FbMapping: AN AUTOMATED SYSTEM FOR MONITORING FACEBOOK DATA

S. Kamal[*], *N. Dey*[†], *A.S. Ashour*[‡], *S. Ripon*[*], *V.E. Balas*[§], *M.S. Kaysar*[¶]

Abstract: In recent modernized era, the number of the Facebook users is increasing dramatically. Moreover, the daily life information on social networking sites is changing energetically over web. Teenagers and university students are the major users for the different social networks all over the world. In order to maintain rapid user satisfactions, information flow and clustering are essential. However, these tasks are very challenging due to the excessive datasets. In this context, cleaning the original data is significant. Thus, in the current work the Fishers Discrimination Criterion (FDC) is applied to clean the raw datasets. The FDC separates the datasets for superior fit under least square sense. It arranges datasets by combining linearly with greater ratios of between – groups and within the groups. In the proposed approach, the separated data are handled by the Bigtable mapping that is constructed with Map specification, tabular representation and aggregation. The first phase organizes the cleaned datasets in row, column and timestamps. In the tabular representation, Sorted String Table (SSTable) ensures the exact mapping. Aggregation phase is employed to find out the similarity among the extracted datasets. Mapping, preprocessing and aggregation help to monitor information flow and communication over Facebook. For smooth and continuous monitoring, the Dynamic Source Monitoring (DSM) scheme is applied. Adequate experimental comparisons and synthesis are performed with mapping the Facebook datasets. The results prove the efficiency of the proposed machine learning approaches for the Facebook datasets monitoring.

Key words: *Fisher Discriminant Criterion, Sorted String Table, Bigtable, Dynamic Source Monitoring, aggregation*

Received: April 1, 2016
Revised and accepted: October 3, 2016

DOI: 10.14311/NNW.2017.27.002

[*]Sarwar Kamal – Corresponding author; Shamim Ripon; Department of Computer Science and Engineering East West University, Bangladesh, E-mail: sarwar.saubdcoxbazar@gmail.com, dshr@ewubd.edu

[†]Nilanjan Dey; Techno India Institute of Technology, Kolkata, India, E-mail: neelanjandey@gmail.com

[‡]Amira S. Ashour; Department of Electronics and Electrical Communications Engineering, Faculty of Engineering, Tanta University, Egypt, E-mail: amirasashour@yahoo.com

[§]Valentina E. Balas; Faculty of Engineering, AurelVlaicu University of Arad, Romania, E-mail: balas@drbalas.ro

[¶]Mohammad Shibli Kaysar; Chittagong University of Engineering and Technology, Bangladesh

1. Introduction

In the era of cyber networking, billions of datasets are interacting over webs, industries, institutes and companies. These interactions are increased rapidly with time as well as the datasets reflect the problems and situations of the individuals' current society as well as status. Real world datasets are collected from thousands of web browsers and end users through the social media, such as Facebook, LinkedIn and Twitter. Facebook is one of the dominant media that allow users to have efficient and optimal communication. Both Facebook and LinkedIn are popular social media that helps users to assess known people from their list of database [33]. Facebook facilitates this feature by showing message as add friends. Netflix also arrange large datasets for users by maintaining recommender system [2]. Moreover, Twitter organizes the datasets in another ways as real world spelling mistakes and query suggestions [26]. Facebook and Twitter are the key ways to understand the real world, science, markets and politics. Consequently, information mining from these social media methods is very challenging as well as costly, where outcomes become significant after adequate analysis. Thus, network interactions' monitoring is imperative [30] along with video qualities [39], where users' communication is maintained through text-/video- chat. This monitoring is required to find out meaningful information and users' interest.

Recently, machine learning procedures are important for data processing to guarantee high information quality on the web. Data and information on the Facebook have been changing rapidly over time due to the new information generation and methods that control data transmission via networks from one web to web. Arranging processes for managing such mechanical data at large scale creates many challenges. Thus, researchers and scientists are interested with the way of changing and altering the contents and information during communication over dedicated networks.

Now-a-days, a wide variety of approaches have been imposed to track and monitor the Facebook and social media interactions towards smooth and easy communication. These techniques and hybrid architectures helps to maintain such monitoring and tracking interrelated datasets. Simultaneously, machine learning and efficient data retrieving techniques helps to fulfill the demands of the accurate management of Web information. Retrieving data and information from Big data is a demanding task. The existing machine learning techniques' complexities are growing exponentially, which provide very less support for big data handling process. Some popular framework such as Hadoop and standard data warehouse are not worked for large big data processing. However, there are limited works that deals with individual interest [10,11,24,29]. In this regards, technology independent machine learning system are vital for data analysis.

Moreover, it is significant to have balanced interaction among all the nodes in the connected networks for better Web communications. However, this interaction balance is a challenging process as the interactions among the nodes are very much complex due to the continuous increase of the nodes. Fig. 1 illustrated that the relationships are always many to many in the regular networks. In such cases, monitoring of the networks is a critical task.

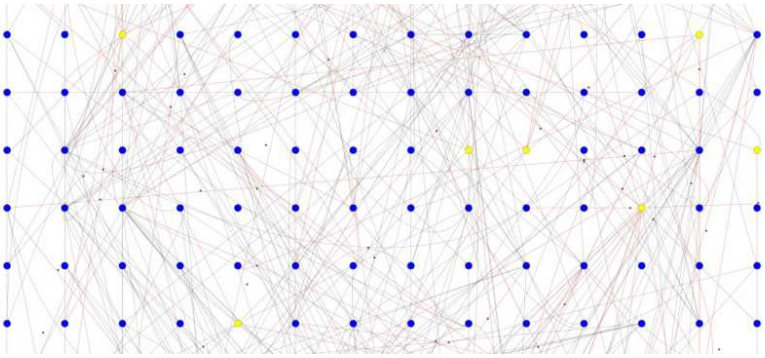


Fig. 1 *Nodes in Facebook networks.*

Dynamic monitoring is essential for repeatedly check the interactions among nodes. These nodes perform the automated communications as well as priority based communications. However, priority based communications are not performed always. Thus, in order to achieve faster communications as well as secure message passing, dynamic source monitoring (DSM) and dynamic network communication maintenance (DNCM) are considered. The DSM and DNCM help to track a very large set of the network's nodes that exist in the Web pages. The multi-face nodes are easily monitored by the DSM. This enables self-management/auto-configure the network without extra governmental support, which allow loop free assessment. In order to achieve reliable Facebook tracking, detailed simulation is required. Currently, it is compulsory to have a system that manages and processes millions of Web datasets/documents per day with continuous increased complexity. General approaches such as route finding and route assignment suffer from time consuming as well as space requirements. Thus, it is critical to track a very large datasets with efficient and cost-friendly manner.

Till April 3, 2016 there are about 1.59 billion Facebook users [1]. This number of users is huge and it is complicated to track the communication factors as well as the users' interest. Consequently, the current work proposed a dynamic tracking system with information monitoring association over social media's especially for Facebook. Dynamic associations as well as relationships might help to have sufficient satisfactions over large datasets and interactions. This work is interested with grouping the datasets that collected from active Facebook users to monitor the interactions among thousands of users. The grouping/monitoring processes are performed for big datasets with mapping based machine learning techniques such as Fisher Discriminant Analysis, Canonical correlations, Maximum likelihood and Dynamic Source Monitoring. Thus, the main contribution of the current work is to categorize the interest of Facebook users, gender and age interest as well as to monitor the interactions among the Facebook users irrespective of the gender.

The structure of the remaining sections is as follows. Section 2 includes the related work, followed by the methodology in Section 3. Afterward, the Bigtable data storage system is depicted in Section 4. The experimental results along with the discussion are presented in Section 5. Finally, the conclusion is presented in Section 6.

2. Literature review

Researcher focused on establishing social relationships by using online social networks rather than offline relations to provide strong and effective relations [12]. Teenager people who are suffering inhumanity or unsocial emotion, social network provide the opportunity to develop such friendship and social activities [34]. It was established that social network is important to express identity, which varies from male to female. Social network is essentially flexible to promote individual ethic and customs [6, 7, 19, 32, 35]. Furthermore, individuals can express their views and opinions with their friends or other supports in critical peer based sociality [37] . Such processes of socialization are imperative for psychosocial development. Smith et al. [31] reported that 37% of 18–29 year old youth use blogs and Social network for civic and political engagement. Political candidates are increasingly utilizing social network/media for election publicity and issue-orientated groups [17, 22, 27].

Various parameters and factors are considered for Social Network (SN) research. Recently, online social sites become popular in the web, where people are searching for friendships and social activates [12]. Hence, the researchers provide their interest toward different social media users on different age groups. They try to find out personal views and reaction that vary with different age groups and gender. Gross and Acquisti [17] presented a survey on 4000 Carnegie Mellon University Facebook profiles from their personal views, opinions and activates. The authors concluded that teenagers have a potential threat for their profile and personal life. Teenager spent most of the time in social media such as Facebook, twitter and WhatsApp for chatting or other unproductive activity. As a result teenagers suffer from less concentration on their academic life. However, social media has positive effect on education through e-learning and integrated framework for education purpose [28]. Some users access the e-learning facilities by using social network to express their knowledge that leads to formal education by using the social networks [3]. Several pilot studies emphasized on the potential of social software, services on school and higher education frame work [15]. Social network provides an opportunity for formal and higher education across geographical context.

Social networking practices are a routine part of teenager and adult people's daily lives. Researchers are interested with the overcoming of the online threat due to the social networks [25]. Thus, several studies are conducted to monitor the social networking using various machine learning approaches. Kamal and Arefin [21] proposed the Apriori algorithm based analysis to assess amount time spending by teenager students in school and college level. An association rules categorized the maximum numbers of hours spent by English medium students. This work considered only one factor, namely the study number of hours. Venkatesan et al. [36] presented a protocol management system based on grouping data privacy on specific database. However this work does not categorized the collected datasets and there is no indications how they have achieved the data in database.

Consequently, the current work (FbMapping) is proposed an automated approach that groups the datasets in certain groups. Moreover, the monitoring system depicted in same work under relations with connected users. Furthermore, several factors are considered to group the interest and monitor the interactions among users. Recently a community detections algorithm has proposed that iter-

atively update node information in a local area. This approach works only for a small area or local area. However, our work covers both local and global area [9].

3. Methodology

The integration of this work was done with the coordination's of multiple machine learning approaches to design an automated system for monitoring the Facebook Data. Fig. 2 demonstrated the proposed system procedure, where initially data was collected from Web and Facebook pages' users. These collected data contains noise and irrelevancy.

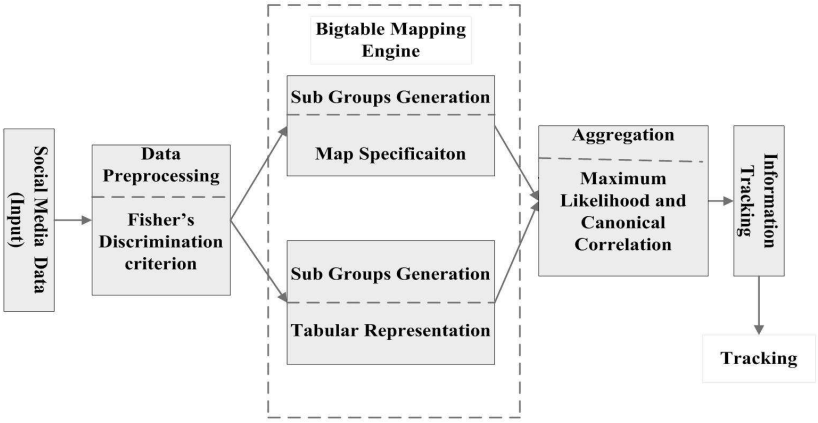


Fig. 2 *Proposed system architecture.*

Fig. 2 depicted that data preprocessing phase was involved to remove the noise and data mismatches. Fishers Discriminant analysis is efficient due to its dynamic managements of large datasets calculations. It helps to measures large datasets iterative over limited contains. Thus, it was used in the current work for noise removal as well as for data classification. Moreover, mapping control and large datasets grouping over individual classes are one of the important and key parts. Map specifications and tabular representations are the pivotal analysis for mapping. In specifications, mapping generates the whole datasets in some data management fields as Row, Columns and table content. These three are the primary instances of mapping. Sorted String Table (SSTable) is another significant element of mapping that ensures dynamic data allocations over limited datasets. It defines the tabular representations repeatedly until it satisfies all the collected data. Aggregation is the intermediate part that associates all the processing and filter the datasets for last level. The aggregation is done with Maximum likelihood and Canonical correlations analysis. Both these two methods allows automated filtering as well as grouping accurate choice of interest that are the pivotal objective of this research. Finally, dynamic tracking system permits automated network tracking with certain approach. Different subspaces of learning methods [18, 23] are used for feature selection and classification.

3.1 Fisher's discriminate criterion

Classification is a machine learning approach that aim is to assign a predefined class for every instance. It classifies the existing data set into classes for new instances. Due to the required analysis for large data sets, it is significant to extract features from the entire data set before classification. Afterward, feature selection is used to select the most informative features for the original data set. Several techniques can be used for features selection from large data sets [9,18]. One of the powerful adaptive learning approaches for data clustering/feature selection techniques is the Linear Discriminate Analysis (LDA). LDA achieves good performance for classification by using covariance matrix among the groups [13]. The LDA has used for preprocessing phase that classify the data sets into different class groups. In preprocessing phase, data are selected according to specified data of interest. Fisher Discriminate Analysis (FDA) [14, 16, 38] is a popular reduction approach which maximizes between class scatter and minimizes within class scatter. Data preprocessing indicates noise removal and irrelevant data removal. There are sets of machine learning approaches that help in cleaning large datasets. All the methods are appropriate for specific datasets. In this research work, Facebook datasets have been considered that are wide and discrete. In this regard, FDA covers whole datasets in both local and global level. Other methods can handle the problems but unable to consider in separate manner. LDA is classified the observation into two different approaches: Two classes approach and Multi class approach.

3.1.1 Two classes approach

The LDA was introduced by Fisher [14] for two classes to transform multivariate observations $\mathbf{x}$ to univariate observations y. The y is classified into different groups that derived from the two possible classes. Suppose there is a set of m samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ that belong to two different classes $\mathcal{C}_1$ and $\mathcal{C}_2$. The scatter matrix for the class i is given by

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{C}_i} (\mathbf{x} - \bar{\mathbf{x}}_i)(\mathbf{x} - \bar{\mathbf{x}}_i)^{\mathrm{T}},$$

where

$$\bar{\mathbf{x}}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x}$$

and m_i is the number of samples of $\mathcal{C}_i$. The total within classes scatter matrix for two classes is then given by

$$\mathbf{S}_{\mathrm{w}} = \mathbf{S}_1 + \mathbf{S}_2$$

and the inter-class scatter matrix is given by

$$\mathbf{S}'_{\mathrm{b}} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}}.$$

The purpose of the LDA of Fisher discriminator is to find the projected vector $\mathbf{w}$ that maximizes the Fisher separation criteria, which is expressed by

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}'_{\mathrm{b}} \mathbf{w}}{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{w}} \mathbf{w}}$$

To determine the value of $\mathbf{w}$, the eigenvalues problem of $\mathbf{S}_\mathrm{b}\mathbf{w} = \lambda\mathbf{S}_\mathrm{w}\mathbf{w}$ with its eigenvalue was generalized. Assume n number of the original feature set to be $\{f_1, f_2, \ldots, f_n\}$. Then, feature selection is required to select certain number of features, d, $F_d = \{f_{d1}, f_{d2}, \ldots f_{dn}\}$ from the original features that have the largest Fisher's selection value. Here, $d(i)$ is the selected feature index in the features subset. The selected feature set F_d was denoted for the class scatter and within class scatter as $\mathbf{S}_\mathrm{b}(f_d)$ and $\mathbf{S}_\mathrm{w}(f_d)$ respectively. The Fisher selection criterion $J(F_d)$ is based on separation criterion that is formulated by

$$F_d = \arg\max J(F_d),$$

where $J(F_d) = J(F_1, F_2, \ldots, F_d)$ is defined as

$$J(F_d) = \frac{\mathbf{w}^\mathrm{T}\mathbf{S}_\mathrm{b}(F_d)\mathbf{w}}{\mathbf{w}^\mathrm{T}\mathbf{S}_\mathrm{w}(F_d)\mathbf{w}}.$$

The equations compute all the weights, their standard deviations, weight deviations and associated weight variations with accurate multiplication. A priori algorithm checks the specific facts based on rule base associations. On the other hand, FDA covers whole data sets according to the statistical probabilities and interactions. From the experiment result it has been noticed that, FDA has wide scope to handle both local and global datasets.

3.1.2 Multi-class approach

The multi-class approach is used when the exit observation contains more than two classes. Thus, Fisher's Linear Discriminate will be multiple discriminate analyses (MDA) [11]. As in two classes approach, the multi-class approach will classify the observation into multiple classes rather than two classes. However, the maximum value is computed for several competing classes. The within classes scatter matrix for n classes is calculated by

$$\mathbf{S}_\mathrm{w} = \mathbf{S}_1 + \cdots + \mathbf{S}_n = \sum_{i=1}^{n} \sum_{\mathbf{x}\in\mathcal{C}_i} (\mathbf{x} - \bar{\mathbf{x}}_i)(\mathbf{x} - \bar{\mathbf{x}}_i)^\mathrm{T}$$

and the between classes scatter matrix is computed by

$$\mathbf{S}_\mathrm{b} = \sum_{i=1}^{n} (\bar{\mathbf{x}} - \bar{\mathbf{x}}_i)(\bar{\mathbf{x}} - \bar{\mathbf{x}}_i)^\mathrm{T},$$

where $\bar{\mathbf{x}}$ is the total mean vector,

$$\bar{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^{n} m_i \bar{\mathbf{x}}_i.$$

After obtaining $\mathbf{S}_\mathrm{w}$ and $\mathbf{S}_\mathrm{b}$, the linear transformation matrix $\mathbf{W}$ can be calculated by the generalized the eigenvalue problem

$$\mathbf{S}_\mathrm{b}\mathbf{W} = \lambda\mathbf{S}_\mathrm{w}\mathbf{W}.$$

By solving the eigenvalue problem, the data were classified into multiple classes, where the MDA provides an optimal classification. Once the transformation matrix $\mathbf{W}$ is obtained, the classification is performed based on distance matrix that is calculated by the Euclidian distance using the following expression

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

In addition, the cosine distance is used

$$d(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}.$$

The cosine distance is used to measure the similar interests among Facebook users. Any new instance $\mathbf{z}$ is classified into class $\mathcal{C}_k$ that is generated by

$$\mathcal{C}_k = \arg\min_{\mathcal{C}_k} d(\mathbf{z}\mathbf{W}, \bar{\mathbf{x}}_k \mathbf{W}),$$

where $\bar{\mathbf{x}}_k$ is the central point in class $\mathcal{C}_k$. Generally, the data is classified based on the centre classified point.

This procedure was used for Facebook data classification and monitoring as in the following section.

3.2 Facebook dataset processing

Based on the first phase in the proposed system, the FDA was applied for the Facebook user data mining in the pre-processing to achieve good accuracy. In the current work, initial processing is performed with two class approach and later multi-class approach is applied for global data sets. All data sets were divided into two parts based on the gender and age (teenager/adult). Different factors and parameters were inferred from a set of social media users. The important factors are time, people interests such as food habit and the travelling, the number of male/female users, and the number of teenager/adult people users. The pre-processing reduces the data impurities or data redundancy. All the datasets have been cleaned by using FDA with the intra and inter classes. These two types of classes make a boundary of the used mining data sets in the current work. These filtered data sets were used in Bigtable [8] approach in the next phase. To reduce the mining complexity, big table spilt was performed to split the table into multiple tables. For example, the primary categories were the age and gender. Every category had two sub-categories such as gender (male/female) and age (teenager/adult). Assume the number of users based on the two categories in different data records which checked automatically was as follows:

Gender ($\mathcal{C}_1$): $\mathbf{x}$ = (male x_1, female x_2) = $\{(4,2),(2,4),(2,3),(3,6),(4,4)\}$,

Age ($\mathcal{C}_2$): $\mathbf{x} = (\text{teenager } x_1, \text{ adult } x_2) = \{(9,10),(6,8),(9,5),(8,7),(10,8)\}$.

It is possible to choose multiple classes during the Facebook users monitoring. The total of n classes could be available, where the value of n is varies from ten to twelve.

The classes' mean values are

$$\bar{\mathbf{x}}_1 = \frac{1}{m}\sum_{i=1}^{n} x_i = \frac{1}{5}\left[(4,2)+(2,4)+(2,3)+(3,6)+(4,4)\right] = (3,3.8), \tag{1}$$

$$\bar{\mathbf{x}}_2 = \frac{1}{m}\sum_{i=1}^{n} x_i = \frac{1}{5}\left[(9,10)+(6,8)+(9,5)+(8,7)+(10,8)\right] = (8.4,7.6). \tag{2}$$

By using Eqs. (1) and (2), the scatter matrices of the two classes are as follows:

$$\begin{aligned}
\mathbf{S}_1 &= \sum_{\mathbf{x}\in\mathcal{C}_1} (\mathbf{x}-\bar{\mathbf{x}}_1)(\mathbf{x}-\bar{\mathbf{x}}_1)^{\mathrm{T}} = \\
&= [(4,2)-(3,3.8)]^2 + [(2,4)-(3,3.8)]^2 + [(2,3)-(3,3.8)]^2 + \\
&\quad + [(3,6)-(3,3.8)]^2 + [(4,4)-(3,3.8)]^2 = \begin{pmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{pmatrix}, \\
\mathbf{S}_1 &= \sum_{\mathbf{x}\in\mathcal{C}_2} (\mathbf{x}-\bar{\mathbf{x}}_2)(\mathbf{x}-\bar{\mathbf{x}}_2)^{\mathrm{T}} = \\
&= [(9,10)-(8.4,7.6)]^2 + [(6,8)-(8.4,7.6)]^2 + [(9,5)-(8.4,7.6)]^2 + \\
&\quad + [(8,7)-(8.4,7.6)]^2 + [(10,8)-(8.4,7.6)]^2 = \begin{pmatrix} 2.3 & -0.50 \\ -0.50 & 3.3 \end{pmatrix}.
\end{aligned}$$

Thus, within classes scatter matrix between the two categories is

$$\begin{aligned}
\mathbf{S}_{\mathrm{w}} &= \mathbf{S}_1 + \mathbf{S}_2 = \\
&= \begin{pmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{pmatrix} + \begin{pmatrix} 2.3 & -0.50 \\ -0.50 & 3.3 \end{pmatrix} = \begin{pmatrix} 3.3 & -0.50 \\ -0.50 & 5.5 \end{pmatrix}
\end{aligned}$$

and similarly the between classes scatter matrix

$$\begin{aligned}
\mathbf{S}_{\mathrm{b}} &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} \\
&= [(3,3.8)-(8.4,7.6)]\,[(3,3.8)-(8.4,7.6)]^{\mathrm{T}} = \begin{pmatrix} 19.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix}.
\end{aligned}$$

These matrix values indicate the maximum and minimum value for data filtering, which consider the data range for every category. The valid data range within the class and between the classes are generated and allowed in the current study.

4. Bigtable data storage system

The Bigtable is a new distributed concept for big data handling as it is considered a cloud computing that operates on thousand commodity servers. It is a

distributed file system that manages large size data, such as petaByte structural data. The Bigtable has several advantages including: (i) it uses more than sixty Google product a nd project including Google finance, Google analytics, Google earth, Orkut, Personalized Search and Writely [8], (ii) Cloud Bigtable allows large scale of single key data without latency, (iii) it is an ideal data source for the map reducing operations, (iv) it achieves several advantages such as simple administration, incredible scalability, wide applicability, cluster resizing, high ability and high performances, (v) it resembles the database and applies many database implementations strategy, (vi) it provides different interface strategy in parallel and main memory database [4,5], (vii) it supports a simple data model for client. It provides dynamic control over the data format and layout that allow clients optimal operation in underlying storage, and viii) it can handle structure and unstructured data sets. Typically, the Bigtable's parameters provide the client dynamic control that serve the data from different dedicated servers.

The key element of the Bigtable is the Sorted String Table (SSTable) as illustrated in Fig. 3. Bigtable is distributed, sparse, and persistent distributer maps. The Bigtable data is indexed by a rows, columns and contents. Bigtable settled this approach after examining a variety of potential uses. Google use their web table for Bigtable processing. In the web table, Google placed the URL (Uniform Resource Locator) as row keys, various web page aspects placed in columns and store the web content in content field. Web data or content are fetched from content field.

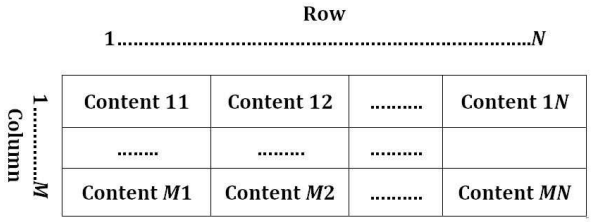


Fig. 3 *A slice of pattern for big table data processing.*

The row, column and content are the basic element that generates Bigtable operation. They are selected based on user demand. Parameters of Bigtable generate the ultimate searching string pattern are represented by

$$(\text{row:string}, \text{column:string}, \text{content:string}) \rightarrow \text{string}.$$

Bigtable and MapReduce handle large datasets in parallel mode. Bigtable organize whole datasets into parallel tables with multiple rows and columns. On the other hand, MapReduce deploy set of methods in parallel levels. In this regards, Bigtable generates better result for homogenous datasets. All the datasets considered here are homogenous data. As a result, Bigtable outperform MapReduce for this Facebook data monitoring.

4.1 Rows

The row keys in a table consists arbitrary data with size from 1bit to 64KB. Bigtable designs decision approach which is easier to the client. Bigtable allows con-current

data to update and analyze the system behavior. The data rows maintain the lexical order, where the rows are portioned dynamically. The rows' range is known as a tablet that reduces the overload and units of distribution. For that small range of rows communicate efficiently and provides less overload. Users can easily access the data from small range rows. Moreover, they can reduce the data set based on rows that are selected by the dynamic approach. The selected rows consist of different column families and contents. These contents are selected based on rows and columns using the dynamic approach.

4.2 Column families

The column keys are grouped in similar sets called column families, which indicate the user basic controls. Same types of data are stored and compressed in the same column families. Before data insertion, a column is created and then other column is created for another type of data sets. During data retrieve operation, a small number of distinct columns are created with rarely changed families. Each column family consists of anchor and qualifier, where the Anchor represents the cell content and the qualifier may be arbitrary string. Access control and memory as well as the disk management are performed in the column family level. The small part of a table is named as web table that provides multiple operational facilities for column families. Some new data is added, create new derived column to view data for column families operation.

4.3 Table content

Each cell in the Bigtable contains different types of data sets that known as table content. This content can have different sizes. They can be assigned by the Bigtable. Typically, the table contents have unique data to avoid data collisions. Stored content data facilitates the user searching process. For optimal user data search in the Bigtable, an ordering procedure is followed, where the top priority data exists in the first content followed by the second priority data in the second cell and so on. The Bigtable considers and collects automatically the less used data as garbage. Thus, the user is always provided by update facilities, where the Bigtable is always updating the data sets. Ultimately total content generates the Bigtable interface.

4.4 Building blocks

In order to support parallel processing, big tables build several table block infrastructure. A Bigtable cluster share the table blocks for various types of applications. Big table blocks are concurrently shared for different application. Bigtable performs a cluster management system that allows job scheduling, resources failure, managing resources and memory/machine status. Bigtable is used a persistent and ordered distributed file system, which contains a sequence of blocks. Every block has a size of 64 KB though it's configurable. In order to find the blocks' locations, a blocks index is used. When the distributed file is opened, the block index is loaded in the memory. A lookup operation is used for accessing the table block.

First binary search is used for searching block index, and then find out the block location in the memory.

Generally, the distributed file system has several tasks to i) ensure that master copy of the file is always active, ii) store the bootstrap location, iii) discover the tablet locations to store the Bigtable schema, and iv) store access control lists. However, the main task of the distribution file is to balance the tablet files, to collect the file's garbage and to handle the schemas changes such as rows or column families' creations. A Bigtable clutters store a number of tables, which are small pieces of the file that handles large data sets. Each table contains a set of tables and this tables consists an associated data in row and column range. In the initial phases, Bigtable consists of a single table, and then as the data size increases, multiple tables are added. Bigtable spilt the master tables into multiple table blocks; each table blocks have a size of 100–200 MB. Fig. 4 demonstrated the hierarchy of the table blocks for Bigtable design.

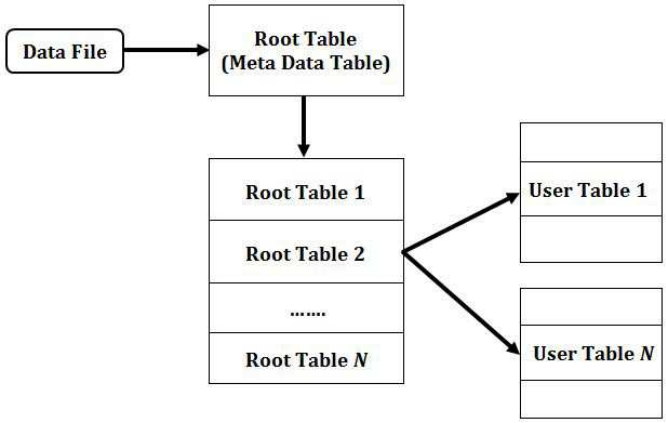


Fig. 4 *Table blocks hierarchy for Bigtable design.*

4.5 The proposed Bigtable for Facebook datasets

In the current work, the Bigtable is deployed for Facebook data mining approach as well as for handling large volume of data sets. The proposed scheme generates a prediction approach for Facebook based on different categories, such as the time spent in social media (Facebook), age of the Facebook users, and the users interest (e.g. food habit, traveling, and gaming). The proposed approach was applied to data sets stored in a file, which divided into different rows, column families and contents as illustrated in Fig. 5a. Every row contains the record and the column families indicate the attributes. Every cell contains the numeric value that indicates the user number in a certain category. The used data sets were previously preprocessed by using fisher discriminate. Every rows, column families and content generate a Bigtable for the used excremental data set. In preprocessing phase, this table reduces the redundant data and the garbage data.

The Bigtable was divided into several known tables as depicted in Fig. 5b. Basically, tablets are table blocks that reduce mining complexity due to their size.

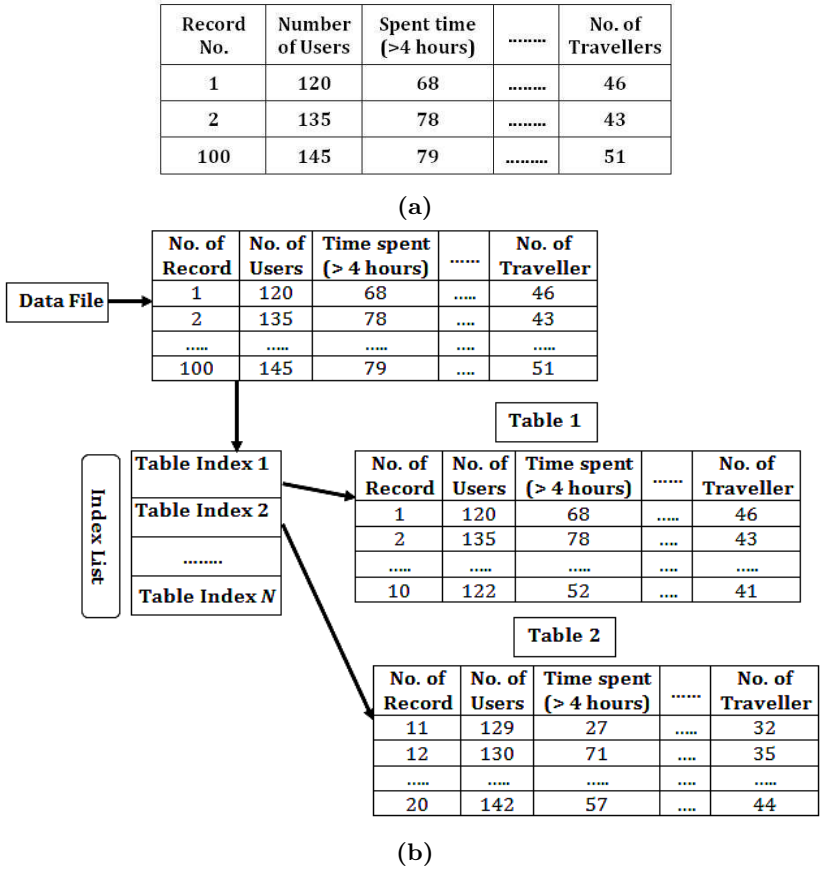


Fig. 5 *(a) Bigtable schema for Facebook mining table (b) Table blocks for mining process.*

Every table blocks contain similar column families as their master copy. In the proposed system, a mining operation was applied on table blocks in concurrently. This approach enhances the job scheduling, manages large data, and enhances the accuracy of mining approach. The used table blocks are different in size and maintain a block lists. The block lists contain the physical address of the blocks. From the block lists, the desired table block address was selected and accessed the table for mining approach. Moreover, a table schema was maintained for the table. The used table schema was FbTable (UserNo, Spent_Time, No_of_male, No_of_female). Furthermore, every table blocks maintain the similar table schema. In the current work, the table block had a size of 100–200 Mb. This small size table reduces the data volume and enhances the data classification accuracy.

Multiple rows allow set the data in a single plate form. Similar rows are easily managed by removing the common entry from large data table. This is one kind of reduction that occurs in MapReduce.

4.6 Canonical Correlations Analysis

The interactions among Web and nodes can be adjusted using the statistical analysis. One of the very popular estimation method is the Canonical Correlation Analysis (CCA) [20]. This is a multivariate demonstration that leads to data sets with common items as well as features. Facebook data sets grouping as well as interests on common fields are important to measures the common group. For large networks, it will be easy to handle the complete network with statistical measurements. For example, suppose there are nodes as p and q in two different dimensions as u and v respectively. All nodes in a certain Web networks have been divided into two classes as $\mathcal{A}$ and $\mathcal{B}$, where $\mathcal{A} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and $\mathcal{B} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$. The complete relations and interactions were measured in the proposed work in the form of correlations formulas as follows:

$$p = \frac{p' \sum_{\mathbf{xy}} q}{\sqrt{p' \sum_{\mathbf{xx}} p \sqrt{q' \sum_{\mathbf{yy}} q}}},$$

where $\sum_{\mathbf{xx}} = E[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^{\mathrm{T}}]$ and $\sum_{\mathbf{yy}} = E[(Y - \mu_y)(Y - \mu_y)']$ are the covariances of $\mathcal{A}$ and $\mathcal{B}$ respectively. The cross-covariance of $\mathcal{A}$ and $\mathcal{B}$ is expressed by $\sum_{\mathbf{xy}} = E[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{y}})^{\mathrm{T}}]$. In a certain Web networks, the continuous nodes should have some features with appropriate facilities with certain lengths. Two social networks nodes $\mathcal{A}$ and $\mathcal{B}$ were used to interact with certain j. Moreover, the neighbor nodes were followed the same procedure, where

$$\hat{p}(j) = \frac{\hat{x}(j)^T \sum_{xy} \hat{y}(j)}{\sqrt{x(j) \sum_{xx} x(j)^T \sqrt{y(j)^T \sum_{yy} \hat{y}(j)}}}.$$

Canonical Correlation Analysis (CCA) associates common features of Facebook users in a certain families. It helps to merge all the features with less efforts and time.

4.7 Maximum likelihood estimating

Maximum likelihood method is a well known statistical estimation method, where the Maximum likelihood (ML) is an underlining approach that operates on fixed data size. The ML generates a set of probabilities from similar datasets and then selects the dataset with the highest probability. The similarity parameter values are based on the variance and mean for some sample data. The ML used normal distribution to find out the maxim parameters. The maximized value is the agreement of the selected data item with observed data for discrete random variable. The maximized probability was calculated from the distributed random variables, where the ML estimation provides a unified and well defined estimation function for the normal distribution.

Since, the ML principle is a straightforward approach, thus it was used with the Facebook sample data. The 450 Facebook users were denoted by $X = \{x_1, x_2, \ldots x_n\}$ of certain random variables (age group, gender, public view, etc.) The probability of every variable family is P_θ (collection of variables). In addition, $f(\mathbf{x}|\theta)$,

$\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are the density functions of the random variable and θ is the true value of the parameters. The estimation value θ' for every class variable in the proposed work Facebook monitoring approach was calculated. The likelihood function for the used data set is

$$P(\theta|x) = f(x|\theta), \theta \in \phi,$$

where θ is the threshold value for density function. The ML estimator is denoted as

$$\theta'(x) = \arg\max P(\theta|x).$$

In the proposed work, the ML was inclined for large data set that has different desirable variables, such as the age group, gender, travelling interest and food habit. For high volume data analysis, the ML generates multiple local maximum values, where it is challenging to find global maximum. The $\theta'(x)$ is estimated for the variance and standard deviation to remove the unbiased estimator.

A joint probability density function for all the 450 Facebook users' observations should be specified before using the ML approach. For an independent and identically distributed sample, this joint density function is given by

$$P(\theta|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = f(\mathbf{x}_1|\theta) \times f(\mathbf{x}_2|\theta) \times \ldots f(\mathbf{x}_n|\theta). \tag{3}$$

Afterward, the ML estimator for the age group, gender or personal view of multiple observations can be calculated by finding the similarity among the discrete data sets as follows:

$$P(\theta \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

This likelihood function returns maximum value for multiple variables for discrete data sets. The equation (3) estimates the probabilities of all datasets exist in the training data table. This probability helps to measure the similarities among all the common features and interest exist on Facebook communications and sharing.

k-NN

a. Basic data classifications are easy and simple. Only distances among points are needed to be considered. Euclidean distance is one of the common ways to solve the similarities among set of points.

b. Datasets that are used here are non parametric. These datasets helps to handle homogenous features and interactions.

c. Since distances are easily measureable, there are very less error rate in distances measuring.

d. All the distances among features and data points are slow in counting. There are discrete options to counts all the points exist. This approach makes the system slow and idle.

e. For counting each points and their associated distances, there consume excessive time.

f. Sometimes unable to find an optimal result from whole data processing area and datasets.

g. Big datasets processing are not suitable by k-NN for a whole. It can take support with MLE.

Maximum Likelihood

a. Maximum Likelihood Estimations (MLE) counts the probabilities for each and every datasets counting for getting desired values. In this work MLE combines whole homogenous data points and features in a specific tables for faster and desired features counting.

b. MLE can cover both parametric and non parametric datasets. There are both options to assess the whole datasets used in this approach.

c. Error rate are also minor and its outcomes are better than that of k-NN.

d. Log normal counting helps to generate faster results and decisions. There are better outcomes than that of k-NN.

e. MLE consumes less time in all respects of data processing.

f. Optimal values are available in Maximum Likelihood estimation. Log normal analysis generates set of optimal result

g. MLE is a good option for big datasets processing. It can cover up to five hundred mega byte DNA base pair with single iterative program.

4.8 Tracking nodes

Dynamic Source Monitoring (DSM) is interactive demonstrations that automatically create relationships among nodes exist in a specific Web. It is a systematic process with limited memory, time and system processing capacities. In a specific Web, there are huge Facebook users are used to communicate with each other irrespective of the geographical location. These interactions are multi hops exchange approaches to reach the end users communications. When any user or nodes in the Facebook network join or leave the Web, their communications path is easily determined and monitored by the DSM process. Basically, in a specific Web, Facebook users or nodes are existing in very congested way or large in size. Sequences of the next users or nodes are always important due its availability every time to deliver the information. Fig. 6 demonstrated the large Facebook nodes in a Web with the assumption that there are thousands of Facebook users in the Web.

Assume a group of users $P, Q, R, R, S, T, U, V, W, X, Y$ and Z are interacting over a network. A primary user P intended to communicate with U or other network. In this case, the DSM finds a path or route that help to monitor the interaction among nodes exist in a given network as shown in Fig. 7. This path

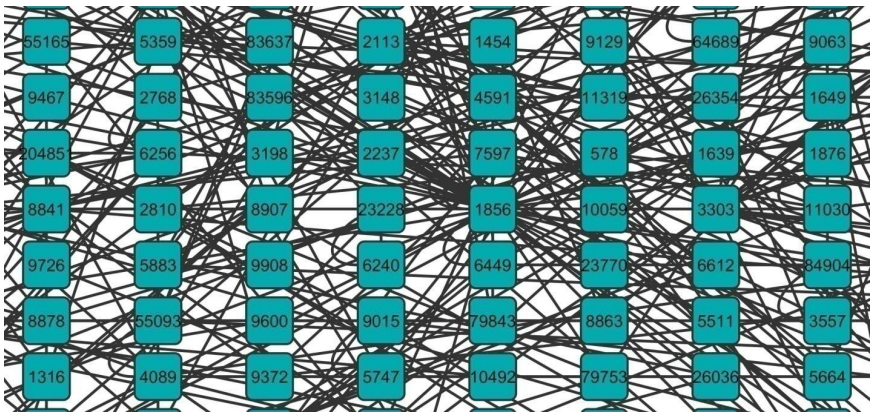
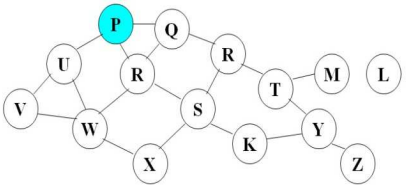
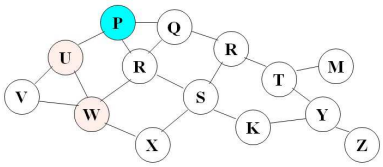


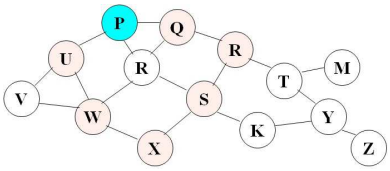
Fig. 6 *Large Facebook nodes in a web.*



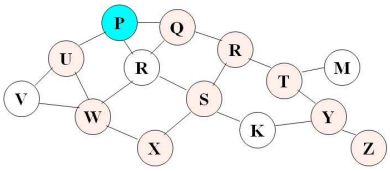
(a) *Web Facebook network with initial node P.*



(b) *Web Facebook network with initial node P and connections with U and W.*



(c) *Web Facebook network with initial node P and connections with U, W, X, S, Q and R.*



(d) *Web Facebook network with initial node P and connections with U, W, X, S, Q, R, T, Y and Z.*

Fig. 7 *Web Facebook networks.*

finding and interactions investigations are significant processes in the current globalization to maintain faster and secure communications. The DSM maintains a buffer to place the intermediate nodes that are part in the interactions or not in the given network.

Consequently, the proposed automated Facebook monitoring is designed to categorize the Facebook users' interest by gender and age. The monitoring process is used to check the interactions among the Facebook users irrespective of their gender. The Naïve Bayes automated classification process for social media data set is employed in the current work. Java tools were involved for the filtering process.

Interactions between two adjacent nodes depend mainly on the outcomes of CCA between the nodes. During the experiments, features of node P were similar

with features of Node U. since both Q and U are two adjacent nodes of P however the propagation starts from P to U due to common Facebook features and interest.

5. Result and discussion

Consider 450 users data for the present experiment in different categories. In order to test the proposed method, Java environment is used during the classification rate, execution time and system performance measurements. Tab. I included the users' record for different classification attributes, where the attributes of the data sets under concern is considered to be 1 or 0. If the users satisfy the attributes condition, then 1 is assigned otherwise 0 is used.

UID	Teenager	Adult	Time Spent≥ 4	Male	Female	Travelling Interest	Food Habit
U1	1	0	0	1	0	0	0
U2	1	1	1	1	1	1	1
U3	0	1	1	0	0	0	
...	...	...	...	...	...	...	...
U200	1	1	1	1	1	1	1
U201	0	1	1	0	0	0	
...	...	...	...	...	...	...	...
U449	1	1	1	1	1	1	1
U450	1	1	1	0	1	1	1
...	...	...	...	...	...	...	...

Tab. I *Master data file for classification with different categories.*

In Tab. I, UID indicates the user's ID (identification documents) as unique column. The Bigtable schema contains a primary field and assigns individual number for the 450 users. In the current study, two age groups (teenager and adult) are considered and are indicated in column two and three. In addition, the gender (male and female) and the time spent in the Facebook are considered. Column seven indicates the travelling interest of the Facebook users for choose travelling zone or place. Food habit indicates the preferred users' choices for the restaurants or hotels. Tab. II contains the count of the total users in individual attributes or columns. It indicates the total users count for individual attributes after data pre-processing.

Tab. II depicts the datasets size for mining approach, which is the Bigtable mining approach and Naïve Bayes classification. Since, Fisher discriminate approach requires less time than other approaches. Thus, prior to the classification process, pre-processing phase for the data sets by using fisher discriminate is performed.

5.1 The classification execution time

Fisher discriminate used classification techniques for filtering. Furthermore, with the increased social media size, Fisher discriminate speeds up the processes by en-

Attribute	User count
Teenager	223
Adult	227
Male	265
Female	185
Spent Time $\geq$ 4 hours	290
Travelling Interest	209
Food Interest	231

Tab. II *User count table for classification.*

abling intra- and inter- class classification. Two classes approach support faster calculation for larger data volumes. Consequently, for the proposed system evaluation, the effort rates along the original values are measured. Fisher classification reduces the data size that reduces the data processing time. Fig. 8 illustrates a comparison of the pre-processing execution time by using the fisher discriminate, Naïve filtering and without filtering. Without filtering is a simple string operation approach that required linear time for generates class group.

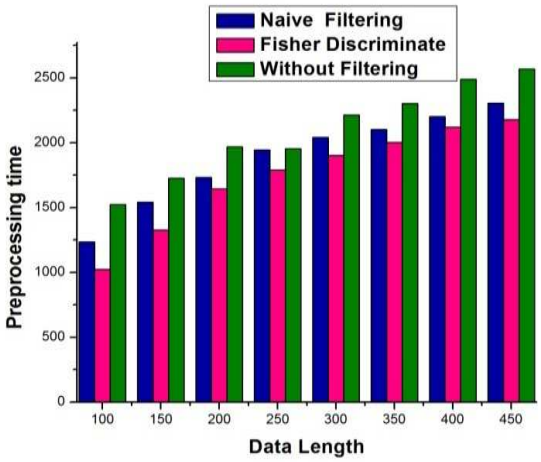


Fig. 8 *Comparisons among without filtering, Naïve filter and fisher discriminate based on preprocessing time.*

Fig. 8 depicts that the Naïve filtering achieves better relationship between time and user data length than without filtering results. However, the Fisher discriminate requires the least time compared to the other methods. The used fisher discriminate approach reduces sample data in intra-class and between classes. Diagonal linear discriminate is used for sub-groups generation for classification. Fisher discriminate reduces the sample data than the Naïve filter and without filtering. Since, data size become large, then the system complexity of Naïve filter becomes high for classification for social media.

The filtering table is divided by using Bigtable approach. Bigtable generates different table blocks with different data length. The small table blocks requires less classification time. The Bigtable needs less classification time compared to the naïve classification time. Since, the Bigtable blocks are different in size, thus it is assumed that the classification time is measured for similar data length. Tab. III contains the required classification time for the Bigtable and Naïve Bayes classifiers.

Data Length	Canonical Correlation Aggregation	Without Aggregation
100	234	210
150	334	289
200	432	367
250	567	489
300	612	567
350	699	612
400	734	698
450	845	778

Tab. III *Classification time for Bigtable and Naïve Bayes classifier.*

In Tab. III, the data lengths indicate the number of Facebook users in the Bigtable blocks. A minimum 100 Facebook users and maximum 450 users are considered in the current experimental data sets. It is established that, the increase of the data sets size leads to the increase of the classification time for both classifiers. However, the Bigtable requires less time compared to the Naïve classifier. For example as depicted in Tab. III, for 300 user classifications the Naïve Bayes requires 612 seconds for its classification, while the Bigtable requires 567 seconds. Thus, the Bigtable classifier is 7.35% faster than the Naïve classifier for 300 users' classification. Fig. 9 illustrates the required classification time for the two classifiers.

Fig. 9 along with Tab. III depict that the required classification time for the two classifiers (Bigtable and Naïve classifiers). As the number of users is increased, the classification time of both approaches is increased. The maximum required classification time is obtained in the case of 450 users, which is 778 seconds for the Bigtable and 845 seconds for the Naïve classifier. The least required classification time is measured for 100 users. It is clear that for every data set the Bigtable requires less time than the Naïve classifier.

5.2 The classification accuracy evaluation

Since, the used data sets are classified based on various attributes, such as the gender (male or female), age (teenager or adult), and the people views (food habit, travelling, sharing emotions). The data set classifications for these different attributes are illustrated are Fig. 10. Thus, the proposed classification is applied on different table block with different data size. The classification approach is

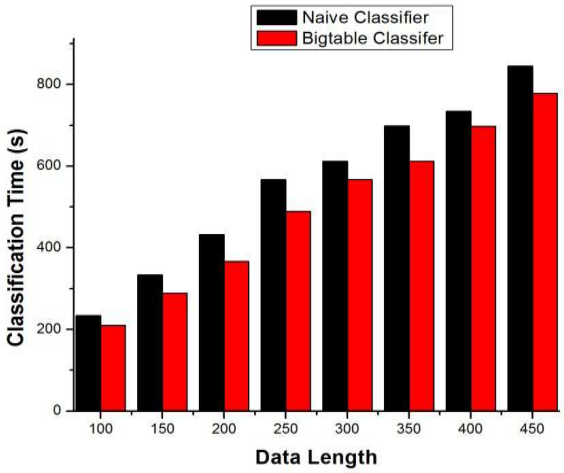


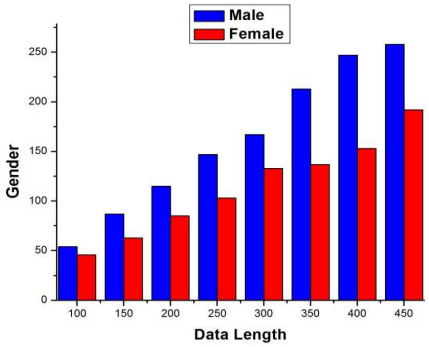
Fig. 9 *Classification time by using Bigtable approach and Naïve classifier.*

performed on tablet blocks in parallel approach, which enhances the classification accuracy.

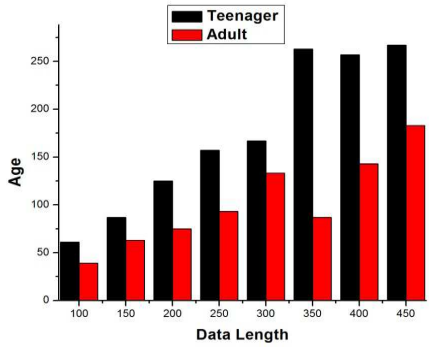
Fig. 10a illustrates the relation between the data set of different Facebook users and the gender attribute. The figure depicts that most of the males spent their time and share their emotion or personal views than female. For 450 users, about 42% female spent their time in face book, which is close to the male users' number. Fig. 10b includes the same relation when classifying the age group. The user age is considered in two categories: teenagers ($\leq$ 18 years) and adults ($>$ 18 years). It is established that the teenager users addicted in social media (Facebook) than adults. Most of the teenagers spent their time for sharing their personal feelings and emotion, while adults only use social media for communication or needed works. It is near about two third of the user are teenager and others are adult. Since, most of the users share their food habit, travelling interest, political or personal views by using Facebook. Fig. 10c depicts that most of the users use Facebook for sharing their daily activates emotion or fallings. The users choose Facebook to express their food habit. They share their desire food item or choose restaurant by using Facebook. However, the users have less interest for sharing the political views. About 10% of total users share their political views, while about 25% people prefer Facebook for choose historical or travelling place. They share their travelling zones with their friends or followers by using Facebook.

Moreover, the similar classification measurements are performed by using Naïve Bayes to classify the used data based on gender, age group and people view. Tab. IV illustrates the accuracy rate for different data sets. This accuracy is measured for the age/gender group classification using the Bigtable approach as well as the Naïve Bayes.

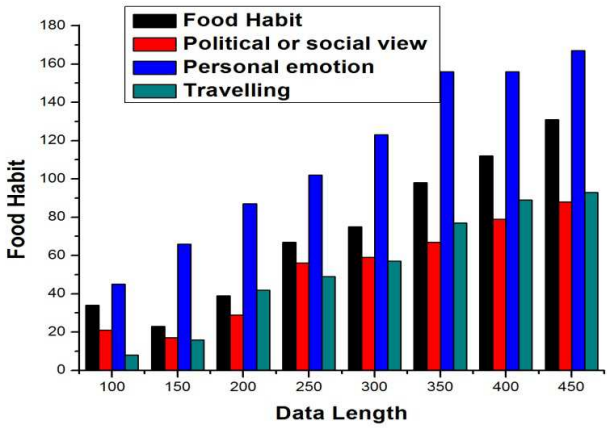
Tab. IV illustrates the different number of users in the first column. It is depicted that for 300 users, the Bigtable approach is more accurate by about 3.1%



(a) *Classification based on gender.*



(b) *Classification based on different age.*



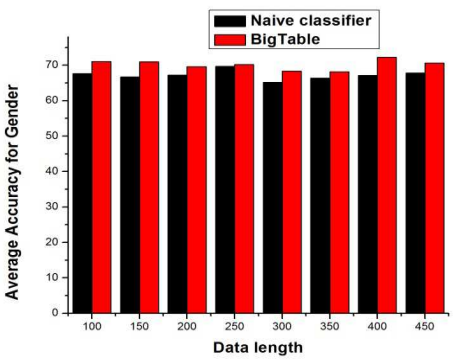
(c) *Classification based on people view.*

Fig. 10 *Classification of Facebook users.*

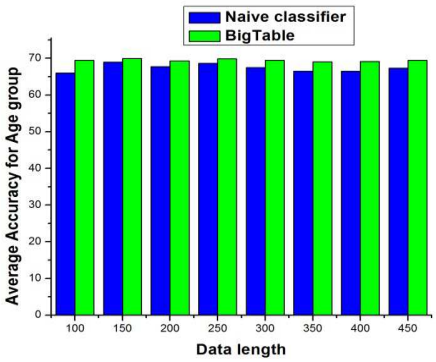
Data Length	BigTable Approach		Naïve Bayes		BigTable Approach		Naïve Bayes	
	Gender Group		Gender Group		Age Group		Age Group	
	Male [%]	Female [%]	Male [%]	Female [%]	Teenager [%]	Adult [%]	Teenager [%]	Adult [%]
100	70.0	71.2	68.0	67.3	71.5	67.5	64.7	67.3
150	71.5	70.5	64.7	68.7	69.4	70.5	69.2	68.7
200	69.8	69.4	69.2	65.2	69.2	69.4	70.2	65.2
250	71.2	69.2	70.2	69.2	71.0	68.8	68.0	69.2
300	67.8	68.8	65.7	64.7	71.5	67.5	64.7	70.2
350	68.9	67.5	67.3	65.4	68.9	69.2	67.3	65.7
400	69.5	70.5	68.2	66.1	69.5	68.8	68.2	64.7
450	70.3	70.9	69.3	66.3	71.5	67.5	69.3	65.4

Tab. IV *Accuracy measurement based on age group and gender.*

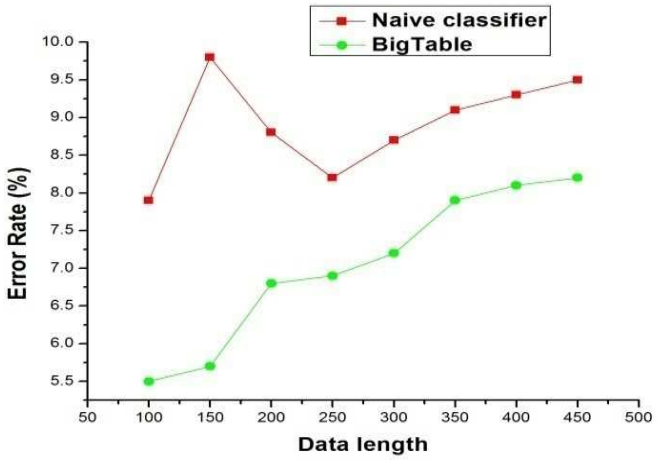
than the Naïve classifier for male user prediction. For similar data sets, the Bigtable achieves about 5.9% improved prediction accuracy than the Naïve Bayes. Moreover, the Bigtable is also achieves more classification accuracy for age group prediction than the Naïve classifier. Fig. 11 illustrates the Average accuracy rate as well as the classification error rate.



(a) *Average accuracy rate for gender attribute.*



(b) *Average accuracy rate for age group attribute.*



(c) *Measure error rate*

Fig. 11 *Average accuracy rate comparison.*

Fig. 11a depicts that the Naïve classifier and Bigtable achieves similar accuracies classification based on gender. Though, Bigtable is slightly accurate than the Naïve classifier due to parallel and small size of data processing. The average accuracy for gender domain of Bigtable and Naïve classifier are 71% and 69%; respectively. Fig. 11b depicts that the Naïve classifier is less accurate than the Bigtable for classification based on age group. Though, the Naïve classifier is a learning approach and good classifier, but its resultant accuracy is fall down when

data length is increased. The average accuracy for age domain of Bigtable and Naïve classifier is 69.5% and 68.0%; respectively.

The classification error rate is measured for both classifiers. Basically, the error rate measure in percentage that indicates the number of miss-classified member in a class. When every data element in the classified class is similar to the original data of the same class, the accuracy is closed to 100% and error rate is almost 0% (this result is only for a small portion of dataset) as illustrated in Fig. 11c. The Naïve Bayes error rate is higher than that obtained by the Bigtable approach. Naïve classifier's error rate is high due to high data volume processing. Bigtable is more accurate and less error rate than Naïve classifier. For concurrent data processing and same size of tabular data enhance the Bigtable accuracy and reduce error. For big data processing Bigtable is more optimal than other classifiers.

5.3 The aggregation results

Aggregation is a process in which information is gathered and generates a summary form for further statistical data analysis. A common aggregation indicates the collection of data in efficient and effective way for data analysis. It combines the different data blocks in efficient way. Data aggregations have several advantages for large data handling as it manages the user demand for better information, lower transaction cost, efficient data retrieve and query. The aggregation process is performed into several phases, namely i) the first phase, where the relative data are selected for analysis, ii) the second phase, at which the data is integrated from various sources, iii) the third phase that performs the data transformation for specific format, and iv) the last phase is for data reduction, which compresses the data for easiness of operation. The canonical correlation and maximum likelihood operation for the proposed method aggregation are used. The conditional operation is done using the ML to find out the data similarity. Data similarity indicates the similar elements of a class. The ML performs better operation for small data set. The table blocks are aggregated by using coefficient correlation and ML operation for faster transaction. The data sets are compared by using aggregation process using the ML and without aggregation. Tab. V illustrated the comparison of the classification time by using the ML aggregation and without aggregation.

In Tab. V, the data lengths refer to the number of Facebook users in different table blocks named tablet. The other two columns in the table indicate that classification time after aggregation by using the ML and without aggregation; respectively. Different number of table blocks is considered with minimum 100 Facebook users and maximum 450 users for the present experimental data sets. These table blocks are aggregated by using the ML, which performs better result for small data sets, while with the increased data sets; the classification time is increased for likelihood aggregation and without aggregation. The ML requires less time compared to the without aggregation case. For 250 user classification without aggregation, 423 seconds are necessary for its classification, however 397 seconds are required when using the ML for its classification. Thus, the ML is 6.14% faster than the without aggregation case when classifying 250 users. Fig. 12 demonstrates the required classification time using the ML aggregation and without aggregation.

Data Length	Canonical Correlation Aggregation	Without Aggregation
100	134	231
150	214	302
200	312	389
250	397	423
300	422	442
350	513	482
400	592	687
450	623	739

Tab. V *Classification time by using maximum likelihood aggregation and without aggregation.*

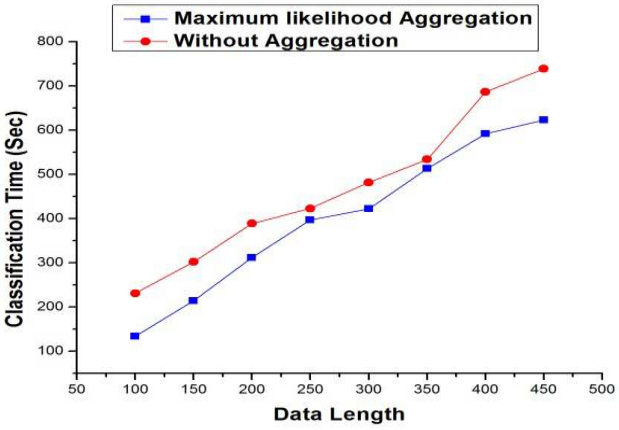


Fig. 12 *Classification time by using maximum likelihood aggregation and without aggregation.*

Fig. 12 depicts that the ML provides a feasible solution for classification, while without aggregation approach requires more time. With increasing the number of users, the aggregation is required more classification time of both approaches. For example, foe 450 users, the classification time is maximum and has the value of 623 seconds and 739 seconds for ML and without aggregation; respectively. The least classification time is measured for 100 users, when both aggregation and without aggregation are involved. It is established that for every data set, the ML needs less time compared to the without aggregation approach.

Moreover, the canonical correlation is employed, which is an aggregation approach. The canonical correlation is more effective in large data analysis. Thus, it is used to find and to identify association among the data sets. The canonical correlation is appropriated for accurate multiple regressions. It determines the cor-

relation variant among the data sets. In the proposed approach, the classification time is measured with canonical correlation aggregation and without aggregation as illustrated in Tab. VI.

Data Length	Canonical Correlation Aggregation	Without Aggregation
100	174	231
150	254	302
200	362	389
250	407	423
300	442	442
350	517	482
400	611	687
450	645	739

Tab. VI *Classification time by using maximum likelihood aggregation and canonical correlation.*

Tab. VI demonstrates that for 300 users, the canonical correlation approach is 8.30% faster than the without aggregation approach. With the data sets increase both approaches (canonical correlation and without aggregation) have increased classification time. In addition, the canonical correlation is 12.8% faster than without aggregation in the 450 users' case. Fig. 13 depicts the relationship between the classification time and the user data length for canonical correlation aggregation and without aggregation.

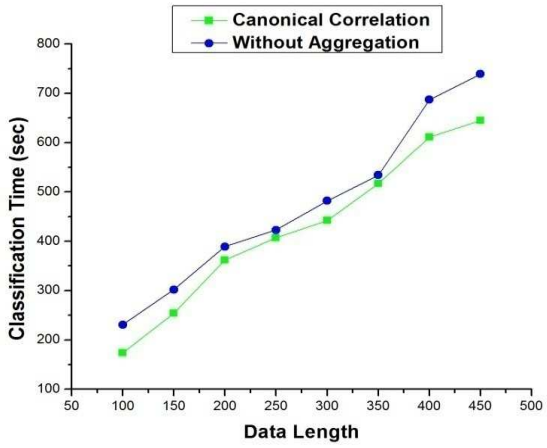


Fig. 13 *Classification time by using canonical correlation aggregation and without aggregation.*

Fig. 13 established that the canonical correlation performs better than without aggregation. In addition, the canonical correlation requires less time compared to the without aggregation case for every data length. For 450 users, the classification time is maximum and has the value of 445 seconds and 739 seconds for canonical correlation and without aggregation; respectively. The least classification time is attained with 100 users for both aggregation and without aggregation. It is clear that for every data set canonical correlation requires less time than without aggregation.

Moreover, Tab. VII depicts the classification time comparison for both aggregation approach, namely the canonical correlation and ML. The canonical correlation performs better with large data, while the ML handles the small size of data sets. For small data sets, the ML requires less time compared to the canonical correlation.

Data Length	Maximum likelihood Aggregation	Canonical Correlation Aggregation
100	134	174
150	214	254
200	312	362
250	397	407
300	422	442
350	513	517
400	592	611
450	623	645

Tab. VII *Classification time by using maximum likelihood aggregation and canonical correlation.*

In Tab. VII, table blocks are aggregated by using the ML and the canonical correlation. The ML performs outperforms the canonical correlation for small data sets. As the data sets are increased, the classification time is increased using the ML aggregation or canonical correlation aggregation. The ML requires less time than the canonical correlation aggregation. For 350 users' classification, the canonical correlation aggregation requires 517 seconds for its classification, while the ML requires 513 seconds. The ML is 0.78% faster than the without aggregation for 350 users' classification. Thus, it is clear that maximum likelihood is little bit faster than canonical correlation. Fig. 14 reflects the classification time for different data length by using canonical correlation and maximum likelihood aggregation.

Fig. 14 established that the classification time for every approach is increased with the data size increase. The ML is efficient with small data sets as it requires less time than the canonical correlation (as in the 100 users' case). However, the canonical correlation requires less classification time with large data sets (as in the 450 users' case).

The preceding results established the efficiency of the proposed system for different classes' classification of the Facebook data sets for monitoring. This monitoring system works by making inference and reasoning on grouped datasets. However,

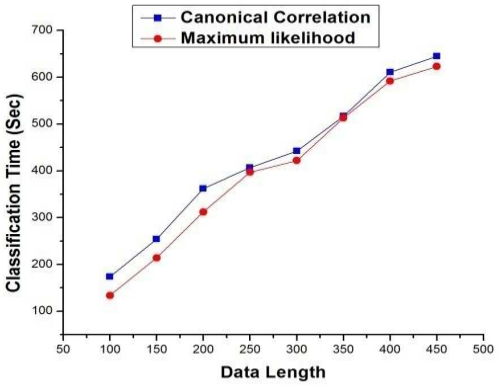


Fig. 14 *Classification time by using canonical correlation aggregation and maximum likelihood aggregation.*

dynamic source monitoring limitations are not considered. During monitoring, there might be loss of data or synchronizations. In that case it will be difficult to track the nodes. This drawback will be overcome in future. Moreover, comparisons between Bigtable mapping and MapReducing mapping will be verified in future work.

6. Conclusions

This research is focused on grouping the datasets collected from active Facebook users to monitor the interactions among thousands of users. Both grouping and tracking are done for handling big datasets with mapping based machine learning techniques such as Fisher Discriminant Analysis, Canonical correlations, Maximum likelihood and Dynamic Source Monitoring. The FDA was efficient for redundant datasets removal from the training datasets. Pre-processed datasets are then sent to Bigtable phase for efficient mapping. Mapping was used as Bigtable orientation where collected and processed data were sent to group automatically in the specific database. Bigtable also permitted the refreshment of the database tables to repeatedly update with certain period of time. Maximum likelihood and canonical correlations aggregate the mapped data to certain group. For large datasets, canonical correlation outperformed the ML analysis. Alternatively, for marginal and small datasets, the ML was superior to the canonical correlations. Automated Facebook nodes and interactions among users were monitored by dynamic source monitoring.

References

[1] Dmr yelp statistics report. Tech. rep., March 2016. Available from: http://expandedramblings.com/index.php/downloads/dmr-yelp-statistic-report/

[2] AMATRIAIN X. Big & Personal: Data and Models Behind Netflix Recommendations. In: *Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine '13*, ACM, 2013, pp. 1–6, doi: 10.1145/2501221.2501222.

[3] ANDERSON P. What is Web 2.0? Ideas, technologies and implications for education. *JISC Technology & Standards Watch 1*, 2007, pp. 1–64. Available from: http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf.

[4] BANGA G., DRUSCHEL P., MOGUL J.C. Resource containers: a new facility for resource management in server systems. In: *OSDI '99: Proceedings of the third symposium on Operating systems design and implementation*, Berkeley, CA, USA, USENIX Association, 1999, pp. 45–58, doi: 10.1145/224056.225831.

[5] BARU C.K., FECTEAU G., GOYAL A., HSIAO H., JHINGRAN A., PADMANABHAN S., COPELAND G.P., WILSON W.G. Db2 parallel edition. *IBM Syst. J.* 1995, 34(2), pp. 292–322, doi: 10.1147/sj.342.0292.

[6] BLANCHARD M., METCALF A., DEGNEY J., HERMAN H., BURNS J. Rethinking the digital divide: findings from a study of marginalised young people's information communication technology (ict) use. *Youth Studies Australia*. 2008, 27(4), p. 35. Available from: http://academics.hamilton.edu/ebs/pdf/rtdd.pdf.

[7] BOYD D.M., ELLISON N.B. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*. 2007, 13(1), pp. 210–230, doi: 10.1111/j.1083-6101.2007.00393.x

[8] CHANG F., DEAN J., GHEMAWAT S., HSIEH W.C., WALLACH D.A., BURROWS M., CHANDRA T., FIKES A., GRUBER R.E. Bigtable: A distributed storage system for structured data. In *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation - Volume 7*, Berkeley, CA, USA, OSDI '06, USENIX Association, 2006, pp. 15–15. Available from: http://dl.acm.org/citation.cfm?id=1267308.1267323

[9] CORDEIRO M., SARMENTO R.P., GAMA J. Dynamic community detection in evolving networks using locality modularity optimization. *Social Network Analysis and Mining*. 2016, 6(1), p. 15 doi: 10.1007/s13278-016-0325-1.

[10] CUESTA C.E., MARTÍNEZ-PRIETO M.A., FERNÁNDEZ J.D. *Towards an Architecture for Managing Big Semantic Data in Real-Time*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 45–53 doi: 10.1007/978-3-642-39031-9_5.

[11] DEMCHENKO Y., NGO C., MEMBREY P. Architecture framework and components for the big data ecosystem. Sne technical report, University of Amsterdam, September, 2013. Available from: http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf

[12] DONATH J., BOYD D. Public displays of connection. *BT Technology Journal*. 2004, 22(4), pp. 71–82, doi: 10.1023/B:BTTJ.0000047585.06264.cc.

[13] DUDA R.O., HART P.E., STORK D.G. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[14] FISHER R.A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. 1936, 7, pp. 179–188, doi: 10.1111/j.1469-1809.1936.tb02137.x.

[15] FITZGERALD R., BARRASS S., CAMPBELL J., HINTON S., RYAN Y., WHITELAW M., BRUNS A., MILES A., STEELE J., MCGINNESS N. Digital learning communities (dlc) : investigating the application of social software to support networked learning (cg6-36), 2009. Available from: http://eprints.qut.edu.au/18476/1/__staffhome.qut.edu.au_staffgroupb%24_bozzetto_Documents_2011009485.pdf.

[16] FUKUNAGA K. *Introduction to statistical pattern recognition*, second ed. Computer Science and Scientific Computing. Academic Press, 1990.

[17] GROSS R., ACQUISTI A. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society* New York, NY, USA, WPES '05, ACM, 2005, pp. 71–80, doi: 10.1145/1102199.1102214.

[18] HE X., CAI D., NIYOGI P. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems*, Y. Weiss, B. Schölkopf, and J. C. Platt, Eds., MIT Press. 2005, 18, pp. 507–514. Available from: http://papers.nips.cc/paper/2909-laplacian-score-for-feature-selection.pdf.

[19] HILLIER L., HARRISON L. Building realities less limited than their own: Young people practising same-sex attraction on the internet. *Sexualities*. 2007, 10(1), pp. 82–100, doi: 10.1177/1363460707072956.

[20] KAMAL M.S., KHAN M.I. Chapman-kolmogorov equations for global ppis with discriminant-em. *International Journal of Biomathematics*. 2014, 7(05), doi: 10.1142/S1793524514500533.

[21] KAMAL S., AREFIN M.S. Impact analysis of Facebook in family bonding. *Social Network Analysis and Mining*. 2016, 6(1), doi: 10.1007/s13278-015-0314-9.

[22] KANN M.E., BERRY J., GRANT C., ZAGER P. The internet and youth political participation. *First Monday*. 2007, 12(8), doi: 10.5210/fm.v12i8.1977.

[23] LIAO S., ZHU X., LEI Z., ZHANG L., LI S.Z. Learning multi-scale block local binary patterns for face recognition. In: *Proceedings of the 2007 International Conference on Advances in Biometrics,* Berlin, Heidelberg ICB'07, Springer-Verlag. 2007, pp. 828–837, doi: 10.1007/978-3-540-74549-5_87.

[24] MAIER M. *Towards a big data reference architecture.* Master's thesis, Eindhoven University of Technology, 2013. Available from: www.win.tue.nl/~gfletche/Maier_MSc_thesis.pdf

[25] MCGRATH H. *Young people and technology: a review of the current literature.* South Melbourne, Vic.: Alannah and Madeline Foundation, 2009. Available from:http://apo.org.au/node/19311

[26] MISHNE G., DALTON J., LI Z., SHARMA A., LIN J. Fast data in the era of big data: Twitter's real-time related query suggestion architecture. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data,* New York, NY, USA, SIGMOD '13, ACM. 2013, pp. 1147–1158, doi: 10.1145/2463676.2465290.

[27] MONTGOMERY K.C. *Generation Digital: politics, commerce and childhood in the age of the internet.* MIT press, 2007.

[28] NOTLEY T.M., TACCHI J.A. Online youth networks: Researching the experiences of 'peripheral' young people in using new media tools for creative participation and representation. *3CMedia: Journal of Community, Citizen's and Third Sector Media and Communication 1*. 2005, 1, pp. 73–81. Available from: http://eprints.qut.edu.au/3788/

[29] SCHMIDT R., MÖHRING M. Strategic alignment of cloud-based architectures for big data. In *2013 17th IEEE International Enterprise Distributed Object Computing Conference Workshops*. 2013, pp. 136–143, doi: 10.1109/EDOCW.2013.22.

[30] SIMONCELLI D., DUSI M., GRINGOLI F., NICCOLINI S. Stream-monitoring with blockmon: Convergence of network measurements and data analytics platforms. *SIGCOMM Comput. Commun. Rev.* 2013, 43(2), pp. 29–36, doi: 10.1145/2479957.2479962.

[31] SMITH A., SCHLOZMAN K.L., VERBA S., BRADY H. The internet and civic engagement, 2009. Available from: http://www.pewinternet.org/2009/09/01/the-internet-and-civic-engagement/

[32] STEPHENS-REICHER J., METCALF A., BLANCHARD M., MANGAN C., BURNS J. Reaching the hard-to-reach: How information communication technologies can reach young people at greater risk of mental health difficulties. *Australasian Psychiatry*. 2011, 19(1), pp. S58–S61, doi: 10.3109/10398562.2011.583077.

[33] SUMBALY R., KREPS J., SHAH S. The big data ecosystem at linkedin. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data,* New York, USA, SIGMOD '13, ACM. 2013, pp. 1125–1134, doi: 10.1145/2463676.2463707.

[34] THIRD AND RICHARDSON. Connecting, supporting and empowering young people living with chronic illness and disability. The Livewire Online Community, Report prepared for the Starlight Children's Foundation, 2010.

[35] VALTYSSON B. Access culture: Web 2.0 and cultural participation. *International Journal of Cultural Policy*. 2010, 16(2), 200–214, doi: 10.1080/10286630902902954.

[36] VENKATESAN S., OLESHCHUK V.A., CHELLAPPA, C., PRAKASH S. Analysis of key management protocols for social networks. *Social Network Analysis and Mining*. 2015, 6(1), 3, doi: 10.1007/s13278-015-0310-0.

[37] VROMEN A. Inclusion through voice: Youth participation in government and community decision-making. In: *Social Inclusion and Youth Workshop Proceedings* Melbourne, Brotherhood of St Laurence, 2008. Available from: http://library.bsl.org.au/jspui/bitstream/1/6663/1/Vromen_paper_29Oct08.pdf

[38] WESTON J., WATKINS C. Multi-class support vector machines. Tech. rep., Department of Computer Science, University of London, 1998. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.50.9594&rep=rep1&type=pdf

[39] ZENG C., JIANG Y., ZHENG L., LI J., LI L., LI H., SHEN C., ZHOU W., LI T., DUAN B., LEI M., WANG P. Fiu-miner: A fast, integrated, and user-friendly system for data mining in distributed environment. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, KDD '13, 2013, ACM, pp. 1506–1509, doi: 10.1145/2487575.2487714.