Abstract

A novel method of efficient spam mail classification using clustering techniques is presented in this research paper. E-mail spam is one of the major problems of the today’s internet, bringing financial damage to companies and annoying individual users. Among the approaches developed to stop spam, filtering is an important and popular one. A new spam detection technique using the text clustering based on vector space model is proposed in this research paper. By using this method, one can extract spam/non-spam email and detect the spam email efficiently. Representation of data is done using a vector space model. Clustering is the technique used for data reduction. It divides the data into groups based on pattern similarities such that each group is abstracted by one or more representatives. Recently, there is a growing emphasis on exploratory analysis of very large datasets to discover useful patterns, it is called data mining. Each cluster is abstracted using one or more representatives. It models data by its clusters. Clustering is a type of classification imposed on a finite set of objects. If the objects are characterized as patterns, or points in a n-dimensional metric space, the proximity
A Novel Method of Spam Mail Detection using Text Based Clustering Approach

measure can be the Euclidean distance between pair of points or similarity in the form of the cosine of the angle between the vectors corresponding to the documents. In the work considered in this paper, an efficient clustering algorithm incorporating the features of K-means algorithm and BIRCH algorithm is presented. Nearest neighbour distances and K-Nearest neighbour distances can serve as the basis of classification of test data based on supervised learning. Predictive accuracy of the classifier is calculated for the clustering algorithm. Additionally, different evaluation measures are used to analyze the performance of the clustering algorithm developed in combination with the various classifiers. The results presented at the end of the paper in the results section show the effectiveness of the proposed method.

Reference

- http://www.informationretrieval.org/
- http://www.aueb.gr/users/ion/publications.html
- http://www.cl.cam.ac.uk/users/bwm23/
- http://www.wikipedia.org
- Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, Second Edn.
A Novel Method of Spam Mail Detection using Text Based Clustering Approach


Index Terms

Computer Science Data Mining

Key words

Spam Mail Detection
Text Based Clustering Approach
K-Nearest