International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.5, pp : 597-600 1 May 2014

Churn Prediction using MAPREDUCE

S. Ezhilmathi Sonia, Prof. S. Brintha Rajakumar, Prof. Dr. C. Nalini

Department of CSE, B. I. S. T. (Bharath University), Chennai, TN, India
ezhilmathisonia@yahoo.com, brintha.ramesh@gmail.com, DrNaliniChidambaram@gmail.com

ABSTRACT – ***The data mining process to identify churners has concern with size of the dataset. This paper analyzes the telecom customer complaints and call quality datasets using Mapreduce to predict the customer churn. HDFS and Mapreduce make it possible to mine larger data sets without the constraints of the data size.***

Keywords – Hadoop MapReduce, Telecommunication, Churn Analysis, Data mining

I. Introduction

Telecommunications companies create enormous amounts of data every day. The data include call detail data that describes the calls that traverse the telecommunication networks, network data, which describes the state of the hardware and software components in the network, and customer data, which describes the telecommunication customers.

With the rapid expansion of data, the data storage has moved to Petabyte Age. At the same time, new technologies make it possible to organize and utilize the massive amounts of data currently being generated. This paper is about telecommunication customer churn analysis using HDFS (Hadoop Distributed File System) and Mapreduce based on the customer satisfaction and call quality details.

Churn, which is defined as the loss of customers to another company, is a crucial problem in the telecommunication industry. As the telecom market has matured and opportunities for growth are limited, retaining existing customers has become a higher priority. Ultimately churn occurs because the customers are dissatisfied with the quality of service, usually as compared to competing industry.

The open source Hadoop system consists of HDFS and Mapreduce. MapReduce infrastructure has provided data mining researchers with a simple programming interface [i] for parallel scaling up of many data mining algorithms on large data sets.

Motivation

Telecommunication data pose several interesting issues [vi] for data mining.

a) Telecommunication databases may contain billions of records and are amongst the largest in the world.
b) The raw data is often not suitable for data mining.

The extraction of the raw data from the huge data files and summarization for the data mining is the time consuming tasks. Data mining using Hadoop mapreduce [i] will provide high performance, reliable and fault- tolerant framework for the processing of vast amount of data. Hadoop handles variety of unstructured, structured or semi structured data. Hadoop mapreduce programming used for customer classification data mining will provide better performance than any other data mining applications. The result of this implementation using mapreduce programming will provide rapid decision making for the business marketing and customer churn analysis.

This paper focuses on the customer relation data and call quality data to identify the customers who may or may not leave the company service. The customer dissatisfaction [iii] due to the call quality affect customer churn. The analysis of these datasets will help to understand customer satisfaction with the company services. The dissatisfaction indicators such as number of complaints and call drop rate due to poor connectivity quality have significant impact on the customer churn.
Next, the results and analysis of the experiments are discussed in detail.
This paper is structured as follows: In section II, an overview of related works in data mining using the mapreduce and the proposed architecture, data model, the rule based methodology and experimental set up used are described. In section III, the data sets being considered and the results obtained based on the sample dataset has been discussed in detail. Finally conclusions are drawn and future work is outlined in section IV.

II. Material and Methodology

The MapReduce techniques will tackle the problems in the large scale data mining domain. Taking advantages of the Mapreduce, it is very clear that the MapReduce framework provides good performance with respect to scalability, reliability and efficiency [ii,v]. Utilizing MapReduce is one of major methods along with sampling and optimal algorithm design in large scale and parallel data mining. MapReduce has been intensively used for large scale machine learning problems at Google Inc. Other researchers have used MapReduce in many data mining applications. Panda et al presented a framework using a paralleling decision tree algorithm with MapReduce. These researches confirm that the Mapreduce will provide better performance and easily processes the massive datasets used for data mining. Hence the churn analysis using the data mining technique can be used to extract hidden predictive information in the huge telecommunication data.

Ali [iii] has discussed about the major features that can be utilized in order to build a predictive model for customer churn. The attrition of the customers in the telecommunication industry depends on the level of satisfaction with alternative

International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.5, pp : 597-600 1 May 2014

specific service attributes including call quality and the customer relationship management. The customer dissatisfaction, customer status, switching costs and service usage are the major factors that can be considered to develop efficient predictive model for telecom churn prediction.

Proposed Work

In the proposed scheme using hadoop mapreduce will resolve two major issues of the telecommunication data mining.

Architecture

The architecture for this implementation is illustrated below. Following are components:

1. Input Data files - Customer Complaints, Call Quality
2. Hadoop Framework – HDFS and MAPREDUCE
3. Output Data file - Churners list

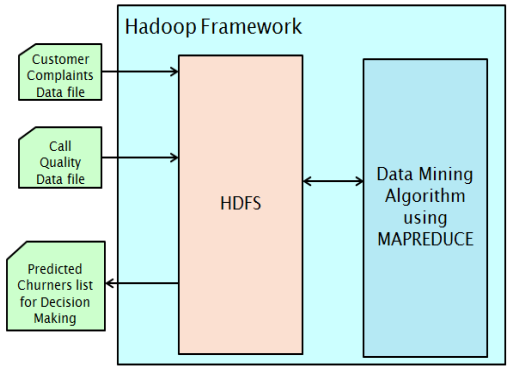

Figure 1: Architecture

Methodology used

Data mining using the mapreduce can be applied in big dataset.

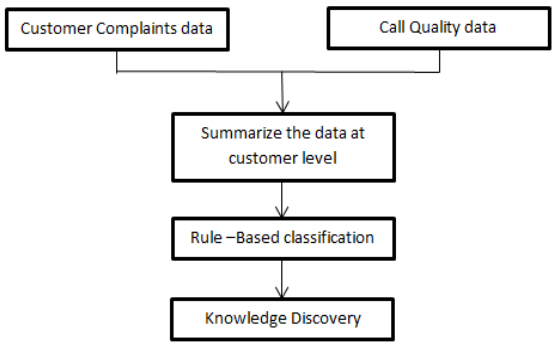

Figure 2: Block Diagram

Below are the attributes considered for the customer complaints entity:

Customer Complaints Entity
YearMonth(YYYYMM)
Week Number
Customer Id
Count of CS complaint calls
Count of CS complaint emails

Figure 3: Customer Complaints Entity

Below are the attributes considered to know the call quality entity:

Call Quality Entity
Customer Id
YearMonth(YYYYMM)
WeekNumber
Total Incoming calls
Total Incoming call dropped
Total Outgoing calls
Total Outgoing connectivity dropped
% of failed In calls
% of failed Out calls

Figure 4: Call Quality Entity

The rules that have been defined are mentioned below.

- CustomerComplaints
  TotalNumberofComplaints > 5 = Yes
  TotalNumberofComplaints <= 5 = No
- CallsFailures
  Average%ofCallFailures > 30% = Yes
  Average%ofCallFailures <= 30% = No
- Prediction
  - Churners
    CustomerComplaints = Yes AND
    CallFailures = Yes
  - Non-Churners
    CustomerComplaints = Yes OR
    CallFailures = No = Non-Churners

Experimental Set up:

Hadoop [8] is a powerful framework for automatic parallelization of computing tasks. Java 1.6 and Eclipse Europa 3.3.2 are the prerequisites for the installation and working of the Hadoop[9]. Cygwin is a set of Unix packages ported to Microsoft Windows 7.0. It is needed to run the scripts supplied with Hadoop because they are all written for the Unix platform. Hadoop 0.20.2 is configured with Eclipse. For HDFS nodes we have the NameNode, and the DataNodes. For MapReduce nodes we have the JobTracker and the TaskTracker nodes. The Hadoop namenode is formatted to

International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.5, pp : 597-600 1 May 2014

create the Hadoop Distributed File System. Once all the nodes are started, Hadoop cluster will be operational.

III. Results and tables

The data model and input datasets has been created self to predict telecom churners. The telecommunication customer complaints and the call quality datasets has been considered for the churn analysis. The Customer Service and Relationship Management team can provide the telecom customer complaints data. The Network Connectivity Maintenance team can provide the call quality data being tracked for every incoming and outgoing call.

The sample data considered for last quarter of the year 2013 has been self-generated for the analysis. The snapshot of the customer complaints sample data is given below.

Year Month	Week No.	Customer Id	Count of CS complaint calls	Count of CS complaint emails
201310	w1	1110012345	10.00	4.00
201310	w1	1110012345	3.00	2.00
201311	w2	1110067890	5.00	3.00
201312	w3	1110067891	3.00	8.00
201310	w4	1110056565	0.00	5.00
201311	w1	1110056565	1.00	3.00
201312	w2	1110034543	1.00	3.00
...	...	...	...	...

Table 1: Customer Complaints sample data

The snapshot of the call quality sample data is given below.

Customer ID	Year Month	Week No.	Total In calls	Total In calls dropped	Total Out calls	Total Out calls dropped	% Failed In Calls	% Failed Out Calls
1110012345	201306	w1	40	20	20	10	50	50
1110012345	201307	w1	45	25	30	25	56	83
1110067890	201308	w2	4	2	4	2	50	50
1110067891	201309	w3	20	10	20	5	50	25
1110056565	201310	w4	25	15	25	15	60	60
1110056565	201311	w1	3	3	3	3	100	100
1110034543	201312	w2	3	3	3	2	100	67
...	...	...	...	...	...	...	...	...

Table 2: Call Quality sample data

The input datasets will be copied to the HDFS. The mapreduce program will read the data from the files placed in HDFS. The mapreduce program will perform the following and produces corresponding output datasets:

1. Summarization of the customer complaints via calls and email and group the customers having total no. of complaints greater than 5.

2. Calculate the average of the % of the Incoming and Outgoing call failures and group the customers having the Average % of failures greater than 30.

3. Apply the defined rules on the merged datasets and classify the customers as churners and non-churners.

The Map function takes a pair of < key; value > as input, and emits intermediate < key; value > pairs to the Reduce function. The reduce function processes all intermediate values associated with the same intermediate key. The Customer ID is set as the key for both the datasets.

Below is the sample of the output dataset generated in HDFS:

```
1110010012      N
1110010018      N
1110012340      Y
1110012341      N
1110012342      N
1110012343      N
1110012345      Y
1110012378      N
1110013210      N
1110013456      N
1110030023      N
1110034543      Y
1110036789      Y
1110037654      N
1110050014      N
1110050015      N
1110050016      N
```

Figure 4: Churners & Non-Churners list

The customers flagged as Y are predicted as churners and the customers flagged as N are predicted as non-churners.

IV. Conclusion

This paper helps to predict churn of telecom customers using Mapreduce. With better understanding of customers, organization can develop a customized approach to customer retention activities. In future, Mapreduce can be used to implement and explore new methodologies for churn prediction by integrating with different data mining techniques and data sets.

Acknowledgement

We thank the Management, the Principal/Director, Professors of B.I.S.T (Bharath University), Chennai, TamilNadu, India for encouraging us for this journal.

References

i. *N.Kamalraj, Dr.A.Malathi,[IJARCSSE] Applying Data Mining Techniques in Telecom Churn Prediction, Oct 2013*

ii. *Amy Xuyand Tan, Valerie Li Liu, Murat Kantarcioglu and Dr. Bhavani Thuraisingham "A Comparison of Approaches for Large-Scale Data Mining" (Utilizing MapReduce in Large-Scale Data Mining), The University of Texas at Dallas, Technical Report UTDCS-24-10 August 2010.*

iii. *Ali Tamaddoni Jahromi, Lulea University of Technology, Predicting Customer Churn in Telecommunications Service Providers, 2009: 052 - ISSN: 1653-0187*

iv. *Dragana Camilovic, Dragana Becejski-Vujaklija and Natasa Gospic, "A Call Detail Records Data Mart: Data Modelling and OLAP Analysis" , December 2009*

v. *http://bigdatauniversity.com/*

vi. *Data Mining in Telecommunication: http://storm.cis.fordham.edu/~gweiss/papers/kluwer04-telecom.pdf*

vii. *http://www.researchgate.net/publication/225904834_Data_Mining_in_Telecommunications/*

viii. *Hadoop: http://hadoop.apache.org/*

ix. *Installation and Hadoop configuration: http://v-lad.org/Tutorials/Hadoop/00%20-%20Intro.html*

x. *D.Gillick, A.Faria, and J.DeNero. Mapreduce:Distributed computing for machine learning*

xi. *B. Panda, J. S. Herbach, S. Basu, and R. J. Bayardo.Planet: Massively parallel learning of tree ensembles with mapreduce. Lyon, France, 2009. ACM press*

xii. *http://stackoverflow.com*