrometeorol* 2003. 4 p. .

[Adler and Negri ()] 'A satellite technique to estimate tropical convective and stratiform flora'. R F Adler , A J Negri . *J. Appl. Meteorol* 1988. 27 p. .

[Degaetano ()] 'A spatial grouping of United States climate stations using a hybrid clustering approach'. Degaetano . *International Journal of Climatology* 2001. 21 p. .

[Jain and Dubes ()] *Algorithms for Clustering Data*, K Jain , R C Dubes . 1988. New Jersey: Prentice-Hall.

[Anderberg] M R Anderberg . *Cluster analysis for applications*, (London) Academic Press, Inc. p. 973.

[Bailey ()] K D Bailey . *Cluster analysis*, 1975. p. .

[Bradley et al. ()] 'Clustering very large databases using EM mixture models'. P Bradley , C Reina , U Fayyad . *Proceedings of 15th International Conference on Pattern Recognition (ICPR'00)*, (15th International Conference on Pattern Recognition (ICPR'00)) 2000. 2 p. .

[Guha and Rastogi ()] 'Cure: An efficient clustering algorithm for large databases'. S Guha , . R Rastogi , Shim . *Proceedings of the ACM Sigmod Conference*, (the ACM Sigmod ConferenceSeattle, WA) 1998. p. .

[Hand et al. ()] *Principles of Data Mining*, David J Hand , Heikki Mannila , Padhraic Smyth . 2001. MIT Press.

[Macqueen et al. (1967)] 'Some methods for classification and analysis of multivariate observations'. J Macqueen , I Lindsay , Smith . *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, (of the Fifth Berkeley Symposium on Mathematical Statistics and Probability) 1967. February 26, 2002. p. .

[Little and Rubin ()] *Statistical analysis with missing data*, R J Little , D B Rubin . 1987. John Wiley & Sons.

[Fisher ()] 'The Use of Multiple Measurements on Taxonomic Problems'. R A Fisher . *Annals of Eugenics* 1936. 7 p. .

[Berry and Browne ()] *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, M Berry , Browne . 1999. SIAM.

[Cooley ()] *Web usage mining: discovery and application of interesting patterns from web data*, R W Cooley . 2000. University of Minnesota, USA (PhD thesis)