International Journal for Rapid Research in Engineering Technology & Applied Science
Vol 5 Issue 9 November 2019
ISSN Online:- (2455-4723)

Cloud Running Cost Optimization using Inactive VM Identification by Data Mining Technique

Vaidehi Bakshi[1], Mohit Jain[2]
1 M.Tech Scholar, 2 Head of Department Computer Science
1,2 Department of Computer Science & Engineering
1,2 BM Group of Collage of Engineering And Technology ,Indore, Madhya Pradesh ,India
vaidehibakshi35@gmail.com* , hod.computers@bmcollege.ac.in**

Abstract: **Cloud is a huge infrastructure; it is composed of networks, data centers and the brokers. Among broker is third party entity which is associated with a number of cloud infrastructures. Additionally different data centers are associated with other infrastructures. That mess is created to offer scalability on computational resources. The service providers borrow the resources from other infrastructure when the self resources are in high workload, and the resources are ideal when the fewer loads arise for computation. Both the conditions increase the cloud infrastructure running cost. In this context this issue is observed when the resources are not properly managed. Thus the cloud computing is focused on appropriate utilization of computational resources for achieving the higher performance from the cloud infrastructures. The management of resources can be possible by adopting the scheduling strategy or by managing the virtual machines. The proposed work is focused on management of virtual machines in cloud servers. In this context the proposed work offers a data mining technique that is helping us to identify the ideal or inactive VMs in infrastructure. Basically the inactive VMs are degrading the services of cloud data centers because these machines engage the cloud resources but it not functioning as requirements. Thus the inactive VMs are increases the server running cost, to maintain the performance it is necessary to identify and repair the VMs. The proposed technique usages the k-means and SVR (support vector regression) to classify the cloud server log for identifying the target types of VMs. The implementation of the presented technique is performed using JAVA technology and it is observed the proposed technique works more efficiently as compared to the similar available techniques..**

Keywords: ***Data mining, server running cost optimization, inactive VM classification, performance improvement of cloud, SVR, K-means.***

I Introduction

The cloud computing enhancing the computational experience, it offers the computational as well as the storage resources. Due to its characteristics it offers shared resources, scalable computational and storage resources, and more. In order to achieving this, different resource management techniques are applied. Among them the virtualization, load balancing and resource scheduling techniques are much popular. In this work virtual machine management is simulated for recovering performance on existing cloud infrastructure. Therefore, the aim of the cloud is to optimize the resource utilization for maintaining the service quality. The proposed work is involved for resource management of cloud servers. The data centers are basically configured with the multiple VMs for effective resource utilization, but the less number of active VM not provide effective productivity.

Thus identification and repairing of VMs are required. In this context for classifying the VMs the data mining technique is used. That technique help to analyze the server log and provide the outcomes based on classification of active and inactive VMs. The proposed work involves the study of data mining learning methods supervised and unsupervised respectively. The data mining techniques are employable over the data in different manner for recovering the target patterns. This target patterns are help to recognize the properties of VM which are not working or damaged. However the proposed work involves the supervised learning approach for efficient and accurate classification of inactive VMs.

II PROPOSED WORK

The proposed work is intended to design and develop a data mining model for recognizing the inactive or ideal VM (virtual machine) in cloud infrastructure. In this context this chapter provides the details about the proposed system and their functional aspects.

A. System Overview

Data mining techniques are become popular due to it's data processing ability. In this context the supervised and unsupervised learning techniques are used to learn over some predefined patterns and then the algorithms classify or categorize the data based on their previous experience or training experience. However the performance of pattern recognition of supervised learning techniques is more accurate then unsupervised learning techniques. In this presented work an application of data mining is provided in the domain of cloud computing. The cloud computing is accepted now in these days for efficient, scalable and low cost solutions. In this context a significant efforts on management level (SLA) required. For example when the server is over loaded then the resources are barrowed from other service providers and when the resources are ideal. In both the cases the server running cost is wearied by the cloud service providers.

In this context various strategies are utilized for managing the cloud resources i.e. cloud load balancing, cloud job scheduling and VM management. This presented work involves the VM management techniques for improving the performance of data centers. The cloud data centers involve the VMs (virtual machines) which are used for efficient computing experience. But when the VMs are compromised or damaged due to some reasons then the performance of entire data center is affected, because the inactive or ideal VM capture the cloud resources but non-functioning for execution of the user request. In this context we need a technique that identifies the active and inactive VMs. By estimating the inactive VMs the service providers can improve their productivity and cloud service running cost. This section provides the basic understanding about the proposed system. The next section provides the details how the proposed model works to identify the inactive VMs in cloud data centers.

B. Methodology

The proposed inactive classification model is demonstrated using figure 2.1, in this diagram the different modules of the system is simulated and this section contains detailed discussion of each module:

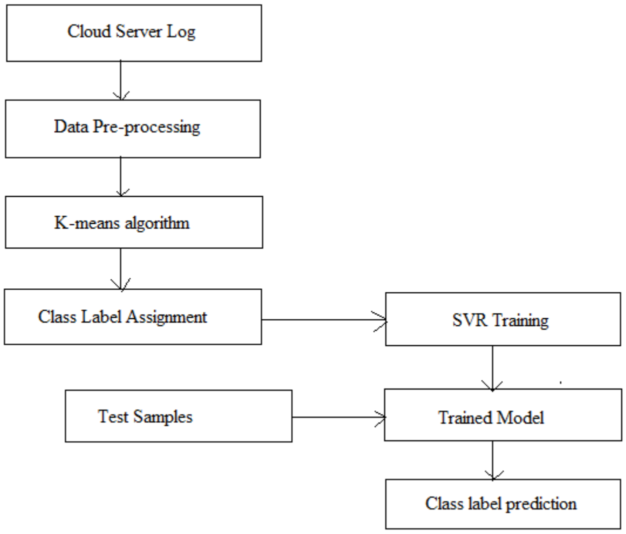

Figure 2.1 proposed system

Cloud server log: the cloud data centers are preparing a log file for every event which is happen in server. This log file contains the different attributes which are relevant to the task, VM input, output, time taken, job description, bandwidth and others. Using these log attributes the available VMs are estimated to get their functioning. The required cloud server VM log file is taken from the [36].

Data preprocessing: the data preprocessing is an essential step in data mining. The available raw data is processed before utilizing the data with the machine learning algorithms for optimizing the quality of data. In this context the data and attributes are removed, added, transformed or mapped with the similar values which are acceptable with the target machine learning algorithm.

k-means algorithm: however the data available in cloud server log files are not 2eighbou as the active and inactive VMs. It only contains attributes but to use with the supervised learning algorithm it is required the data instances has the predefined class labels. Therefore the proposed work involves the k-means clustering algorithm to define class labels for all the dataset instances. The working of k-means clustering algorithm is given as:

The classical k-means algorithm is given using table 2.1. This is an unsupervised learning algorithm that works directly on the supplied data to recover the patterns. In this context the algorithm usages the distance function for computing the pattern similarity. The less distance demonstrates the most similar object in a given dataset.

The k-means algorithm initially selects the random centroids and compare with all the available data instances in dataset. After comparing each pattern with the centroids the decision is made to group them. If the solution is not much suitable then the algorithm computes the new centroid based on previous steps, this process is continued till the best solution is not found.

Input: N objects to be cluster ($x_1, x_2 \dots x_n$), amount of clusters to compute k;

Output: k clusters, sum of dissimilarity among objects and its 3eighbour centroid;

Process:

1. Algorithm select k random instances as initial centroids $(m_1, m_2, \dots, m_k)$;

2. measure distance among each data instance X_i and centroid, and create label as nearest centroid, calculating distance using following formula:

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} (x_i - m_{j1})^2}, i = 1 \dots N, j = 1 \dots k$$

$d(x_i, m_i)$ denotes distance between data instance x and m.

3. compute mean of data instances in each centroid to update cluster centers,

$$m_i = \frac{1}{N} \sum_{j-1}^{n_i} x_{ij}, i = 1,2, \dots, K$$

N_i is number of instances in a cluster i;

4. reiterate 2) 3) solution is achieved, and algorithm return $(m_1, m_2, \dots, m_k)$ clusters.

Table 2.1 K-means clustering

Class label assignment: the k-means clustering algorithm clusters the entire dataset instances using the internal pattern similarity. Additionally during the clustering it assign a class label to instance which is depends upon the nearest cluster center or centroids.

SVR training: support vector regression (SVR) is an improved version of SVM (support vector machine). That is designed to classify the data in multiple classes. Therefore this technique also utilizes the concept of SMO classification. The key difference between SVM and SVR is that SVM usages the basic distance based approaches to deploy the hyper plan, on the other hand the SVR usages the regression based optimization technique to classify data more accurately.

Trained model: after training the SVR algorithm returns a data model instance which is ready to use for classification problem. Therefore it is the outcome of SVR training with the training data samples.

Test samples: in order to verify the learning of SVR, some preprocessed log data samples are used here for testing of the system. These samples are classified using the trained model of SVR.

Class label prediction: the trained SVR accept the test samples one by one, additionally for each instance of data generates the class labels. Based on this class label prediction the performance of classifier is measured.

III Proposed Algorithm

This section summarizes the proposed algorithm for classification of cloud server log for identification of active and inactive VMs. The table 2.2 contains all the required steps.

Input: cloud server log L, number of clusters k=2, Test samples T

Output: predicted class label C

```
Process:
    1. $R_n = ReadData(L)$
    2. $P_n = preProcessData(R_n)$
    3. $[index, centroid] = kmeans.Clusters(P_n, k)$
    4. $Model_{train} = SVR.Train(P_n, centroid)$
    5. $C = Model_{train}.Classify(T)$
    6. Return C
```

Table 2.2 proposed algorithm

IV Result Analysis

The proposed VM classification system for finding the inactive VMs is implemented successfully. This chapter provides the understanding about the performance of the implemented system.

A. Accuracy

The accuracy of any data mining algorithm is measured using the ratio of total correctly recognized data samples and the total samples to be recognized. That is a measurement of correctness. For estimating the accuracy the following formula is used:

$$accuracy = \frac{total\ corrrectly\ recognized\ samples}{total\ samples\ to\ classify} X100$$

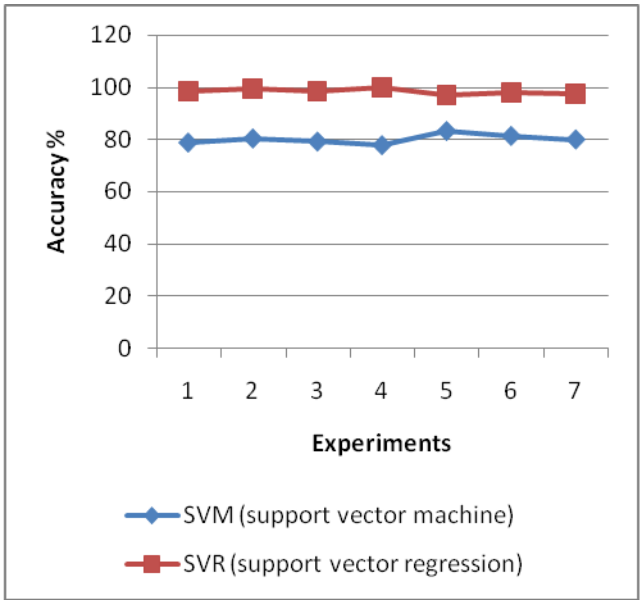

Figure 3.1 accuracy %

Experiments	SVM (support vector machine)	SVR (support vector regression)
1	78.91	98.27
2	80.42	99.31
3	79.25	98.54
4	77.93	99.92
5	83.19	97.07
6	81.38	98.21
7	80.06	97.38

Table 3.1 accuracy %

The classification accuracy of two classifiers for classifying the inactive VM classification data set is demonstrated in table 3.1 and figure 3.1. The table contains the obtained observations of experiments conducted with the system and their visualization using the line graph is given figure 3.1. According to demonstrated results both the algorithms providing the consistent accuracy for the classification system among the proposed SVR based classification technique outperform as compared to the SVM based traditional classification technique. Therefore the proposed system is acceptable for accurate VM identification.

B. Error Rate

The error rate of a data mining model shows the incorrectly recognized instances during data processing using any machine learning algorithm. That is a ratio of incorrectly classified instances and the total data instances provided for classification. In order to calculate the error rate of algorithm the following formula can be used:

$$Error\ rate = \frac{incorrectly\ recognized\ instances}{total\ instances\ to\ recognize} X100$$

Or

$$error\ rate = 100 - error\ rate$$

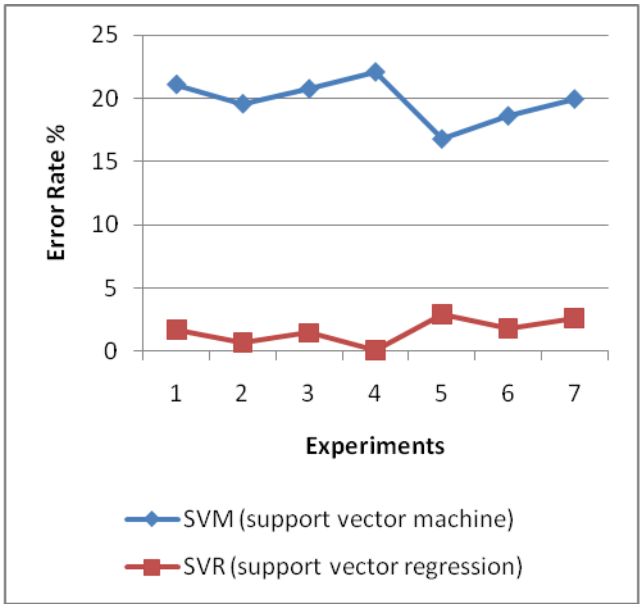

Figure 3.2 error rate %

Experiments	SVM (support vector machine)	SVR (support vector regression)
1	21.09	1.73
2	19.58	0.69
3	20.75	1.46
4	22.07	0.08
5	16.81	2.93
6	18.62	1.79
7	19.94	2.62

Table 3.2 error rate %

During the experiments the obtained percentage error rate is given in figure 3.2 and table 3.2. The error rate of the proposed system and the traditional SVM based classification system is compared using table 3.2 and obtained error rate is reported. These values are used with the figure 3.2 for providing the line graph. According to this figure X axis contains the experiments and the Y axis shows the obtained percentage error rate. According to the outcomes out proposed system demonstrate the stable and accurate classification outcomes are compared to the traditional SVM based technique.

C. Memory Usages

An algorithm what amount of main memory resources are usages is known as the memory usages of the system. The memory usages of an algorithm are also known as space complexity of the system. The memory usages of a java based algorithm are computed using the following formula:

$$memory\ usages = total\ assigned\ memory - free\ memory$$

The memory usages of the proposed inactive VM classification system and the traditional SVM based technique are demonstrated in this section. In this context the table 3.3 contains the memory usages of the algorithm during different number of experiments. These obtained values are visualized using the line graph as given in figure 3.3. This diagram shows the number of experiments conducted in X axis and Y axis denotes the obtained memory usages in terms of kilobytes (KB). According to the obtained results the proposed system consumes additional amount of memory as compared to the traditional SVM based algorithm because the proposed algorithm for accuracy improvements involve the technique of K-means algorithm too.

Experiments	SVM (support vector machine)	SVR (support vector regression)
1	20184	16739
2	19373	17827
3	19883	16827
4	20171	17882
5	18830	18301
6	19382	17392
7	19928	18837

Table 3.3 memory usages

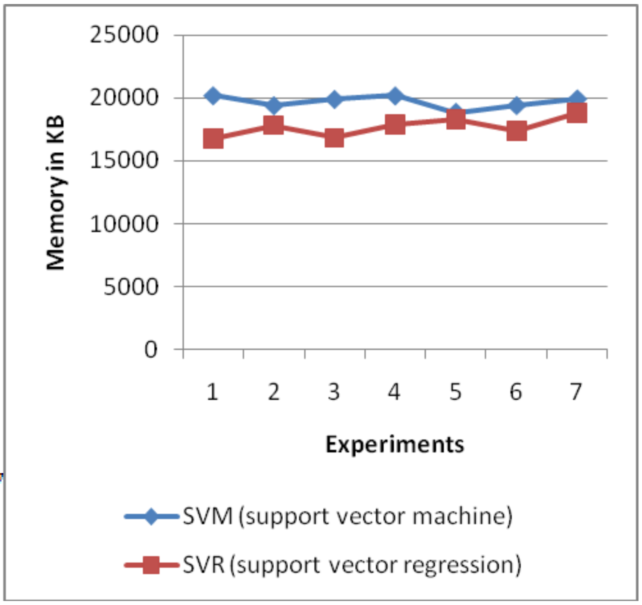

Figure 3.3 memory usages

A. Time Consumption

In order to train a supervised learning algorithm an amount of time required to process the data. This amount of time requirement is known as time consumption of the system. To calculate the time consumption the following formula is used.

$$time\ consumed = end\ time - start\ time$$

Experiments	SVM (support vector machine)	SVR (support vector regression)
1	278	309
2	252	281
3	266	293
4	271	284
5	262	273
6	277	291
7	284	309

Table 3.4 time usages

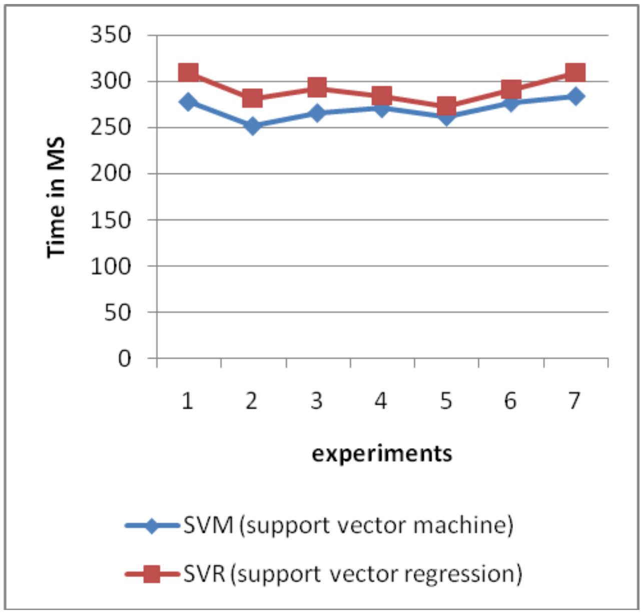

Figure 3.4 time usages

The time consumption of the proposed system is described in figure 3.4, this figure is prepared using the observed values given in table 3.4. This table includes the time requirements for calculating class labels. Additionally in figure 3.4 the required time is notified in terms of milliseconds (MS). The X axis of diagram shows the number of experiments performed and Y axis shows the expense of time. According to the given diagram the SVM consumes less amount of time as compared to proposed SVR based classification system.

V.Conclusion And Future Work

The proposed work is intended to design and develop a data mining model that help to improve the productivity of the cloud computing. In this context summary of research work is submitted in this chapter, additionally some essential future extensions are also reported.

A. Conclusion

The cloud computing offers scalable computing, plague and play resources. Additionally according to the use of application that is able to arrange the required resources. But to maintain the service quality a significant amount of background efforts are required. The cloud computing is a huge infrastructure, using this large computational problems are solved. In this context the efficient resource management techniques are required. Additionally unused resources are adding penalty over the cloud service providers. Thus the virtualization techniques are used for obtaining higher computational throughput. In this context multiple VMs are created over a single host or data center. But when the VMs are not working properly or damaged due to some reasons it is necessary to repair these deficiencies. Thus recognition of inactive VMs is required to improve the performance of cloud infrastructure. The proposed work involves the data mining techniques for this task. The data mining techniques are able to analyze the generated huge log files and can successfully identify the target VMs.

Basically data mining techniques enable us to accept data in bulk and produce the analysis outcome. The data mining techniques for this task usages the different kinds of computational algorithms for estimating the pattern similarities. Therefore different kinds of learning approaches are applied on data i.e. supervised learning and unsupervised learning. These algorithms are helpful for understanding the patterns and relationships among available data attributes. The proposed technique of VM classification aimed to improve the cloud computing productivity by recognizing the inactive VMs. Therefore a data mining technique is proposed for pattern identification. In order to experiment and system design the cloud server logs are used. That logs are used for identifying the inactive VMs. In this context the logs files are not containing the predefined data samples. Therefore to create class labels the k-means clustering algorithm is used. This algorithm help to create 2 different patterns groups. After pre-labeling of data the SVR (support vector regression) is used for final classification of inactive VMs.

The proposed system is implemented using the JAVA based technology. Java enable us to use different other technologies and libraries for implementing the required task. Additionally to manage performance data the MySql server is used. The performance of implemented techniques is provided in table 4.1.

S. No.	Parameters	Proposed SVR	Traditional SVM
1	Accuracy	High	Low
2	Error rate	Low	High
3	Memory usages	High	Low
4	Time consumption	High	Low

Table 4.1 performance summary

According to the obtained performance summary, as given in table 4.1 the proposed system provides the accurate classification or recognition as compared to traditional SVM based technique. The proposed SVR classifier based inactive VM recognition technique is suitable when accuracy is key aim not preservation of computational complexities. In near future it is tried to improve the method for time and memory usages.

B. Future Work

The aim of designing an accurate data model for identifying inactive VMs is accomplished successfully. In near future the following work is proposed for extension of the work.

1. The proposed technique usages the hybrid technique for classifying the data accurately, in near future the ensemble learning technique used for extending the work
2. The current data contains a limited attributes for identifying the inactive VM, in near future more attributes are collected for more precise estimation.

References

[1] In Kee Kim, Sai Zeng, Christopher Young, Jinho Hwang, Marty Humphrey, "A Supervised Learning Model for Identifying Inactive VMs in Private Cloud Data Centers", Middleware Industry '16, December 12-16, 2016, Trento, Italy c 2016 ACM. ISBN 978-1-4503-4664-1/16/12.

[2] Kainaz Bomi Sheriwala, "Data Mining Techniques in Stock Market", INDIAN JOURNAL OF APPLIED RESEARCH, Volume 4, August 2014.

[3] S. L. Pandhripande and Aasheesh Dixit, "Prediction of 2 Scrip Listed in NSE using Artificial Neural Network", International Journal of Computer Applications (IJCA), Volume 134, No.2, January 2016.

[4] Alexa Huth and James Cebula, "The Basics of Cloud Computing", © 2011 Carnegie Mellon University Produced for US-CERT, a government organization

[5] V. Bharath, D. Vijaya kumar, R. Sabarimuthu kumar, "AN EFFICIENT LOAD BALANCING ALGORITHM FOR CLOUD ENVIRONMENT", International Journal of Advance Research In Science And Engineering http://www.ijarse.com, IJARSE, Vol. No.4, Special Issue (02), March 2015

[6] Deepak Puthal, B. P. S. Sahoo, Sambit Mishra, and Satyabrata Swain, "Cloud Computing Features, Issues and Challenges: A Big Picture", 2015 International Conference on Computational Intelligence & Networks (CINE 2015)

[7] Deepika Khithani, Akshata Mohite, "Cloud computing-Future of IT", International Journal of Computer Science and Information Technologies, Vol. 7 (5) , 2016, 2290-2292

[8] "What is cloud computing? A beginner's guide", online available at: https://azure.microsoft.com/en-in/overview/what-is-cloud-computing/

[9] "The Benefits of Cloud Computing", available at: https://www.ibm.com/ibm/files/H300444G23392

G14/13Benefits_of_Cloud_Computing_634KB.pdf, Dynamic Infrastructure July (2009).

[10] Bhavani B H and H S Guruprasad, "Resource Provisioning Techniques in Cloud Computing Environment: A Survey", International Journal of Research in Computer and Communication Technology, Volume 3, Issue 3, March- 2014

[11] U. Sharma, P. J. Shenoy and S. Sahu, and A. Shaikh, "A Cost-Aware Elasticity Provisioning System for the Cloud", in Proc. International Conference on Distributed Computing Systems, July 2011, pp. 559-570.

[12] K. Tsakalozos, H. Kllapi, E. Sitaridi, M. Roussopoulous, D. Paparas, and A. Delis, "Flexible Use of Cloud Resources through Profit Maximization and Price Discrimination," in Proc of the 27th IEEE International Conference on Data Engineering(ICDE 2011),April 2011,pp.75-86.

[13] L. He, D. Zou, Z. Zhang, K. Yang, h. Jin and S. Jarvis, "Optimizing Resource Consumption in Clouds", in Proc. of the 12th IEEE/ACM International Conference on Grid Computing(Grid 2011), 2011, PP.42-49.

[14] Y. Hu, J. Wong, G. Iszlai, and M. Litoiu, "Resource Provisioning for Cloud Computing", in Proc. of these 2009 Conference of the center Studies on Collaborative Research, 2009, pp. 101-111.

[15] Petre, Ruxandra - Stefania, "Data mining in cloud computing", Database Systems Journal 3.3 (2012): 67-71.

[16] Sebastiani, F., "Machine learning in automated text categorization", ACM Computing Surveys, Volume 34, Number 1, 2002, pp. 1–47