

# Improved Information Retrieval IIR-Model for Pile of data sets in Large Repositories

Dr. Tarigoppula V S Sriram<sup>1</sup>, Dr. Madhavi Kolukuluri<sup>2</sup>, Dr. V. V. Hari Babu<sup>3</sup>, Mr. P Ratna Kumar<sup>4</sup>

Associate Professor, Department of Computer Science and Engineering, NSRIT, Visakhapatnam, AP, India<sup>1&2</sup>  
Department of Physics, Bapatla Engineering College, Bapatla, AP, India<sup>3</sup>, Assistant Professor, Koneru Lakshmaiah Education  
Foundation, Hyderabad<sup>4</sup> Email: rameesis@gmail.com<sup>1</sup>, kolukulurimadhavi@gmail.com<sup>2</sup>, vvhbphy@gmail.com<sup>3</sup>,  
rk30111972@klh.edu.in<sup>4</sup>

## ABSTRACT

Big Data Analytics is the area which is growing very fast around the world. Billions of dollars of money is spent on research and development by American government. America like nations they depend upon the computer systems, as per growth of population in future, data will be more in above GBs. After some days or months or years later it will reach PBs also. This mass of data is stored in the larger repositories and these repositories are also growing day by day and it will continue like that, it is never ending process till the scope of systems exist. Retrieving data from the large data tombs or data buckets or large repositories is also a challenge. Improved Information Retrieval (IIR) should fast and accurate. Data will spread in distributed databases for any organizations around the globe. End user will interact with the web server and web server stores the user details in an index that is available in database relation, that process is said to be WUM. Challenge is chased by maintaining index with a format and storing information in sequential manner. After issues of identity, software will chose one database where information is linked with linked list in sorted manner. User interaction is dumped into storage and some information is transformed into large data sets. The information can be end user search, or showing interest towards product and his bills. Based on store index we can retrieve information fast and accurate. This process is said to be improved information retrieval (IIR), application end support is required to add message to persistent data.

**Keywords:** Improved Information retrieval, IIR, large data sets, index, big data, WUM;

## 1. INTRODUCTION

Information retrieval is basic task that should be considered under large data sets. Everyday large and huge information is added to the volumes. Now day's volumes are added with web network data. Large data is shared using websites, around the world websites are 1,518,207,412 according to Net Craft, January 2019. According to web server survey compared 1,805,260,010 in January 2018, the data volumes are 800 billion in 1999, 11.5 billion in 2005 and 18 billion in 2019. Information retrieval challenge based on interaction, business intelligence (more than TBs) and sentiment analysis. History of image and images are started from bytes to KB, KBs to MBs, MBs to GBs, GBs to TBs, TBs to PBs, PBs to more than. Volumes are increasing daily; they should not be turning to data tombs. Main storage of data takes place on web. Facebook daily uploads are 100 TBs, Twitter process of 4 million tweets daily, linkedin and g+ of 10 TB daily base and youtube each minute 48 fresh videos are updates around average size of 150MB. Along with this data every device can generate other information related to the system. Making use of that information is focused on this paper. We have to use various techniques to information retrieval on sophisticated large devices and map them to understandable way. It is big head-ache to software users to retrieve or manage data. Data management task is a big deal to software professionals.

## 2. EXISTING SYSTEM

We have servers, client machines and other electronic devices which are interconnected to do some facilities to the users. Similarly there is lot of data which is useful to us is generated through networks, servers, machines and systems. Such data can be used more informatively for better understanding of the information retrieval. If they are used properly then we can have more information understandable system. Technology is growing very fast; everyday millions of data is added to it. Business and technology and user needs are parallel they are growing. Inceptions of any business there will be no need of much technology, as it will growing and increasing in branches at various places, the requirement of technology is needed to help their business as friend. Same way

customers will increase, dealers will increase, suppliers will increase and other issues. Where memory is growing and query applied to it may take more time give resultant. As business organization requires information from the resultant data very fast. The information what I required is from billing system, where it stores bill\_id, item\_name, unit\_price, price and total\_amount. This persistent data is stored in the billing\_system table. Let us assume that there are 20,000 locations around the globe, 720 records per month (hourly measurements, approximately 720 hours per month), 120 months (for 10 years back) and many years into the future. A simple calculation yields the following results.

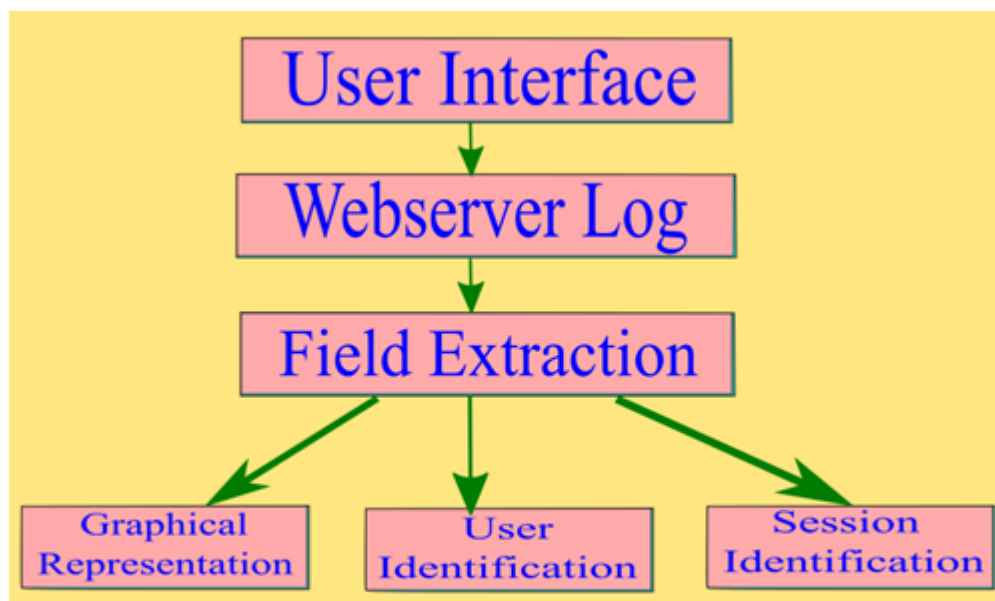


Fig. 1 Preprocessing under WUM.

20,000 locations X 720 records X 120 months (10 years back) = 1,728,000,000 records. These are the past records; new records will be imported monthly, so that's approximately 20,000 X 720 = 14,400,000 new records per month. The total locations will steadily grow as well. Data mining is an excellent topic that is related to the computer science and engineering, data is growing day by day and it should be mined for information or message. WUM is one of the data mining techniques used for web related business to know the user interests, most searched or ordered products and other information on persistent data. WUM can be used for better understanding of the customer or end user who uses the system for he/she needs. Preprocessing plays a key role in the WUM. Preprocessing extract raw data and converts it to the necessary information using pattern discovery. Preprocessing related WUM is shown in the below fig. 1. The purpose of data preprocessing is to improve quality and increase mining accuracy. There are two phases in this process: first phase is to collect the raw web log files and place it in a relational database table in order to make it available for mining. This process is to clean the raw data. Second phase of this process is given as follows: Process 1: Extract the web log files that are collected from web server. Process 2: Clean the web log files and remove duplicates. Process 3: Collect the data and paste it into a relational database table or data warehouse and reduce to be frequency analysis to create summary reports. After having data frequency we can use it for k-means or clusters or other way. Statistical information of the data that collected from web server is shown in below fig. 2. We can extract useful data from web server and we can use it for study of customers. Whatever data we are collecting can be handled very well if we organize them carefully.

Date=2012-03-22	0.0333	0	0.037
Date=2012-03-23	0.0333	0	0.037
Date=2012-03-24	0.0333	0	0.037
Maximum actions in one visit	60.1	74.3333	58.5185
Actions	2528.4333	3896	2376.4815
Unique visitors	807.7667	1070	778.6296
Visits	927.8	1254.6667	891.4815
Bounce Rate	63.9	61	64.2222
Actions per Visit	2.7133	3.1	2.6704
Avg. Visit Duration (in seconds)	182.6667	221.3333	178.3704
Actions by Returning Visits	1094.7667	1655.6667	1032.4444
Unique returning visitors	209.6	286	201.1111
Returning Visits	288.0333	407.6667	274.7407
Bounce Rate for Returning Visits	47.9333	45	48.2593
Avg. Actions per Returning Visit	3.79	4.0667	3.7593
Avg. Duration of a Returning Visit (in sec)	306.5667	364	300.1852
Conversions	70.3667	107.3333	66.2593
Visits with Conversions	47.8	73.6667	44.9259
Conversion Rate	5.1663	5.8633	5.0889
Revenue	123.8333	185.3333	117
Outlinks	94.2667	145.3333	88.5926
Pageviews	2231.0667	3735.6667	2063.8889
Unique Outlinks	289.1333	135.6667	306.1852
Unique Pageviews	1738.9667	2831.3333	1617.5926
Downloads	9.9	0	11
Unique Downloads	0	0	0

**Fig. 2 K-means with neural networks**

We have server logs and system logs, now what I am using system will also have logs, but we do not know about the logs. They are not in the readable format. But they are important. It can be a computing device or a non-computing device. It contains information about when we logged and other device related information.

#### 2.1 Server logs

Source of IP traffic can be easily notified when end user or customer interact with client system via server. Where they are located and where they access from geographically.

Security treats can be easily notified, when there is any vulnerability to their servers to their servers to the network?

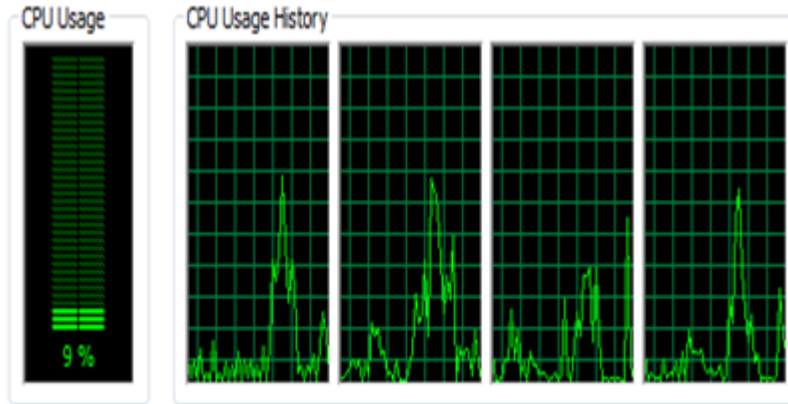
#### 2.2 Network Vulnerability

Any basic system will have log which stores at one location and directory. But we can read those logs and they are not in the readable format. They have some really important stored in them.

#### 2.3 System performance

How people are accessing your CPU and at what time, what they try to access and which application they want accesses. How your system is responding to the access. Any transaction or any operation that is happening on your device at gets registered. It is treasure; it is like a treasure hunt game.

CPU usage & load is shown in form of graph, as shown in below fig. 3. CPU usage and history can be shown for a single machine or a server or a web server.



**Fig. 3 CPU usage and history.**

Performance of the total CPU is shown in a graph manner. If it is dual core processor, then each processor performance is shown in the each graph. Task manager, previously it is windows task manager, gives the total management of the system, free, fast, open source, daily backups and customizable unlimited users. User access logs are shown figure (Fig. 4) by task manager, by displaying which process is used by which user. User access log (UAL) is a striking feature of windows server, which shows the client and client using process. This helps in windows server to administrate easily client computers.

File Options View						
Processes Performance App history Startup Users Details Services						
Name	PID	Status	User name	CPU	Memory (p...	Description
armsvc.exe	3452	Running	SYSTEM	00	796 K	Adobe Acrobat Update Service
aswEngSvc.exe	3704	Running	SYSTEM	00	7,912 K	Antivirus engine server
aswidsagent.exe	2844	Running	SYSTEM	00	16,516 K	Avast Behavior Shield
audiodg.exe	4904	Running	LOCAL SE...	00	3,232 K	Windows Audio Device Graph Isolation
AvastBrowser.exe	1124	Running	Anitha	00	39,268 K	Avast Secure Browser
AvastBrowser.exe	8364	Running	Anitha	00	548 K	Avast Secure Browser
AvastBrowser.exe	8892	Running	Anitha	00	192 K	Avast Secure Browser
AvastBrowser.exe	5660	Running	Anitha	00	21,848 K	Avast Secure Browser
AvastBrowser.exe	9172	Running	Anitha	00	16,436 K	Avast Secure Browser
AvastBrowser.exe	1936	Running	Anitha	00	23,992 K	Avast Secure Browser
AvastBrowser.exe	8800	Running	Anitha	00	11,224 K	Avast Secure Browser
AvastBrowser.exe	1600	Running	Anitha	00	20,776 K	Avast Secure Browser
AvastBrowser.exe	680	Running	Anitha	00	29,676 K	Avast Secure Browser
AvastBrowser.exe	9000	Running	Anitha	00	12,300 K	Avast Secure Browser
AvastBrowser.exe	412	Running	Anitha	00	6,840 K	Avast Secure Browser
AvastBrowser.exe	10056	Running	Anitha	00	15,924 K	Avast Secure Browser
AvastBrowser.exe	1676	Running	Anitha	00	12,752 K	Avast Secure Browser
AvastBrowser.exe	10144	Running	Anitha	00	13,796 K	Avast Secure Browser
AvastBrowser.exe	7204	Running	Anitha	00	1,088 K	Avast Secure Browser
AvastBrowser.exe	9448	Running	Anitha	00	44,764 K	Avast Secure Browser
AvastBrowser.exe	3144	Running	Anitha	00	21,452 K	Avast Secure Browser
AvastBrowser.exe	1308	Running	Anitha	00	44,348 K	Avast Secure Browser
AvastBrowser.exe	2568	Running	Anitha	00	7,608 K	Avast Secure Browser
AvastBrowser.exe	3624	Running	Anitha	00	7,468 K	Avast Secure Browser
AvastBrowser.exe	8204	Running	Anitha	00	35,812 K	Avast Secure Browser
AvastBrowser.exe	7072	Running	Anitha	00	71,580 K	Avast Secure Browser
AvastBrowserCrashH...	3344	Running	SYSTEM	00	108 K	Avast Browser Update
AvastBrowserCrashH...	3360	Running	SYSTEM	00	40 K	Avast Browser Update

**Fig. 4 User access details**

## 2.4 Application performance monitoring

Every single transaction and single thing that happens is stored on that device. It is like that it is a golden nugget where we have to dig them, same as coal or gold mining. It is very tough also. When we want get more accurate answers we want to have dig deeper for actual data. Then also it will be useful to us. Any transaction or any operations that happens on your device is that gets registered on your logs. We can get actual data just lines down to it. Now we can go to the next level because we known what are the problems or when you get actual data. Based on operations we can perform what will help your business. It is like that what you understand of business or understand data is called operational intelligence. Logs are the go-to achieves for gaining company-wide operational intelligence. Exploring of the logs is important where internet is ruling. Exploring this logs and understanding of these logs is not so easy. If you take facebook we can have a log who visited our website, the area where we can get visitor information is "initialchatfriendslist" as shown in below fig. 5.

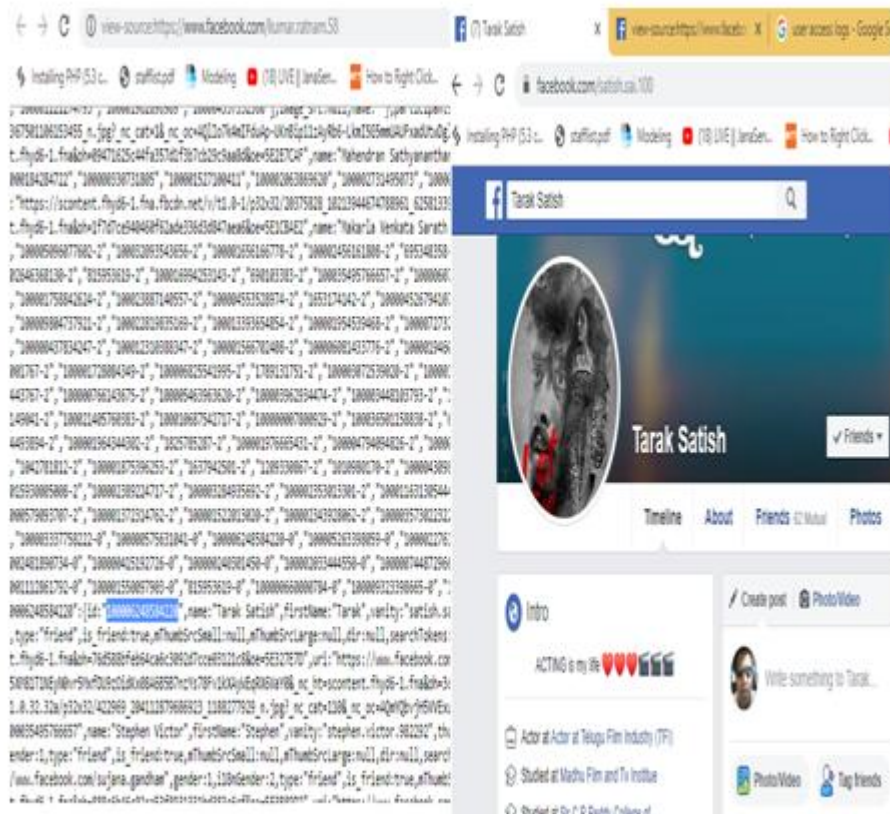


Fig. 5 Visitor list from face book.

## 2.5 Security

1000s of such logs are generated every minute, check security threats in real-time, analyze business metrics in real-time. They are not in readable form. There may be security breach. You know that there is a security breach. How it is present in your logs. How fix this problem and read these logs. What is the meaning of this log? Definite we do not know about it. These logs are generated massively in thousands and more to a system or servers.

## 2.6 Example

Likewise our devices give us some back-end information related to the device. If we handle them carefully, we can make use them in our business intelligence. Especially web server usage plays a key role. Most of the end users they use web site to satisfy their needs. Browse various web sites to get their requirements. A business entrepreneur can easily improve his business by knowing customer needs.

Let us take e-commerce website or amazon website or flipkart website where they are so many users who want to interact with website. Then there IP address is recorded along with what the likes they have opened and what are actions they performed on that website and they are record stored in the logs in web server. At that time we may have security breach and we have to fix that but it is impossible to fix that (check security threats in real-time) or we cannot analyze business metrics in real-time. What customers are searching for other things when they are online. There should be a tool which can understand your logs and give you information about those things in a simple manner.

Fig. 6 for machine generated big data and the ultimate log collection and analysis tool.



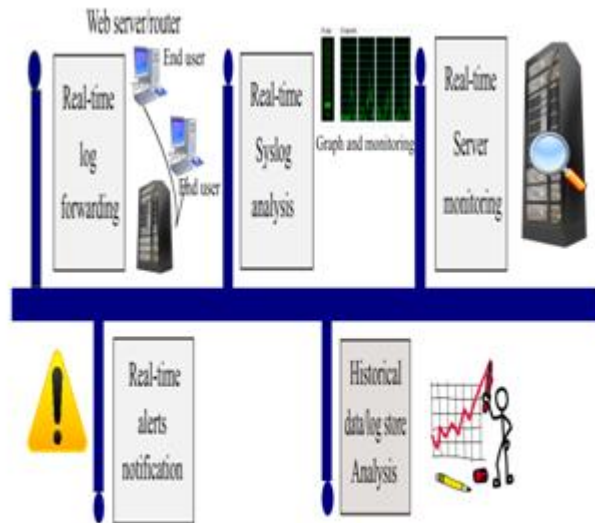


Fig. 6 Machine generated big data.

Every system or device will give log information, which is not user readable form. That information of log is very much important to the business or managers or higher authority in the organization. By just knowing the details, the business can be improved. Here the concern is related to web server. There are so many information can be provided by the web server. We can identify the users IP address and from which geographical area they are interacting with web server. Syslog means system logging protocol, it is a standard protocol used to send system log or event message(s) to a declared server. If it so then it is spelled as syslog server. We can monitor the graph related to the syslog analysis. Real-time server monitoring can provide market analysis. It gives information about customer interest, that means what customer want to buy and what is purpose of watching website. What products really customer likes from the website. Website may collect the details of the customer and when any offers related to the product may be intimated to the customer. Historical data/ log store analysis is related storage of bulk data. What, where data is coming from real-time is send to the historical data and their we can analysis when it appears in the real-time. It can be stored in a bucket. Bucket contains 30 days data. We can have buckets of storage for every 30 days. We can have more than one bucket and delete data after filling of all buckets. It stores in compressed data storage.

Security threats, it can give real-time alerts and notifications and security threat. Something is going to happen that means some person is accessing your server from somewhere. Where he is not reliable person, so immediately software should recognizes the person and throw an alert. We can have a system to monitor CPU performance that cross threshold and may be crash, system can be attend properly.

Where lot of logs are generated everyday because lot of customers, lot of sms, lot of data usage, lot of logs and everything should be stored. Storing them and using them efficiently is related to big data.

Lot of information provided by the device will be stored in buckets; this information can be used efficiently in making business growth via through machine learning.

### 3 Literature Survey

Mrs. Addanki Ramya(2012) et. al. [1] focused on customer related information on the web pages by web usage mining. Now a day's business websites are growing tremendously in the world. Every business owner think that there business should be the first in world market. If they use little business intelligence they can easily have great profits. Here everything is around customers, knowing the customer pulse is important to grow their business. No business owner can go to customers who are geographically located. Customers want products to his doors and business owners want to deliver their product to the customer door steps. It is a cyclic sense on demand and produce. There is only one way to know customers pulse fast, where he/she located, what their interests and so on by web usage mining. Every customers on online have to connect to server to order his/her products, search product related information any time. We can collect the data information from the web site server. This paper focuses on clustering and k-means with neural networks.

K R Sunitha et. al. [2] focused on preprocessing of web data, preprocessing is an important process in data mining that is where it makes data understandable. There are three important phases in it 1. Data cleaning, 2. Data transformation and 3. Data reduction. Really the web data like to have so many errors and required corrections. Hence data preprocessing is a technique

which can make it readable format. Author focuses on two things, 1. Extracting useful data from weblog and 2. Uses NASA web server data for experimental purpose.

Mr. Sanjay Babu Thakare et. al. [3] conforms that preprocessing is very much important for web usage mining. Preprocessing is an important technique where it is available from the web site(s). Authors conforms that no experimental process or lab required in getting the details of weblog. They are easily available from the web server. Anybody can get web data log and can easily identify user from which geographical area he/she belongs and other things. This paper helps in meaningful mining patterns.

Ms. Dipa Dixit et. al [4] today's business owner use website to promote their business around the world. For every business there will be a web server which produces lot of noisy, unambiguous data, errors, missing and inconsistent data. This is because of their huge size of data and preprocessing is necessary on the data to make it in the readable format. Author focus on two approaches XML and text file to give readable data.

Renata Ivancsy et.al. [5] Web server gives lot of huge data from its side. Which contains lot of undiscovered information? This paper focuses on how to discover hidden information in weblog from huge data. This information may be from different users who use the web site. There are lots of methods in discovering hidden information. Data mining is not only a technique to data clean and others. It can be used to identify what are different patterns in knowing customer interest and how many of them have request for various products.

Bikash Mukopadhyay et. al. [6] Data mining is a concept related to mining large data from huge data tomb. The data which is in the data tombs have relationship with other data in the tomb. For example, a supermarket contains large data, huge and lump sum data is added every minute. And retrieval of useful information from those tombs is difficult after data preprocessing. For knowledge discovery of information from huge preprocessed data is maintained as cluster, finding association rules, categorization and statistical analysis.

Mr. S K Chaitanya Rudraraju et. al. [7] Technology is growing very fast but something is lacking in fast information retrieval from large repositories. Paper focuses on preprocessing huge data like cleaning, transforming and compression or reduction. Collect data from large repositories, clean data, remove errors and unknown data is replaced with mid value from 0 to 1, data is compressed as follow-up of the reduction. Data is framed as columnar one after the other, as ants. Based on columnar retrieving is possible.

Shadab Irfan et. al. [8] Data preprocessing is the technique used to extract hidden information from the huge data which is collected from the weblog. It is very difficult in handling such huge data from the weblog. This paper focus on different aspects like WCM, WSM and WUM.

Jan O Pedersen et. al. [9] End user inputs one or more query words which are used to search related information data and display the matching query element as fast as it can. The query words are reformulate the search key, reformulated query performs a subsequent search through the document corpus. Additional non-stop-words are added for each phrase and aligned with each other in a columniation manner.

Niranjana Lal et. al. [10] Information retrieval is not limited to database and text search, it had advanced to multimedia like video, audio, structured and unstructured data, scientific based complex data. Information retrieval from such sources is very difficult and tough job. The basic idea of the paper is to information integration which may be used in dataspace and with heterogeneous data.

#### **4 Proposed System**

Customer plays main role in this concept. Customer can be geographical located anywhere in the globe. When his/her identification comes first time to web server, then web server will identify him by location and ip address as new person. From web log files we store all information about him/her and his active on the web page. Every system will have information of it. That information is grabbed for our domain knowledge purpose. Here the web site may get his details at prior but at some point we get domain information about the person. Attributes may be related to person name, gender, area of interest and some other parameters.

Unique identity is create to each customer on his submit of domain information about he/she. Based on ip past information is added. For every web visit he/she details are extracted from web log files according to WUM. Whatever information is stored in the form of memory storage buckets. Because there will be bulk of data, where customer interacts with web site. Now a day's people depend upon the web site or mobile apps for their daily needs. It can be desktop or laptop or mobile the user should interact with the web server. Web servers have capability of storing information to web log files. This information can be customer details, his/her buying interests, most buying product from the supermarket or giant hyper market, most searched item on web site, highest bill paid and place, salary, age group and soon. Every time customer or end user will not be using search bar as per his/her requirements he/she can use thumbnails for search but that search data is important for business horses. Where we can get that valuable data? It is available behind the screen with errors and lot of noise. If we remove impurities, than we can use the refined data for better business usage in understandable manner. Every data from web logs files is cleaned, transformed and

comprised to relational database table. Tables are arranged in snowflake manner, which is easy to mine. Day to day data is added to the relations and this data is made available for machine learning and is said to be data sets.

Machine learning is the word now-a-days listening more in the society. It is an upcoming domain in computer science. Machine learning is very powerful weapon which gives accurate information based on data sets. We provide available information data to system or machine and provide test data to check how far machine learns your data. Based on some decisions we can predict the future business. We can study the customer's needs and plan the business. Business modeling and machine learning can be combined to get better results. Data sets or datasets is the word heard now and then, data set or dataset is related to the collection of data. They may be scattered under one table or multiple tables. Same as a relational table, this data set table also contains rows and columns. Data sets are related to only numeric data. Raw data which is available from the web log files is cleaned, and then it is transformed from one format to another by removing noise and other things. After transforming, data is reduced or comprised to a fixed range.

$$data = xx' + yy' + zz' \quad --1$$

According to equation 1, data is merged with noise  $x'$ ,  $y'$  and  $z'$ . In data cleaning processes it is cleaned to the following equation 2.

$$data' = x + y + z \quad --2$$

This data is transformed to other data and it is brought to relational table as shown in below figure (Fig. 7),  $x$  is transformed  $a$ ,  $y$  is transformed to  $b$  and  $z$  is transformed to  $c$  with compressed data as shown in equation 3.

$$relation = a + b + c \quad --3$$

Data that is available from the data sets is compressed data based on frequency. Suppose business head want to know total bill made on the each chocolate brand around globe. The data set can be like this as shown in fig. 9, give the data sets of a supermarket which sold chocolates around the London city (LN). Reducer receives  $k$  rows of  $M$  and  $k$  columns of  $N$ , then  $q = 2nk$ , and  $k^2$  outputs are covered. That is,  $g(q)$ , the maximum number of outputs covered by the reducer that receives  $q$  inputs, is  $q^2/4n^2$ .

$$\sum_{i=1}^k qi^2 / 4n^2 \geq n^2 \quad --4$$

Horizontal we have rows of sum of customers from 1 to 49, sum of customers from 50 to 499 and sum of the customers above 500 and it may have columns so on. Second horizontal give the sum from pervious sheet.

Row labels	Sum of group customers 1 to 49	Sum of group customers 50 to 499	Sum of group customers above 500	*****
+	1024	1548	702	*****
- REG1 West				
- New England				
- Item LN				
Dollar/LN	203	7	245	
Wisp/LN	1023	456	2800	
Mars/LN	45	11	789	
Mars Duo/LN	7852	4528	1114	
Wix/LN	927	45	84	
n&M/LN	12	255	941	
*****	*****	*****	*****	*****

**Fig. 7 Example Data set of a supermarket.**

Improved Information Retrieval plays a key role in this mechanism and model for IIR is shown in the below fig. 8. IR used to find relevant information from the database. Index is that which maintains a compressed version of your documents in relational database. It should be very fast in searching data on database.



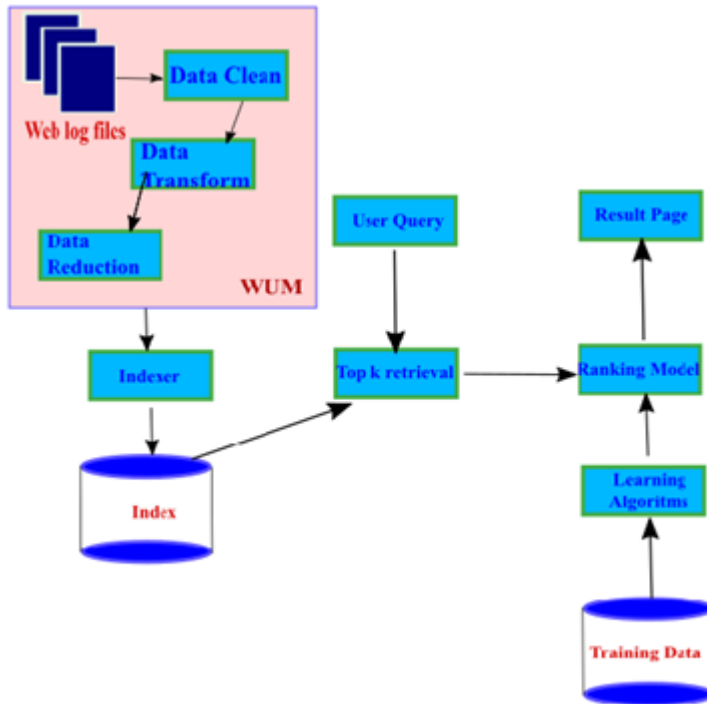


Fig. 8 Model for Improved Information Retrieval

#### IIR Algorithm

- User interacts with web browser and web server captures user needs with noise under one identity like customer or item or bill and so on (index).
- Step 1: Data cleaning for removing the noise.
- Step 2: Data transform.  
 At this stage we can go for normalize the data to make it available for machine training and for decision trees or else we can maintain the same data which may not be working properly.
- Step 3: Get the data to the relational tables with index and which is same for customer id. Maintain database.
- Step 4: IIR Searching process:
1. Get the query for search.
  2. Apply binary search or hashing function.
- Step 5: 3. Based on top k retrieval of search (more number of times search).  
 3.1 According to aggregate function purpose sort the columns.

Step 6: Repeat the process or stop.

After completing the web usage mining, they are indexed with an indexer as shown in below fig. 9. There is a provision for business manager to interact with system at user interface area called user query. Business manager can have query like total sales in all branches, most searched product, most bill paid customer, sales in every region and so on. This process will help very well in the business. Index can be based on hash functions. And data in the index is in sorted order as per dates and it will be easy to retrieve the information very fast. This type of index can be maintained for the new customers or items or products or suppliers or dealers also.

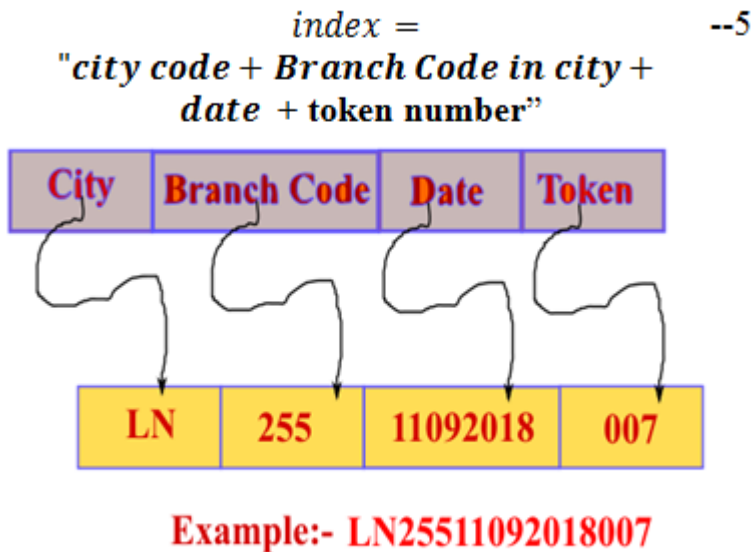


Fig. 9 Index to store customer details.

The search process starts with getting the search details as per data sets for finding highest bill paid by the customer in a year.

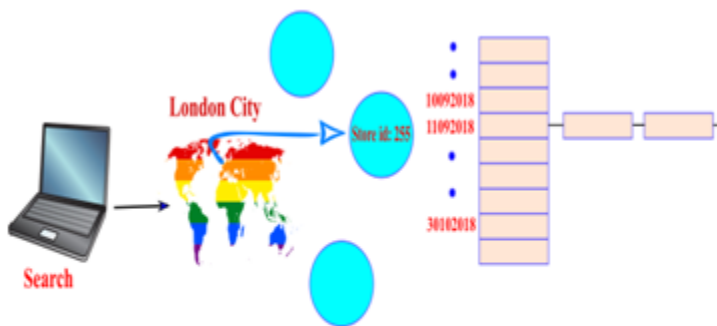


Fig. 10 Search Process of IIR.

The customer id or index is given as query by the business manager to the system as shown in fig. 10. Software gets the id or index which contains city, branch code, date and token number as identity (unique key). Business organization can maintain distributed database and search process starts from city and branch code, based on that it will enter into the date and token number which are in sorted order and by applying binary search it will move to the customer linked list with bills in sorted order. The process will search for the exact bill with maximum payment. Automatically based on customer identity with location will chose it is distributed database. The information of the data is stored in sequential order of customer activities. They cannot be in disorder manner. The internal of customer identity, everything will be in sorted order.

$$Bill_{amount} = S(R(sort(Max(a)))) \quad \text{--5}$$

S is for search of data sets of the relation R, by applying binary search or hashing functions, it is easy to track the customer location by index in the very fast manner. All the bills made by the customer will be mined to one column in a ascending or descending order. Based on max or min, search can find maximum or minimum bill in the data set.

## 5 Implementation

Java code for Super market datasets:

```
import java.io.BufferedReader;
import java.io.FileReader;
import weka.core.Instances;
import weka.associations.Apriori;
Instances data = new Instances (new BufferedReader(new FileReader("datasets/chap5/supermarket.arff")));
Apriori model = new Apriori();
Model.buildAssociations(data);
System.out.println(model);//
```

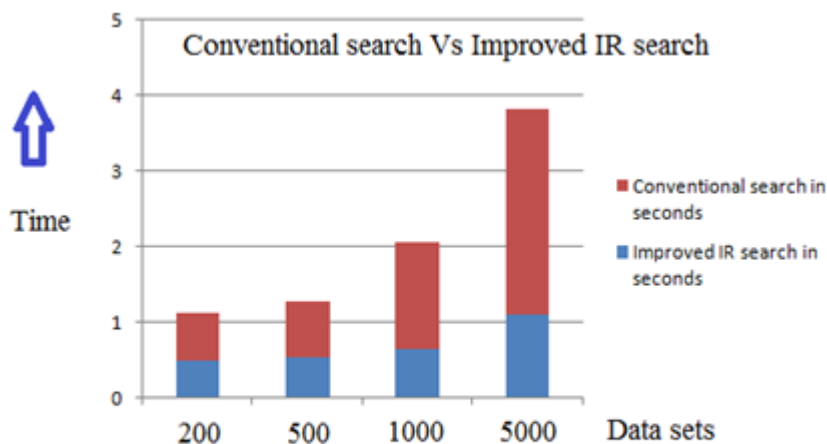
.....

Testing process is carried between the conventional and improved information retrieval searches. Test results are given between the normal search and the proposed search results. Improved IR technique is fast than the conventional search technique as in table 1.

<b>Data set rows</b>	<b>Conventional search in seconds</b>	<b>Improved IR search in seconds</b>
200	0.63	0.49
500	0.74	0.53
1000	1.40	0.65
5000	2.73	1.10

**Table. 1 comparison with conventional searches.**

Graph related to the given to the above readings is given Fig.11.



**Fig. 11 Conventional search versus improved IIR search**

Difference between conventional and improved IR search is given in fig. 11 as a bar graph with data sets on x-axis and time on the y-axis.

## 6 Conclusion


People use computer for various purposes like video, audio, images, multimedia and so on. All this data is stored in repositories. Similarly a business organization will store information in large databases or repositories. This information is growing data by data and there is necessity of retrieval of that data from large repositories and shows it in useful way. Improved Information Retrieval (IIR) is an important concept that is related to big data. Retrieving information from large repositories is a challenging issue. Based on WUM, store data to repositories with the help of index and for retrieving purpose go to the exact location and retrieve information from that location. Index means already it will be sorted order only. It is very easy to find the location and some data is stored as data sets in the repositories and is retrieved with the help of index under distributed databases. This paper consumes storage space for the index because daily data to store is growing day by day.




## 7 References

- [1]. Addanki Ramya, Konda Sreenu, Prathipati Ratna Kumar, "Preprocessing and Unsupervised For Web Usage Mining", International Journal of Social Networking and Virtual Communities, ISSN: 2252-8784, Volume - 1, No. 2, December 2012.
- [2]. K R Sunitha, R Krishnamoorthi, "Classification of Web log Data To Identify Interested Users Using Decision Trees", Research Gate, December, 2010.
- [3]. Mr. Sanjay Bapu Thakare, Prof. Sangram Z Gawali, "A Effective and Complete Preprocessing for Web Usage Mining", IJCSE, Vol. 02, No. 03, 2010, 848-851.
- [4]. Ms. Dipa Dixit, Ms. M Kiruthika, "Preprocessing of web logs", IJCSE, Vol. 02, No. 07, 2010, 2447-2452.
- [5]. Renáta Iváncsy, István Vajk, "Frequent Pattern Mining in Web Log Data", Acta Polytechnica Hungarica, Vol. 03, No. 01, 2006.
- [6]. Bikash Mukhopadhyay, Prof. Sripati Mukhopadhyay, "Data Mining Techniques for Information Retrieval", 2<sup>nd</sup> International CALIBER – 2004, New Delhi, 11-13, February 2004.
- [7]. Mr. S K Chaitanya, Rudraraju, Chalasani Srinivas, Prathipati Ratna Kumar, "Improved Novel Based Ant Colony Clusters for Fast Execution of Large Datasets", IJPAM, Volume. 120, No. 06, 2018, 4675-4692.
- [8]. Shadab Irfan, Subajit Ghosh, "Web Mining for Information Retrieval", IESC, Volume 8, Issue No. 4, April 2018, PP: 17277-17283.
- [9]. Jan O Pederson, Per-Kristian Halvorsen Douglas R Cutting, Jon W Tukey, Eric A. Bier, Daniel G Bobrow, "Iterative technique for phrase query formation and information retrieved system employing name", 1991.
- [10]. Niranjan Lal, Samimul Qamar, Savita Shiwani, "Information retrieval System and Challenges with Dataspace", International journal of Computer Applications, 147(8): 23-28, August 2016.

## 8 Acknowledgements

The authors thanking to the management of NSRIT college for their help in providing resources for the publication of this paper.

S No	Author Details
	Dr. Tarigoppula V S Sriram, working as Associate Professor in department of Computer Science and Engineering, Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam.(NSRIT). He is having total 17 years of teaching experience.

	His area of interest is Big Data Analytics, DWDM and Software Engineering. Guided UG and PG students and published 12 papers in international journals.
	Dr. Madhavi Kolukuluri, working as Associate Professor in department of Computer Science and Engineering, Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam. She is having total 14 years of teaching experience. Her areas of interests are Data Mining, Software Engineering and Computer Networks. She Guided UG and PG students and having 4 papers published in international journals.
	Dr. V V Hari Babu, working as Sr. Assistant Professor, Department of Physics, Bapatla Engineering College, 20 years of teaching experience.
	Mr. P Ratna Kumar, working as Assistant Professor in department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Hyderabad off-campus, since 2 years and having total 25 years of teaching experience.