Current trends of data mining applications in telecom industry

N.P. Singh
Professor (Information Technology Management)
Management Development Institute,
Gurgaon -122001, India
E-mail: npsingh@kimep.kz & knpsingh@mdi.ac.in

ABSTRACT

The telecommunication companies have large number of potential customer due to wide public access of their services and with the invention on web / internet offer large set of the services. It resulted in increasing the complexities of their network and generation of huge stream of data that changes dynamically with each usage of their services by the customers across the world. The companies are employing on line transaction processing (OLTP) systems, data marts, data warehouse and on line analytical processing (OLAP) systems to store their data. In addition, for efficient management of their business data mining tools, business intelligence tools and analytics are implemented by major telecom companies in all part of the world. These tools are basically used to detect frauds, forecasting the growth of networks, risk management, customer churn management, market segmentation, service usage patterns etc. The paper presents, the growth of data mining tools, their usage in telecommunication sector for customer churn management, fraud detection, market segmentation, etc.

Key Words: Data Mining, Telecommunication, Customer Churn Management, Fraud Detection, Market Segmentation

Introduction:

Telecommunication networks are growing fast in size, usage, and complexities making their management with traditional tools more difficult. In general telecommunication services include electronic communication services, subscription computer services, telephone and telegraph services and all other services that involve the transmission of information by wire, radio, cellular, wireless transmission or similar means. The major telecommunication services are telephone (phone plus service, new telephone connection, permanent connection, concession in rentals, shifting of telephones, transfer of telephones, telephone tariff, Caller ID), mobile phones (Unified Messaging, Enhanced Data GSM Environment (EDGE), General Packet Radio Service (GPRS)/ Wireless Application protocol (WAP) / Multimedia Message Service (MMS), tariff, Universal Mobile Telecommunication Services(UMTS(3G)) , Wireless in Local Loops (WLL), Internet ( Web hosting, various products of internet connectivity, Co-location Services), Broadband Services, Multi-protocol label switching (MPLS) Virtual Private Network (VPN), integrated services digital network (ISDN), Intelligent Network (IN- Free Phone Service, Premium Rate Service, telephone cards, Virtual Private Network, Universal Number, Tele

1 Presently working as Professor (IS), Bang College of Business, Kazakhstan Institute of Management, Economics, and Strategic Research, 2, Abai Avenue, Almaty-480100, Kazakhstan
voting), I-Net, Telex/ Telegraph, EPABX, High Speed Satellite based VSAT (Very Small Aperture Terminal) Network, Broadband Services, Toll Free Services, Voice over Internet Protocol (VOIP), Onsite billing with real time monitoring, invoice generation and history report, Voice Mail, Fax, Pager, E-mail, transfer of Computer & Web Data and other data traffic, Short Messing System (SMS), File Transfer, Roaming, etc. One of the major sources of data for a telecom operator is Call Detail Record (CDR) which normally consists of Call Type, Generating System, Recording Office, Connect Date, Timing Indicator, Connect Time, Elapsed Time, GMT Time Zone, Release Cause (Optional), Source Transport Network Address, Source Logical Port Identifier (Optional), Destination Transport Network Address, Destination Logical Port Identifier (Optional), Local Connection Identifier, Encoding Type, Traffic Parameters, Directionality, Generalized Payload Identifier, Service Level, Diversity, Contract ID (Optional).

In a telecommunication company hundreds of millions of CDRs are generated daily and requires collection and mining of CDRs on a continuous basis (To mention: Tele Denmark collects and maintains more than five Terabytes of customer traffic data). The data volumes and data flow rates pose serious scalability and performance challenges. To overcome these problems companies are implementing scalable data mart, operational data stores (ODS), data-warehouse (DW), multidimensional data bases and OLAP tools. Longbing et al (2004) suggested ontology service based integration of DW, OLAP and DM for better performance. These implementations are taking care of issue of scaling the whole operation chain, including data cleansing, loading, maintenance, access and analysis using sophisticated data mining tools.

Data mining tools are used in telecommunication sector to develop strategies for identifying market trends, detecting key characteristics and patterns for market segments, improving the quality of products and services offered, detecting fraud and insolvency early enough and focusing on customers which are likely to stay with the company longer, are which are profitable [Mozer et. al. (2000)]. In addition, data mining tools are used for identifying patterns of behavior of different groups of users, multidimensional analysis, optimization of resources, predicting peak loads and network growth rate, usage of service, etc. For the purpose data from applications such as billing, marketing, sales, fraud management, performance analysis systems, network switches and customer service across the company. The major data mining tools/ algorithm used for various applications in telecommunication are presented in Table 1. Commercial and public domain data mining software’s (Suite, Comprehensive software’s) (Singh and Gupta (2004)) are embedded with sophisticated algorithm to analyze the ever increasing complexities of telecommunication systems. An analysis of this trend is presented in this paper.

References:

**Table 1: Data Mining Tools & their application in Telecommunication**

<table>
<thead>
<tr>
<th>Tools</th>
<th>Type of algorithm &amp; Application in Telecommunication</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Association</strong></td>
<td>Apriori Algorithms (market basket analysis, mining association rules dedicated to estimation and prediction for call tracking data; Sequential Pattern Discovery for telecommunication alarm logs; hash Tree for fast Access; Distributed/ Parallel Algorithms: Parallel Algorithm MLFPT (Multiple Local Frequent Pattern Tree), Dynamic Hashing &amp; Pruning Algorithm (DHP), DIC (Dynamic Item set Counting) Algorithm, Partition Algorithm, SEAR and SPEAR Algorithm, Eclat, MaxEclat, Clique, MaxClique. Web Personalization using Association rules for click stream data. The other popular algorithm are FP Tree Algorithm and Rapid Association Rule Mining (RARM) with SOTrieIT Structure</td>
</tr>
<tr>
<td><strong>Visualization &amp; Link Analysis</strong></td>
<td>Graphical: (using STRUCTURE MINER with Algorithms such as MAJORCLUST, Hierarchical MAJORCLUST, Kohonen, Mincut for analysis of network traffic; SeeNet for visualizing network &amp; its data; Redirection rule for transferring calls; Mineset’s Splat Visualizer, Scatter Visualizer, and map visualizer for Network Analysis, Intrusion Detection, and fraud detection; TGS’s Customer Behavior Analysis (CBA) provide 3D visualization techniques based on an open inventor to map data attributes in to visual dimensions. Geometrical: (Scatter plot metrics, landscapes, projection pursuit techniques, projection views hyper slice, parallel coordinates. Hierarchical (Mineset’s Tree Visualizer for Network Analysis, Intrusion Detection, and fraud detection). The other algorithm of visualization are Icon based, Pixel based and Hybrid Link Analysis: affinity Analysis, Association rules, sequential analysis.</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td>Decision Tree based Methods: The CART family (CART, IND CART, Splus CART, etc., The ML family: ID3, C4.5 (two stage system based on adoption of C4.5 rule generators for rule discovery, C5 and other derivatives, etc., The AID family: THAID, CHAID, XAID, TREEDISC, QUEST etc. (SPSS-Clementine uses CHAID, Extended CHAID, CART, QUEST for handling missing values, Automatic Discretization, Cost Complexity pruning etc.) Rule Discovery – PDAT by Siemens ZFE for intrusion detection; TASA for telecommunication alarm data; Neural Network (Self Organizing Map (SOM)) - User profiling for fraud detection in mobile communications networks and mobile network monitoring: Feed forward neural Network, Density estimation with Gaussian mixture model, and Bayesian network for fraud detection; Churn Management using SPSS’s Clementine:</td>
</tr>
</tbody>
</table>
**Genetic Algorithms:** for planning cable telecommunication network based on exploitation of the Kuhn-Trucker Optimality Condition; EGOIST for optimizing telecom distribution network; for optimizing multi-service convergence in a Metro WDM network; terminal assignment problems of a hybrid Hopfield network, Hybrid Genetic Algorithm Approach for Backbone Design of Communication Networks, etc.

**Bayesian Network:** for managing loss cells in a network to improve QoS; for Modelling the Physical Condition of Copper Access Networks;

**K-nearest Neighbor Classifier, Case Based Reasoning** (Conducting automated tests for Synchronous Digital Hierarchy (SDH) network components)

**Rough Set and Fuzzy Set Approaches** (Fuzzy Grid Based Sequential Mining Patterns Mining Algorithm (FGSPMA))

**Memory Based Reasoning:** Fraud detection and prediction about new customer

**Meta Algorithms (Boosting & Bagging):** Adaboost, Rareboost for mining with rare cases such as telecommunication network failure. The other algorithms are such as **Supper Vector Machines etc.**

**Clustering**

**Hierarchical Methods** (BIRCH, CURE, Chameleon, SAS CLUSTER, MIKCA, NORMIX, Agglomerative and Divisive Hierarchical Clustering)

**Grid Based** (STRING, CLIQUE, Wave Cluster)

**Constraint Based Clustering:** Analysis of incoming calls, sub grouping them such as hold calls, by telephone lines and their intensity etc.

**Partitioning methods** (K-means, K-medoids, CLARANS)

**ML Algorithm:** Monitoring of telecommunication databases,

**Density-Based Methods** (DBSCAN, OPTICS, DENCLUE)

**Disjoint Clustering** (SAS FASTCLUS) for Asynchronous Transfer Mode (ATM) network CDR’s classification. The other methods are such as **Algorithm for H/MDD: Scalable Clustering algorithm and Methods based on Co-Ocurrence of Categorical Data**

**Statistical Analysis**

**Regression Analysis:** Customer satisfaction analysis (Regression analysis, Mean analysis, Gap Analysis, Satisfaction Analysis); Simulation- ARENA Software;

**Multivariate Analysis:** Factor analysis to delineate places with a high degree of centrality, and their zones of influence, using telephone flows between different locations; principle component analysis to delineate urban linkages based on long distance communication data

**Forecasting Techniques** (ARMA- ATM Traffic Modeling; Use of TSE-AX expert system and Time series expert 2.3 for forecasting mobile telecommunication data correction for outliers and replacement of missing observations;

**Sequential pattern Analysis:** Analysis of CDR’s. Adds time comparison between transactions, IBM’s Sequential Mining Algorithm.

**Canonical Discriminant** (SAS CANDISC), ATM CDR’s Analysis

**Hidden Markov Models** – ATM Traffic Modeling

**Poisson Model,** Pareto- Modulated Process, Self Similar Processes (Fractional Brownian Motion (FBM)- ATM Traffic Modeling
| Pre-processing | **Data Cleaning, Feature Engineering** (Feature Extraction, Feature Selection [Unsupervised – Domain Knowledge and Random Sampling, Supervised-Filter Approaches and Wrapper Approaches], Feature Discretization), **Sampling**. Application: Pre-processing for Web personalization analysis. The other methods/algorithms which are part of all the software’s are Semantic **Summarization**, characterization, generalization and **deviation analysis**. |