IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.
www.ijiset.com
ISSN 2348 – 7968

Consummate Cluster Labeling for Spread Network

R. Shantha Mary Joshitta[1], M. Nithya[2] and P. Vidya[3]

[1] Dept. of Computer Science, Jayaraj Annapackiam College for Women (Autonomous), Periyakulam – 625 601, Theni Dt, Tamil Nadu, India. rjoshitta@gmail.com

[2]Dept. of Computer Science, Jayaraj Annapackiam College for Women (Autonomous), Periyakulam – 625 601, Theni Dt, Tamil Nadu, India. nithya.ish.28@gmail.com

[3]Dept. of Computer Science, Jayaraj Annapackiam College for Women (Autonomous), Periyakulam – 625 601, Theni Dt, Tamil Nadu, India. vidhya27.msc@gmail.com

Abstract

Cluster Labeling is a framework or processing large data sets and also used to do distribute computing on clusters of computer. These Cluster Labeling libraries have been written in many programming languages. Here we propose two novel algorithms for an efficient cluster labeling methods, Support vector (Parallel Support vector) and Mining methods (Parallel Mining methods), which facilitate simultaneous participation of multiple computing nodes to construct a boosted classifier. We analyze the problems in the existing system and try to solve it by using these two algorithms in our proposed system. In this paper, we outlined the procedure for applying the two cluster labeling methods and the impact of the methods. We used the Cluster Labeling framework to implement our algorithm and experimented on a variety of synthetic and real-world data sets to demonstrate the performance in terms of classification accuracy, speedup and scale up. By exploiting their own parallel architecture both the algorithms gain significant speedup. Due to the recent overwhelming growth rate of large-scale data, the development of faster processes algorithms with optimal performance as becomes a dire need of the time.

Keywords: *Cluster Labeling; Spread Network; Support Vector Algorithm; Mining Methods; Performance of Real-World Data Sets;*

1. Introduction

The use of Content Distribution Networks (CDNs) has emerged as a popular technique to improve client performance in client-server applications. Instead of hosting all content on a server (or a cluster of servers) at a single location, content providers today replicate their content across a collection of geographically distributed servers or nodes. Different clients are redirected to different servers both for load balancing and to reduce client-perceived response times. The most common redirection method is based on latency, i.e., each client is redirected to the server to which it has the lowest latency in order to reduce the download time to fetch the hosted content [1].

The explosive growth of the Web has imposed a heavy demand on networking resources and Web servers. Users often experience long and unpredictable delays when retrieving Web pages from remote sites. Content Distribution Networks (CDNs) offer users access to a wide variety of services, running on geographically distributed servers. Many web services are delay sensitive interactive applications (e.g., search) [2]. CDNs are quite vulnerable to increases in the wide-area latency between their servers and the clients, due to inter-domain routing changes or congestion in other domains. To detect and diagnose latency problems, CDNs could deploy a large-scale active-monitoring infrastructure to collect performance measurements from synthetic clients all over the world. Instead, this system explores how CDNs can diagnose latency problems based on measurements they can readily and efficiently collect—passive measurements of performance traffic, and routing from their own networks. Thus, in order to analyze the latency increases for groups of requests, we need to define the metrics to distinguish the changes from an individual router or server [3].

2. Problem Definition and Description

We present several events in greater detail to highlight the challenges of measuring and managing wide-area performance. We propose metrics that quantify the latency contributions across sets of servers and routers. Based on the design, we implement the LatLong tool for diagnosing large latency increases for CDN. After detecting large increases in latency, our classification must first determine whether client requests shifted to different front-end servers, or the latency to reach the existing servers increased [4]. Consummate Cluster Labeling for spread Network system will have the following subroutines. a) Initialize CDN b) Cluster Labeling Framework, c) File Download on CDN, d) Search on CDN and e) Admin Role.

ISSN 2348 – 7968

The working of these subroutines are explained briefly here.

2.1 Initializing CDN

Here, we initialize a networked computer system that harnesses the resources of several servers to complete the given tasks. It stores data is shaping up to be the next big trend in the computing industry. Since the debut of the personal computer, we've become used to storing information either on an external storage device like a compact disc or on a computer's hard drive. We're also conditioned to buy new machines or upgrade old ones whenever applications require more processing power than our current computers can provide[5]. With CDN, the responsibility of storage and processing power falls to the network, not the individual computer owner. The most obvious of these advantages is that the applications aren't tied to a specific computer. There's no need to download and install software on a particular machine. Any computer connected to the Internet can access our network. Users don't have to worry about which version of a document is the most current -- it will always be saved in our CDN. Users can upload file on server with high speed. It will be stored in many servers.

2.2. Cluster Labeling Framework

Cluster Labeling is a distributed programming system which is capable of processing large data sets. Map function and the Reduce function are the two main functions involved in Cluster Labeling Framework. User can also set the depth for the search performed in the nodes. The Map tasks are processed in parallel by the nodes in the cluster without sharing data with any other nodes.

2.3. File Download on CDN

We propose a decision tree for separating the causes of latency changes from their effects, and identify the data sets needed for each step in the analysis. Our tool LatLong can analyze latency increases and traffic shifts over sets of servers and routers. Once path performance is known, CDNs can optimize inter-domain path selection based on performance, load, and cost. We get the file with high download speed compare to other networks. Another advantage is that multiple users can make download to the same files at the same time. This is called online collaboration.

2.4. Search on CDN

User can search for documents or files using some of the standard CDN query parameters. The full-text query string is used to search the content of all the documents. User can search for resources matching an exact title or portion of a title by using the title-exact and title query parameters respectively. User can search the content of documents by using the query parameter on the feed. For example, to search a user's documents for the words "dog" and "cat".

2.5. Admin Role

Admin can manage all the files on the server. He can delete all files. Admin can view all the information about download and upload which was made by User. Admin can search for resources matching an exact title or portion of a title by using the title-exact and title query parameters respectively. Admin can search for documents or files using some of the standard CDN query parameters.

3. Algorithm Involved

3.1 Support Vector

The standard support vector algorithm ensembles learning method of iteration is strong classifier from a pool of weak hypotheses. In the Final iteration on the classifiers, a weighted linear combination of base classifiers are analyzed where each of them are entered in the array list. The classifiers with lower error rate are entered first in the array list which indicates higher weight.

3.2 Mining Methods

It is a powerful boosting method that uses regression functions instead of classifiers and these functions output real values in the same form as prediction. The base classifiers should have a higher rate than a random classifier. Simple decision trees with only one non-leaf node often perform well for SUPPORT VECTOR [6].

3.3 Support Vector Clustering (SVC)

In our Support Vector Clustering (SVC) algorithm, data points are mapped from data space to a high dimensional feature space using a Gaussian kernel. In feature space, we look for the smallest sphere that encloses the image of the data. This sphere is mapped back to data space, where it forms a set of contours which enclose the data points. These contours are interpreted as cluster boundaries. Points enclosed by each separate contour are associated

ISSN 2348 – 7968

with the same cluster. As the width parameter of the Gaussian kernel is decreased, the number of disconnected contours in data space increases, leading to an increasing number of clusters. Since the contours can be interpreted as delineating the support of the underlying probability distribution, our algorithm can be viewed as one identifying valleys in this probability distribution. SVC can deal with outliers by employing a soft margin constant that allows the sphere in feature space not to enclose all points. For large values of this parameter, we can also deal with overlapping clusters. In this range our algorithm is similar to the scale space clustering method of Roberts (1997) that is based on a Parzen window estimate of the probability density with a Gaussian kernel function.

3.4 Cluster Boundaries

We formulate a support vector description of a dataset that is used as the basis of our clustering algorithm.

Let $\{\mathbf{x}i\} \subseteq \chi$ be a data set of N points, with $\chi \subseteq \mathrm{IR}d$, the data space.

Using a nonlinear transformation Φ from χ to some high dimensional feature-space, we look for the smallest enclosing sphere of radius R. This is described by the constraints:

$$||\Phi(\mathbf{x}j) - \mathbf{a}||2 \leq R2 \ \forall j \ ,$$

where $||\ \cdot\ ||$ is the Euclidean norm and $\mathbf{a}$ is the center of the sphere.

Soft constraints are incorporated by adding slack variables ξj :

$$||\Phi(\mathbf{x}j) - \mathbf{a}||2 \leq R2 + \xi j \quad \text{with } \xi j \geq 0. \qquad (1)$$

To solve this problem, we introduce the Lagrangian

$$L = R2 - j$$
$$(R2 + \xi j - ||\Phi(\mathbf{x}j) - \mathbf{a}||2)\beta j - \xi j\mu j + C\,\xi j \ , \qquad (2)$$

where $\beta j \geq 0$ and $\mu j \geq 0$ are Lagrange multipliers, C is a constant, and $C\,\xi j$ is a penalty term.

Setting to zero the derivative of L with respect to R, $\mathbf{a}$ and ξj , respectively, leads to j

$$\beta j = 1 \qquad (3)$$
$$\mathbf{a} = j$$
$$\beta j\ \Phi\ (\mathbf{x}j) \qquad (4)$$
$$\beta j \ = \ C - \mu j \ . \qquad (5)$$

The KKT complementarity conditions of Fletcher (1987) result in

$$\Xi j\mu j \ = \ 0 \ , \qquad (6)$$
$$(R2 + \xi j - ||\Phi(\mathbf{x}j) - \mathbf{a}||2)\beta j = 0 \qquad (7)$$

It follows from Eq. (7) that the image of a point $\mathbf{x}i$ with $\xi i > 0$ and $\beta i > 0$ lies outside the feature-space sphere. Eq. (6) states that such a point has $\mu i = 0$, hence we conclude from Eq. (5) that $\beta i = C$. This will be called a *bounded support vector* or BSV.

A point $\mathbf{x}i$ with $\xi i = 0$ is mapped to the inside or to the surface of the feature space sphere. If its $0 < \beta i < C$ then Eq. (7) implies that its image $\Phi(\mathbf{x}i)$ lies on the surface of the feature space sphere. Such a point will be referred to as a *support vector* or SV.

SVs lie on cluster boundaries, BSVs lie outside the boundaries, and all other points lie inside them.

Note that when $C \geq 1$, no BSVs exist because of the Eq. (3).

Using these relations, we may eliminate the variables R, $\mathbf{a}$ and μj , turning the Lagrangian into the Wolfe dual form that is a function of the variables βj :

$$W = j$$
$$\Phi(\mathbf{x}j)2\beta j - i,j$$
$$Bi\beta j\ \Phi(\mathbf{x}i)\ \cdot\ \Phi(\mathbf{x}j). \qquad (8)$$

Since the variables μj don't appear in the Lagrangian, they may be replaced with the constraints:

$$0 \leq \beta j \leq C, j = 1, \ldots, N. \qquad (9)$$

We follow the SV method and represent the dot products $\Phi(\mathbf{x}i)\ \cdot\ \Phi(\mathbf{x}j)$ by an appropriate Mercer kernel $K(\mathbf{x}i, \mathbf{x}j)$.

We use the Gaussian kernel

$$K(\mathbf{x}i, \mathbf{x}j) = e^{-q||\mathbf{x}i - \mathbf{x}j||2} \qquad (10)$$

with width parameter q.

As noted in Tax and Duin (1999), polynomial kernels do not yield tight contours representations of a cluster. The Lagrangian W is now written as:

$$W = j$$
$$K(\mathbf{x}j\ , \mathbf{x}j)\beta j - i,j$$
$$\beta i\beta jK(\mathbf{x}i, \mathbf{x}j). \qquad (11)$$

At each point $\mathbf{x}$, we define the distance of its image in feature space from the center of the sphere:

$$R2(\mathbf{x}) = ||\Phi(\mathbf{x}) - \mathbf{a}||2 \qquad (12)$$

In view of Eq.(4) and the definition of the kernel we have:

$$R2(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}) - 2\ j$$
$$\beta jK(\mathbf{x}j\ , \mathbf{x}) + i,j$$
$$\beta i\beta jK(\mathbf{x}i, \mathbf{x}j)$$
$$(13)$$

The radius of the sphere is:

$$R = \{R(\mathbf{x}i) \mid \mathbf{x}i \text{ is a support vector } \} \qquad (14)$$

The contours that enclose the points in data space are defined by the set

$$\{\mathbf{x} \mid R(\mathbf{x}) = R\} \qquad (15)$$

They are interpreted by us as forming cluster boundaries (see Fig. 1 and Fig.3). In view of equation (14),

ISSN 2348 – 7968

SVS lie on cluster boundaries, BSVs are outside, and all other points lie inside the clusters.

4. Architecture of the System

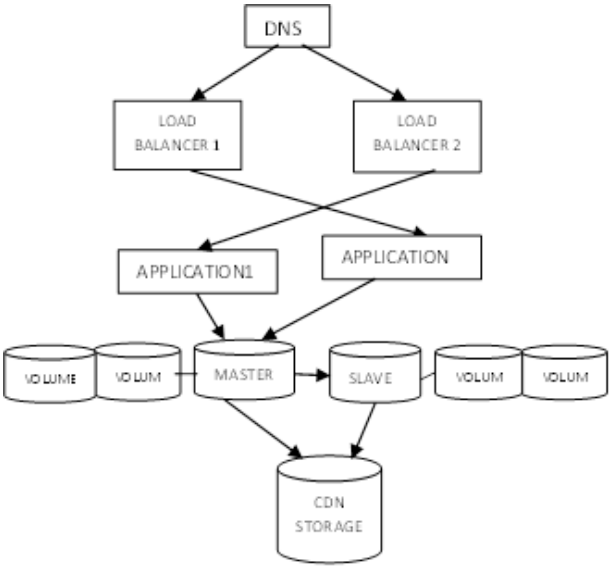

Fig.1: Architecture of the System

5. Implementation of the Algorithm

Step 1: To test the service operation we have to select the listfiles(),datacenter() and serverslist() and invoke them into our system. If we would like to find out the latitude and the longitude of the server, we can select the latlong() and enter the city that contains the server eg. Chicago. The result will be displayed.

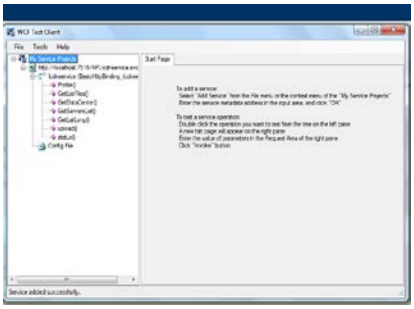

Fig.2: Getting services from the CDN

Step 2: To get the list of files binding with the current CDN we are selected, We can go for getlistfiles() operation. Now our CDN consists of 24 files with the current status.

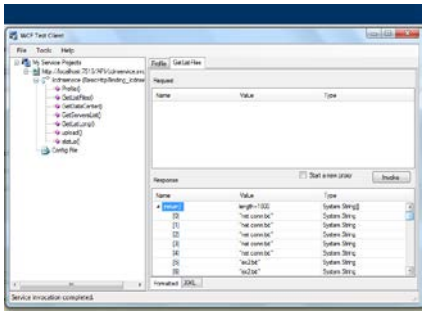

Fig.3: Getting List of Files

Step 3: To find out the status of a particular server in our CDN cluster, We can select status() option. In the below figure, we can easily identify whether a particular server is online or not.

Fig.4: Status of CDN Servers

Step 4: Finding the location of the servers in our CDN clusters. This figure will present you all time statics of the servers.

Fig.5: CloudFlare System Status – All time Statistics

ISSN 2348 – 7968

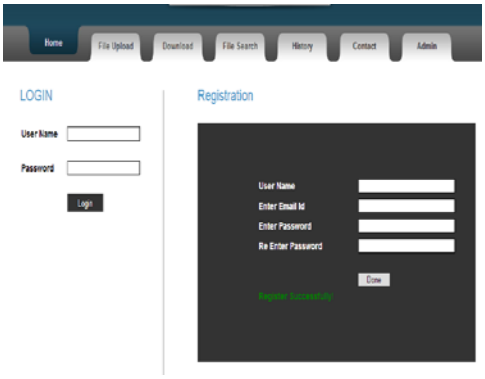

Fig.6: Registering on the cluster

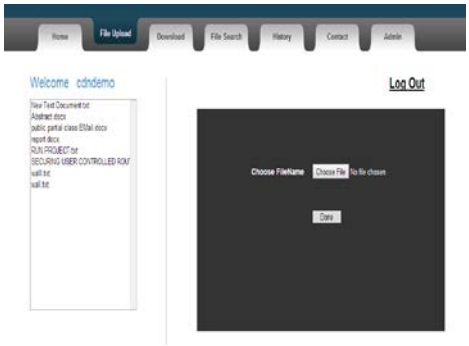

Fig.7: Uploading files into the Cluster

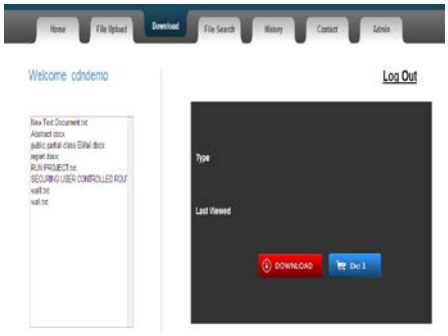

Fig.8: Downloading from the Cluster

6. Limitations

The direction in which this system can be enhanced is implementing this type of application in a widely connected environment. The scale up performance of our algorithms shows that they can efficiently utilize additional resources when the problem size is scaled up. In the future, we plan to explore other data partitioning strategies that can improve the classification performance even further [8].

7. Conclusion

In this paper, we explained the investigation of the problem of data security in distributed network sharing, which is essential to a distributed storage system. The searching process is boosted with efficient algorithm. We prove the efficiency and protect the ability of the clients. The direction in which we have moved our system is to avoid or prevent the data losses and also maintaining the updates in the server for all events which have took place in the client side.

Acknowledgment

We would like to record our thanks Sans Pareil IT Services Private Limited, Chennai for providing financial support to complete this project.

References

[1] Rupa Krishnan, Harsha V. Madhyastha, Sridhar Srinivasan, Sushant Jain, Arvind Krishnamurthy, Thomas Anderson, Jie Gao, “Moving Beyond End-to-End Path Information to Optimize CDN Performance”, http://research.google.com/pubs/pub35590.html

[2] George Pallis, Athena Vakali, Konstantinos Stamos, Antonis Sidiropoulos, Dimitrios Katsaros, Yannis Manolopoulos, “A Latency-based Object Placement Approach in Content Distribution Networks”, IEEE Proceedings of the Third Latin American Web Congress (LA-WEB’05), 2005, 0- -7695-2471-0/05, pp 190-201.

[3] Yaping Zhu, Benjamin Helsley, Jennifer Rexford, Aspi Siganporia, and Sridhar Srinivasan, “LatLong: Diagnosing Wide-Area Latency Changes for CDNs”, IEEE Transactions on Network and Service Management, 1932-4537/12, 2012.

[4] C.Chandravathi, Fauzia Begam. K, Shamili. A, “Verdiction of Time Delay in Wide-Area CDNs”, 2nd National Conference in Emerging Trends in Informative Computing Applications (IWAY) – 2012, Proceedings published by International Journal of Computer Applications® (IJCA)

[5] Strickland, Jonathan, "How Google Docs Works", HowStuffWorks.com. http://computer.howstuffworks.com/internet/basics/ google-docs.htm, 02 June 2008.

[6] Indranil Palit, Chandan K. Reddy, “Scalable and Parallel Boosting with MapReduce”, IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 10, pp. 1904-1916, Oct. 2012, doi:10.1109/TKDE.2011.208.

[7] Geetha Ramani R, Jacob SG, “Prediction of P53 Mutants (Multiple Sites) Transcriptional Activity Based on Structural (2D&3D) Properties”, PLoS ONE 8(2): e55401. doi:10.1371/journal.pone.0055401, 2013

[8] Cloud Security Alliance, “Security guidance for critical areas of focus in cloud computing in 3.0”, 2011.

ISSN 2348 – 7968

First Author completed her M.C.A in Bharadhidasan University. In 2001, she joined as a lecturer in Jayaraj Annapackiam College for Women (Autonomous) and now she is working as the controller of Examination of the college and Assistant Professor in the Computer Science department. She attended many conferences at national and international levels, workshops and seminars and presented many papers. Her areas of interest are Big data Analytics, ICT, e-learning, Green computing and Medical Image Processing. Her poster entitled as "Green technology: Green light for India's growth" presented in the International Conference on Emerging Trends in Computer Science, Communication and Information Technology organized by department of Computer Science and Information Technology, Yeswanth Mahavidhyalaya, Nanded, Maharashtra won first prize with a cash award of Rs. 1500/- . So far, she has published 4 papers in reputed journals. She served as the member of board of studies in two colleges and published many course materials. She is a member of the IEEE.

Second Author is a student of Master degree in Computer Science and Information Technology. She is the receiver of PG Merit Scholarship from UGC for the University Rank Holders for the year 2013 – 2015. She already published one paper in a journal.

Third Author is also a student of Master degree in Computer Science and Information Technology. She already published one paper in a journal.