ISSN: 2320-1363

CONTEXTUAL ADVERTISEMENT MINING BASED ON BIG DATA ANALYTICS

A.Divya[*1], A.M.Saravanan[*2], I. Anette Regina[*3]

MPhil, Research Scholar, Muthurangam Govt. Arts College, Vellore, Tamilnadu, India

Assistant Professor, Department of CS, Muthurangam Govt. Arts College, Vellore, Tamilnadu, India

Associate Professor, Department of CS, Muthurangam Govt. Arts College, Vellore, Tamilnadu, India

ABSTRACT

With the advance of Internet and wireless network technologies, to build and support business centred big data analytic service is becoming a very hot topic. Big data analytic solutions and services are needed to support business people to perform business market analysis and make intelligent decisions. This project focuses on subject relating to big data analytic services on advertising. The primary objective is to develop a web-based data analytics service framework and system to support business users to perform advertising analysis based on available big data resources and web mining techniques.

INTRODUCTION

Behavioural targeting (BT) is a widely used technique for online advertising. It leverages information collected on an individual's web-browsing behaviour, such as page views, search queries and ad clicks, to select the ads most relevant to user to display. With the proliferation of social networks, it is possible to relate the behaviour of individuals and their social connections. Although the similarity among connected individuals are well established (i.e., homophily), it is still not clear whether and how we can leverage the activities of one's friends for behavioural targeting; whether forecasts derived from such social information are more accurate than standard behavioural targeting models. In this paper, we strive to answer these questions by evaluating the predictive power of social data across 60 consumer domains on a large online network of over 180 million users in a period of two and a half months. To our best knowledge, this is the most comprehensive study of social data in the context of behavioural targeting on such an unprecedented scale. Our analysis offers interesting insights into the value of social data for developing the next generation of targeting servicesThis section presents the

ISSN: 2320-1363

former literature work done on the subject and also concepts and materials helping for clear understanding, progress and completion of the project. The purpose of this literature survey is to provide background information on the issues to be considered in this thesis and to emphasize the relevance of the present study.

Random Forest algorithm based on MapReduce framework

There are many problems that can be mapped to a MapReduce program, such as: sorting, searching, indexing and classification. These programs must fit the features of the MapReduce algorithm. For any MapReduce algorithm, processing data must go through a map phase and a reduce phase. With consideration that the output of the map phase as the input to the reduce phase. As in [1] Random Forest is a classification and regression method based on the aggregation of a large number of decision trees. Specifically, it is an ensemble of trees constructed from a training data set and internally validated to yield a prediction of the response given the predictors for future observations. There are several variants of RF which are characterized by 1) the way each individual tree is constructed, 2) the procedure used to generate the modified data sets on which each individual tree is constructed, 3) the way the predictions of each individual tree are aggregated to produce a unique consensus prediction.In the original RF method suggested by Breiman et al , Random forest is described as an ensemble classifier that consists of many decision trees. The method to build the random forest is bagging. Given the independent variable X, Random Forest Classification is an ensemble classification model which combined of K decision tree classifiers hl(X),hiX), ... hdX) [2] . Each of the classifier decision trees votes for one of the classifications and the winner is the final classification results. The main work step of Random Forest Classification: Firstly, select K samples from the original training dataset using bagging method randomly. Then, the K samples will be the training set for growing K trees accordingly to achieve the K classification results . Finally, the K classifiers vote to elect the optimal classification with majority. An important feature of RF is its out-of-bag (OOB) error. Each observation is an OOB observation for some of the trees, i.e. it was not used to construct them and can thus be considered as an internal validation data set for these trees. The OOB error of the RF is simply the average error frequency obtained when the observations from the data set are predicted using the trees for which they are OOB. Through this internal validation, the error estimation is less optimistic and usually considered as a good estimator of the error expected for independent data.

ISSN: 2320-1363

In this project an improved scalable Random Forest algorithm based on Map Reduce model is used [2]. This new algorithm makes data classification in computer cluster or cloud computing environment for massive datasets. SMRF processes and optimizes the subsets of the data across multiple participating computing nodes by distributing. SMRF algorithm is more suitable to classify massive data sets in distributing computing environment than traditional Random Forest algorithm.

MapReduce Programming Framewor

In this project we employ mapreduce programming framework . As in [3] MapReduce is a programming model for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs and a reduce function that merges all intermediate values associated with the same intermediate key .**"Map" step:** The master node takes the input, partitions it up into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller problem, and passes the answer back to its master node. Map takes one pair of data with a type in one data domain, and returns a list of pairs in a different domainMap (k1, v1) → list (K2, v2)**"Reduce" step:** The master node then collects the answers to all the sub-problems and combines them in some way to form the output – the answer to the problem it was originally trying to solve. The Reduce function is then applied in parallel to each group, which in turn produces a collection of values in the same domain:

Reduce (K2, list (v2)) → list (v3)

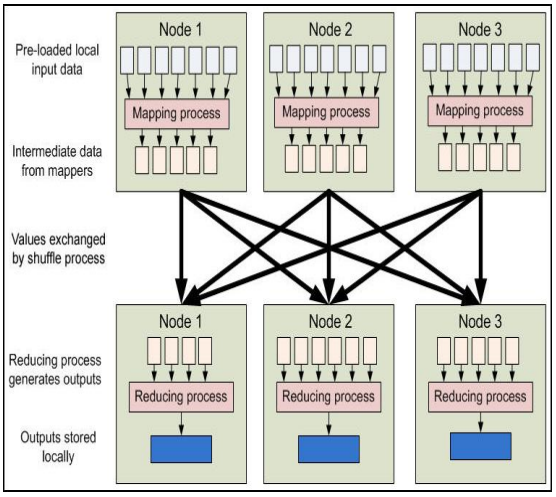


Figure 1.1 Distributed Map and Reduce processe

Google App Engine GAE hosts Web applications on Google's large-scale server infrastructure. It has three main components: scalable services, a runtime environment, and a data store. GAE's front-end service handles HTTP requests and maps them to the appropriate application servers. Application servers start, initialize, and reuse application instances for incoming requests. During traffic peaks, GAE automatically allocates additional resources to start new instances. The number of new instances for an application and the

ISSN: 2320-1363

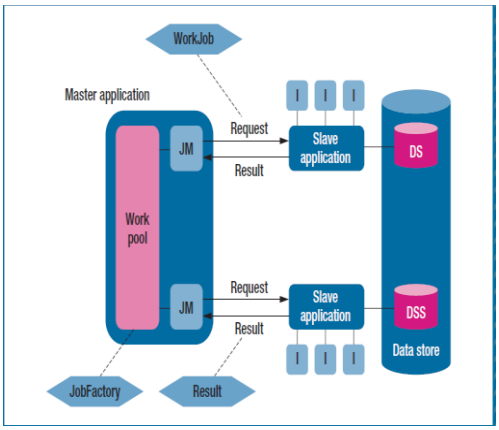
distribution of requests depend on traffic and resource use patterns. So, GAE performs load balancing and cache management automatically. Each application instance executes in a sandbox (a runtime environment abstracted from the underlying operating system). This prevents applications from performing malicious operations and enables GAE to optimize CPU and memory utilization for multiple applications on the same physical machine. Sandboxing also imposes various programmer restrictions:

• Applications have no access to the underlying hardware and only limited access to network facilities.

• Java applications can use only a subset of the standard library functionality.

• Applications can't use threads.

• A request has a maximum of 30 seconds to respond to the client.



Our parallel computing framework architecture. The boxes labeled I denote multiple slave instances. The master application is responsible for generating and distributing the work among parallel slaves implemented as GAE Web applications and responsible for the actual computation.

1.3 Motivation

In the era of Big Data, characterized by the huge volume of data, the velocity of data generation, and the variety of the structure of data, support for large-scale data analytics constitutes a particularly challenging task. To address the scalability requirements of today's data analytics, parallel shared nothing architectures of commodity machines have been lately established as the de-facto solution. Various systems have been developed mainly by the industry to support BigData analysis. Several companies, including Facebook , both use and have contributed to big data analysis and development of mapreduce. MapReduce has become the most popular framework for large-scale processing and analysis of vast data sets in clusters of machines, mainly because of its simplicity. With MapReduce, the developer gets various cumbersome tasks of distributed programming for free without the need to write any code, indicative examples include machine to machine communication, task scheduling to machines, scalability with cluster size, ensuring availability, handling failures, and partitioning of input data. In this project we use the approach of scalable random forest algorithm based on MapReduce framework. we calculate the total

ISSN: 2320-1363

electrical consumption for houses and further classify it based on the values available in dataset in a very effective manner.

1.4 Problem Statement

The goal of this project is to classify the data using scalable random forest approach and also to make load forecasts based on the current load measurements and a model that was learned over historical data.Such forecasts can be used in demand side management to proactively influence load and adapt it to the supply situation, e.g., current production of renewable energy sources.

BIG DATA ANALYTICS

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. Big Data refers to the massive amounts of data that collect over time that are difficult to analyze and handle using common database management tools. The data are analyzed for marketing trends in business as well as in the fields of manufacturing, medicine and science. The types of data include business transactions, e-mail messages, photos, surveillance videos, activity logs and unstructured text from blogs and social media, as well as the huge amounts of data that can be collected from sensors of all varieties. Nowadays, internet has become the main media for advertising. It gives the opportunity to advertise a product which can reach huge amounts of people in a relatively small cost. It introduces a challenge about how to settle the conflict of interests by selecting advertisements that are relevant to the users but also profitable to the advertisers and the publishers. The main form of web advertising is contextual advertising (ad). It involves four parties: publisher, advertiser, ad-network, and user. The publisher is the owner of a web page on which ads are displayed. The advertiser provides ads to promote their products and services. The ad network selects the best ads to place in a web page and thus acts as a mediator between the advertiser and the publisher. The user is an Internet viewer that views the web page. An ad is selected to place in the target page based on its relevance to the page content. When the user opens the Web page, he/she sees the ad as a short textual description and a hyperlink that,

ISSN: 2320-1363

once clicked, takes the user to the ad landing page.

PROPOSED SYSTEM

To place the commercial ads within a webpages based on content in the webpages using big data analytics with Hadoop Mapreduce framework. To design and implement algorithms to adapt large-scale data processing in distributed manner for advertisement generation. To analyse learning Influence Probabilities using Behavioural targeting for mining advertiser-specific user behaviour via ad factors. To construct user profiles from large scale data. To categorize data into clusters such that ads are grouped in the same cluster when they are similar according to specific metrics.

Results and discussion

The temporal-analytics-temporal-data characteristic is observed for many "big data" applications such as behavioural targeted Web advertising, network log querying, and collaborative filtering. This paper proposes the use of temporal queries to write such applications, as such queries are easy to specify and naturally real-time-ready. We propose a framework called TiMR that enables temporal queries to scale up to massive offline datasets on existing M-R infrastructure. We validate our approach by proposing a new end-to-end solution using temporal queries for BT, where responsiveness to user interest variation has high value. Experiments with StreamInsight and SCOPE/Dryad using real data from our ad platform validate the scalability and high performance of TiMR and its optimizations, and the benefit of our BT approach in effective keyword elimination, lower memory usage and learning time, and up

CONCLUSION

Big data is effectively and efficiently captured, processed, and analyzed which can lead to efficiency improvements, increased sales, lower costs, better customer service, and/or improved products and services for advertisement mining ,also it isa new trend in global economy.It is also flexible and general enough to be applied in a multilingual environment, different domains and large data sets. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises.

REFERENCES

[1]Badrish Chandramouli, Jonathan Goldstein, Songyun Duan,"Temporal analytics on big data for web advertising"*IEEE 28th International Conference on Data Engineering,2012.*

[2]Amrit Pal,Pinki Agrawal, Kunal Jain, Sanjay Agrawal,"A Performance Analysis of MapReduce Task with Large Number of Files

ISSN: 2320-1363

Datasetin Big Data Using Hadoop"*IEEEFourth International Conference on Communication Systems and Network Technologies 2014.*

[3]Jyoti Nandimath,Ankur Patil, Ekata Banerjee, Saumitra Vaidya,"Big Data Analysis Using Apache Hadoop"*IEEE IRI 2013, San Francisco, California, USA,*on *14-16 Aug 2013.*

[4].Do Viet Phuong , Tu Minh Phuong, "A Keyword-Topic Model for Contextual Advertising" *Proceedings of the Third Symposium on Information and Communication Technology,* Pages 63- 70 , New York, 2012.

[5]Aditya B. Patel, Manashvi Birla, Ushma Nair"Addressing Big Data Problem Using Hadoop and Mapreduce*" 2012 Nirma University International Conference on Engineering,Nuicone-2012, 06-08december, 2012.*