Journal of Current Development in Artificial Intelligence

Volume No. 12

Issue No. 1

January - April 2024

ENRICHED PUBLICATIONS PVT. LTD

S-9, IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR-5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-47026006

Journal of Current Development in Artificial Intelligence

Aims and Scope

Journal of Current Development in Artificial Intelligence is a Journal addresses concerns in applied research and applications of artificial intelligence (AI). the journal also acts as a medium for exchanging ideas and thoughts about impacts of AI research. Articles highlight advances in uses of AI systems for solving tasks in management, industry, engineering, administration, and education evaluations of existing AI systems and tools, emphasizing comparative studies and user experiences and the economic, social, and cultural impacts of AI. Papers on key applications, highlighting methods, time schedules, person months needed, and other relevant material are welcome.

Journal of Current Development in Artificial Intelligence

Journal of Current Development in Artificial Intelligence

(Volume No. 12, Issue No. 1, January - April 2024)

Contents

A Study of Library Automation and Networking in Engineering College Libraries of Western Uttar Pradesh Region

K.K. Singh[1], Sunil Tyagi[2]
[1]Library & Information Officer, Indian Pharmacopoeia Commission,
Ministry of Health & Family Welfare, Government of India, Sector-23, Rajnagar, Ghaziabad (UP),
E-Mail: kksinghipc@gmail.com
[2]Library & Information Assistant, Indian Pharmacopoeia Commission,
Ministry of Health & Family Welfare, Government of India, Sector-23, Rajnagar, Ghaziabad (UP).

ABSTRACT

The main aim of the work is to present the status of Library Automation and Networking in Engineering College Libraries of Western Uttar Pradesh Region; to study the various barriers of Library Automation and Networking of Engineering College Libraries of Western Uttar Pradesh Region and to provide the feasible and cost effective solution to remove these barriers and allow libraries to cater the standard and quality services to the user community. The study is based on descriptive method of research includes surveys and fact-findings enquiries of different kinds. The major purpose of descriptive research is description of the sate-of-affairs as it exists at present. As the research is descriptive in nature, the present study use the pretested, structured questionnaire and administered observation and interview was also used to generate data to make the study reliable. In order to ascertain the extent of the automation of the libraries the librarians were asked to indicate the extent of automation of the libraries. It is observed from the data that out of 8 libraries, 4 (50%) libraries are completely automated and 3 (37.50%) libraries are partially automated. The Professional qualification, including the technical qualification and training on IT related disciplines are the important aspects, because the librarians are expected to organize technology based information services on one hand, and to impart information seeking skills on the other. It shows that there are 4 (50%) librarians are having Master degree in Library and Information Science, 2 (25%) have M. Phil degree and 1 (12.50%) librarian has Bachelor degree in Library and Information Science, and 1 (12.50%) librarian has Ph. D doctoral degree.

Keywords - Library automation, networking, IT, ICT, modules, operating systems.

INTRODUCTION

The library and information centre is an important component of any educational institution, which is hub of the teaching, and learning activities where students, researcher and teachers can explore the vast resources of information. In the traditional libraries users have to send more time for searching a small piece of information and for that have to depend mainly on the library professional or library staff. But in the age of information communication technology, computers are being used for day-to-day house-keeping activity of the library which saves the time of the end users, and library professional also and at the same time avoid duplication of work and make the library service smooth and effective.

Library automation assumed a great deal of importance in libraries in the mid-1960s. Since then it has become a household word in librarianship. Library automation may be defined as the application of automatic and semi-automatic data processing machines to perform library functions such as acquisition, circulation, cataloguing, reference service, and serials control. Automation of library activities provides the services very efficiently, rapidly, effectively, adequately and economically. As a result of the recent developments, the public has entered cyberspace and expects its information provider, the library, to provide the launching pad. Accordingly, today's integrated system not only must provide modules automating the traditional library functions but also must be capable of connecting through the local systems into systems of other suppliers, databases-bibliography and full content, online and compact disk read only memory (CD-ROM) databases, and the internet. Library users now expect their library systems to be able to do among other things: provide seamless integration between system gateways, remote and local databases through the public catalogue module; allow access by remote users to library's resources via the internet connection; and provide access to resources available on internet using a variety of graphical and multimedia-based software interfaces. The library automation which started in late 70s in few special libraries has now reached at most of the libraries of the India. Research is not satisfactory at engineering college libraries due to various problems. It is essential to identify the barriers, analyze the convenient steps for automation of the libraries and thereby network the libraries for benefit of the society.

STATEMENT OF THE PROBLEM

The information supply in the engineering college libraries is mostly constrained by the major factors as under:

- Poor funding to these libraries;
- Poor bibliographic control of up-to-date literature;
- Inadequate and poor collections;
- Non availability of professional staff; and
- Lack of management support.

The consequences of the above problems hinder engineering college libraries to render effective information services to their users. The lack of services restricts access to information. Though, their role in providing access to information is major, there are no minimum standards ascertained for their libraries. The provision of access to required information is the base for overall development of engineering college libraries. Taking into consideration these factors, the present research work is undertaken.

REVIEW OF LITERATURE

Aswal, (2006) in his study entitled 'Library Automation for 21 Century' highlights Library automation in the 21 century is on interconnecting systems, sharing information resources through innovative networking and ensuring equitable access to a broad range of information and users. These developments include the growth of the use of networks and the internet. This book is a guide for library professionals about the planning process for a library automation system of all sizes. The library needs are more sophisticated, if your current library automation is not fit for latest technology environment of 21 century and your library system does not work properly, you should look for new library systems. Seize this opportunity to assess your system can help you meet your services goals.

Kani, Ghinea and Chen (2008) conducted a study on "User perceptions of Online Public Library Catalogue. The purpose of this study was to establish user suggestions for a typical OPAC applications functionality and features. They also noted that OPACs are widely used electronic library catalogue giving a wealth of remote access to library information resources. Users' role is very important in the OPAC development process to ensure a usable and functional interface, as the integration of user defined requirements of OPACs, along with the other human computer interaction considerations. This facility offers a better understanding of user perceptions and expectations in respect of OPACs, which ultimately result in a truly user centered OPACs. An experiment was undertaken to find out the type of interaction features that users prefer to have in OPAC. The experiment study revealed that regardless of users Information Technology (IT) backgrounds, users expect OPACs to facilitate easier ways to achieve their tasks.

A study on help users search, prototyping an online help system for OPACs by Greifeneder (2008) revealed the difference between the beginners and the experienced and how their mental models fit a particular system. Only if librarians know their users and their behaviour, they can anticipate their problems. This study found that 25 percent of the users already failed at the operating system level, and another 37% got zero results. The first part of this study describes what online help systems are spelling corrections, faceted- browsing, recommender- systems, context sensitive help avatars are only a few examples of the broad field. The second part of the study explains what stands behind the concept of use ware – engineering and how we can apply it to prototype a company independent of OAPC help system.

Modern libraries have become more and more aware of the revolutionary impact of developments in Information and Communication Technology (ICT) on their major activities. The application of ICT facilitates to provide pinpointed, expeditious and exhaustive information at the right time to the right user. It provides opportunities for libraries and information centers to widen the scope of their resources

and services and to increase their significance within the organization they serve. The increasing availability of information in machine readable form allows much information needs to be satisfied with decreased involvement of libraries and librarians. This book has two parts one part is 'Computerization of University Libraries' and second part is 'Application of Information Communication Technology in Libraries'. This book will also be extremely helpful to the students, the researchers and the faculty in library and information science who would like to carry out research studies on library computerization and allied subjects.

Anuradha, (2000) in study entitled "Automated Circulating system using Visual Basic 6.0 discusses salient features of automated circulation system, designed and developed to suit the requirements of a medium sized library using programming language visual basic. It also gives advantages of visual basic based circulation system and objectives of circulation control system and different types of files.

OBJECTIVES OF THE STUDY

The main aim of the work is to study the various barriers of Library Automation and Networking of Engineering College Libraries of Western Uttar Pradesh Region and to provide the feasible and cost effective solution to remove these barriers and allow libraries to cater the standard and quality services to the user community.

More specifically the objectives of the study are enlisted as given below:

- To study the present status of Library Automation and Networking in Engineering College Libraries of Western Uttar Pradesh Region.
- To study the Computer Skilled Professional Staff for Library Automation.
- To evaluate the Library Software/Hardware adopted by the Engineering College Libraries of Western Uttar Pradesh Region.
- To study the Housekeeping Operations (Acquisition module, Circulation, Cataloguing, Serial Control, OPAC/ WEB-OPAC (Intranet/ Internet).
- To find the various barriers in the process the automation and networking faced by engineering college libraries such as inadequate staff, insufficient budget, etc.

METHODOLOGY

The study is based on descriptive method of research includes surveys and fact-findings enquiries of different kinds. The major purpose of descriptive research is description of the sate-of-affairs as it exists at present. As the research is descriptive in nature, the present study use the pretested, structured questionnaire and administered observation and interview was also used to generate data to make the

the study reliable. The Questionnaire was circulated to the 8 engineering college libraries of Western Uttar Pradesh Region. Out of 8 engineering college libraries, all libraries responded to the questionnaire and the response rate is 100%. Hence the analysis of the data collected is based on the responses of these 8 engineering college libraries. This data is collected for the period of 2014-15. The engineering college libraries covered in study are:

- Sir Chhotu Ram Institute of Engineering & Technology (SCRIET), Meerut.
- Meerut Institute of Engineering and Technology (MIET), Meerut.
- Bharat Institute of Technology (BIT), Meerut.
- Radha Govind Engineering College (RGGI), Meerut.
- Kishan Institute of Engineering and Technology (KIIT), Meerut.
- Vidya College of Engineering (VCE), Meerut.
- Galgotias University, Greater Noida.
- Vishveshwarya Group of Institutions, Noida.

Type of Management

S/N	Type of Management	Number of Libraries	Percentage
1	Government	1	12.50%
2	Private Aided	7	87.50%
TOTAL		8	100%

ANALYSIS, INTERPRETATION AND PRESENTATION OF DATA

The analysis of the data collected is based on the responses of 8 engineering college libraries of Western Uttar Pradesh Region. This data is collected for the period of 2014-15.

QUALIFICATION OF LIBRARIAN

The Professional qualification, including the technical qualification and training on IT related disciplines are the important aspects, because the librarians are expected to organize technology based information services on one hand, and to impart information seeking skills on the other.

Table-1: The manpower distribution of libraries based on qualification

S/N	Qualification	Respondents	Percentage
1	Ph.D	1	12.50%
2	M. Lib.& Inf. Sc; M. Phil	2	25.00%
3	M. Lib.& Inf. Sc	4	50.00%
4	B. Lib & Inf. Sc	1	12.50%
TOTAL		8	100%

The analysis of data as shown in the table-1 shows the distribution of libraries based on qualification. It shows that there are 4 (50%) librarians are having Master degree in Library and Information Science, 2 (25%) have M. Phil degree and 1 (12.50%) librarian has Bachelor degree in Library and Information Science, and 1 (12.50%) librarian has Ph. D doctoral degree.

The status of automation in different engineering college libraries

In order to ascertain the extent of the automation of the libraries the librarians were asked to indicate the extent of automation of the libraries. It is observed from the data as shown in the table-2 that out of 8 libraries, 4 (50%) libraries are completely automated and 3 (37.50%) libraries are partially automated.

Table-2: The status of automation in different engineering college libraries

S/N	Type	Respondents	Percentage
1	Completely Automated	4	50.00%
2	Partially Automated	3	37.50%
3	Initial Stages	1	12.50%
4	No Response	0	0%
TOTAL		8	100%

The different types of Library Software's used

Librarians were asked to provide the details about the use of software in their libraries. It is observed from the data as shown in the table-3 that 3 library uses LIBSYS software and that 2 libraries uses SOUL, and SIM software.

Table-3: Types of Library Software's used

S/N	Name of the College Library	Library Software
1	Sir Chhotu Ram Institute of Engineering & Technology (SCRIET)	SOUL
2	Meerut Institute of Engineering and Technology (MIET)	LIBSYS
3	Bharat Institute of Technology (BIT)	LIBWARE
4	Radha Govind Engineering College (RGGI)	LIBSYS
5	Kishan Institute of Engineering and Technology (KIIT)	SOUL
6	Vidya College of Engineering (VCE)	LIBSYS
7	Galgotias University, Greater Noida	SIM
8	Vishveshwarya Group of Institutions	SIM

The number of college libraries using various operating systems

The networking of the different computed of the library will help to give multi user access to the information. The question regarding availability of the supporting software for networking in the libraries was asked. The table-4 shows the number of college libraries using different operating systems. 3 (37.50%) library are using windows 7 operating systems.

Table-4: The number of college libraries using various operating systems

S/N	Operating Systems	Respondents	Percentage
1	Windows 2000	1	12.50%
2	Windows NT	1	12.50%
3	Windows XP	1	12.50%
4	Windows Vista	1	12.50%
5	Windows 7	3	37.50%
6	Windows 8	1	12.50%
TOTAL		8	100%

The areas of automation modules by the libraries

The analysis of data as shown in the table-5 reveals that out of 8 libraries, 100% using acquisition, cataloguing and OPAC modules whereas 87.50% using serial control module.

Table-5: The areas of automation modules by the libraries

S/N	Automation Modules	Respondents	Percentage
1	Acquisition	8	100%
2	Cataloguing	8	100%
3	Serial Control	7	87.50%
4	Circulation	6	75.00%
5	OPAC	8	100%
6	Administration	3	37.50%
7	Web OPAC	3	37.50%

IT Specialized staff for library automation and networking

The table-6 shows IT specialized staff for Library Automation and Networking. It is observed that 7 (87.50%) libraries have sufficient IT specialized staff for Automation and Networking.

Table-6: IT Specialized staff for library automation and networking

S/N	Category	Respondents	Percentage
1	Yes	7	87.50%
2	No	1	12.50%
TOTAL		8	100%

The barriers faced by the library staff during automation

The barriers faced by the library staff during automation, it is found from the table-7 that 5 (62.50%) librarians say lack of IT knowledge on the part of users and 3 (37.50%) say inadequate trained staff.

Table-7: The barriers faced by the library staff during automation

S/N	Automation Modules	Respondents	Percentage
1	Insufficient funds	2	25.00%
2	Inadequate trained staff	3	37.50%
3	lack of co-ordination among the library staff	1	12.50%
4	lack of IT knowledge on the part of users	5	62.50%
5	Lack of support from management	2	25.00%

CONCLUSION

Library automation is the process which needs proper planning, timely implementation and periodical evaluation. The librarian with the administrators has to set the priorities after analyzing the current status and future requirements. Selection of the suitable integrated library management package according to the needs of the users and the library is important. Retrospective conversion, OPAC, circulation and serials control, etc. should be conducted with care. Staff training and user education are keys to the success of the process. Library automation invites realistic approach. Here, those institutions which freed their visions from the traditional shackles of financial insecurities and fears of making proper decisions can only set the pace of journey to excellence. Engineering education and libraries are in a state of transformation in electronic age. Outcome measurements for engineering education including libraries are becoming the norm. Now a day's students need to acquire information skills and critical thinking skills as part of their engineering education. So they can become productive participants in the work and be prepared for lifelong learning. Faculties and librarians can achieve better learning outcomes in terms of critical thinking and lifelong learning skills. They work together on designing curricula to include appropriate courses and modules to teach information skills. The collaboration between faculty and librarians will ensure that faculty members are prepared for electronic information

use, as a result they will be able to integrate technology into their teaching processes and students are taught useable information skills.

REFERENCES

Anuradha, P., (2000). Automated Circulation System using Visual Basic 6.0. Annals of Library science and documentation, 47

Aswal, R.S., (2006). Library Automation for 21 Century. New Delhi: Ess Ess Publication.

Bavakutty, M. Salih M.T.K & Haneefa, Mohamed. (2006). Research on Library Computerization. New Delhi: Ess Ess Publication.

Greifeneder, Elke (2008). Help Users Search! Prototyping an Online Help System for OPACs. In BOBCATSSS-Conference, 129-136.

Kani-Zabihi, Elahe. Ghinea, Gheorghita & Chen, Sherry Y (2008). User perceptions of Online Public Access Catalogues. International Journal of Information Management, 28 (6), 492-502.

Singh, S P (1987). Automation in libraries. New Delhi: Metropolitan.

Sinha, Manoj Kumar. (2008). Scenario of Automation and Networking of Library and Information Centers (LICs) of North Eastern Region of India. An Evaluative study. Ahmedabad and NEHU, Shillong: INFLIBNET Centre,

Uddin, Hanif, (2009). Library Automaton: A study of the AIC, INSDOC and National Libraries of Bangaladesh. Paris: United Nation Educational Scientific and Cultural Organization.

Extract Transform Load Data With Etl Tools Like 'INFORMATICA'

[1]Preeti, [2]Neetu Sharma

[1] M. Tech., Computer Science Engineering, Ganga Institute of Technology and Management, Bahadurgarh-Jhajjar Road, Kablana, Distt. Jhajjar, Haryana

[2] HOD C.S.E Deptt., Ganga Institute of Technology and Management, Bahadurgarh-Jhajjar Road, Kablana, Distt. Jhajjar, Haryana

A B S T R A C T

: As we all know business intelligence (BI) is considered to have an extraordinary impact on businesses. Research activity has grown in the last years [10]. A significant part of BI systems is a well performing Implementation of the Extract, Transform, and Load (ETL) process. In typical BI projects, implementing the ETL process can be the task with the greatest effort. Here, set of generic Meta model constructs with a palette of regularly used ETL activities, is specialized, which are called templates.

1. INTRODUCTION

We all want to load our data warehouse regularly so that it can assist its purpose of facilitating business analysis [1]. To do this, data from one or more operational systems desires to be extracted and copied into the warehouse. The process of extracting data from source systems and carrying it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading. It is an essential phenomenon in a data warehouse. Whenever DML (data manipulation language) operations such as INSERT, UPDATE OR DELETE are issued on the source database, data extraction occurs. After data extraction and transformation have taken place, data are loaded into the data warehouse.

Extraction: The first part of an ETL process is to extract the data from the home systems. Most data warehousing projects amalgamate data from unlike source systems. Each separate system may also use a different data association format. Common data source formats are relational databases and flat files, but may contain non-relational database structures such as IMS or other data structures .Extraction converts the data into a format for transformation processing. The quantity of data is reduced by

omitting any non-relevant data sets. Extraction must not negatively affect the performance of productive systems It runs as a background task or is executed at times of low activity (e.g. during the night).

Transformation: Any transformation desirable to provide data that can be interpreted in business terms is done in the second step. Data sets are cleaned with regard to their dataquality. Eventually, they are converted to the scheme of the target database and consolidated. The transform stage applies a series of rules or functions to the extracted data to derive the data to be loaded. Some data sources will require very slight manipulation of data. Data transformations are often the most difficult and, in terms of processing time, the most costly part of the ETL process. They can range from simple data conversions to extremely complex data scrubbing techniques.

Loading: Now, the real loading of data into the data warehouse has to be done. The early Load which generally is not time-critical is great from the Incremental Load. Whereas the first phase affected productive systems, loading can have an giant effect on the data warehouse. This especially has to be taken into consideration with regard to the complex task of updating currently stored data sets. In general, incremental loading is a critical task. ETL processes can either be run in batch mode or real time. Batch jobs typically are run periodically. If intervals become as short as hours or even minutes only, these processes are called near real time. The load phase loads the data into the data warehouse. Depending on the requirements of the organization, this process ranges widely. Some data warehouses merely overwrite old information with new data.

More complex systems can maintain a history and audit trail of all changes to the data. Designing and maintaining the ETL process is often considered one of the most difficult and resource-intensive portions of a data warehouse project. Many data warehousing projects use ETL tools to manage this process. Data warehouse builders create their own ETL tools and processes, either inside or outside the database.

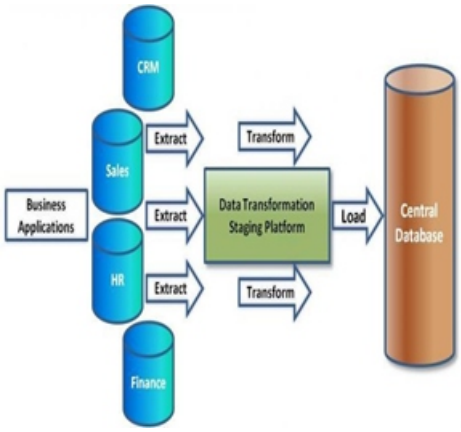


Fig. The ETL Process

ETL tools have been around for some time, have evolved, matured, and now present us with productive environments for Big Data, Data Warehouse, Business Intelligence, and analytics processing. However these are few problems with them .Such tools are very expensive and does not support small size businesses. Configuration of such tools takes lot of time. These are many ETL tools availble in the market .Some of the ETL Tools are:

- DataStage from Ascential Software
- SAS System from SAS Institute
- Informatica
- Data Integrator From BO
- Oracle Express
- Abinito
- Decision Stream From Cognos
- MS-DTS from Microsoft
- Pentaho Kettle

Informatica is the best ETL tool in the marketplace [3]. It can extract data from numerous heterogeneous sources, transforming them as per business needs and loading to target tables. It's used in Data migration and loading projects. It is a visual interface and you will be dragging and dropping with the mouse in the Designer(client Application).This graphical approach to communicate with all major databases and can move/transform data between them. It can move huge bulk of data in a very effective way. Informatica is a tool, supporting all the steps of Extraction, Transformation and Load process. Now a days Informatica is also being used as an Integration tool.

Informatica is an easy to use tool. It has got a simple visual interface like forms in visual basic. You just need to drag and drop different objects (known as transformations) and design process flow for Data extraction transformation and load. These process flow diagrams are known as mappings. Once a mapping is made, it can be scheduled to run as and when required. In the background Informatica server takes care of fetching data from source, transforming it, & loading it to the target systems/databases.

Informatica can talk with all major data sources (mainframe / RDBMS / Flat Files / XML / VSM / SAP etc), [3] can move/transform data between them. It can move hugevolumes of data in a very operational way, many a times better than even bespoke programs written for specific data movement only. It can throttle the transactions (do big updates in small chunks to avoid long locking and filling the transactional log). It can effectively join data from two distinct data sources (even a xml file can be joined with a relational table). In all, Informatica has got the ability to effectively integrate

heterogeneous data sources & converting raw data into useful information.

Some facts and figures about Informatica Corporation:

- Founded in 1993, based in Redwood City, California
- 1400+ Employees; 3450 + Customers; 79 of the Fortune 100 Companies
- NASDAQ Stock Symbol: INFA; Stock Price: $18.74 (09/04/2009)
- Revenues in fiscal year 2008: $455.7M
- Headquarters: Redwood City, CA
- Offices: N. & S. America, Europe, Asia Pacific
- Government organizations in 20 countries
- Partners: Over 400 major SI, ISV,OEM and On Demand

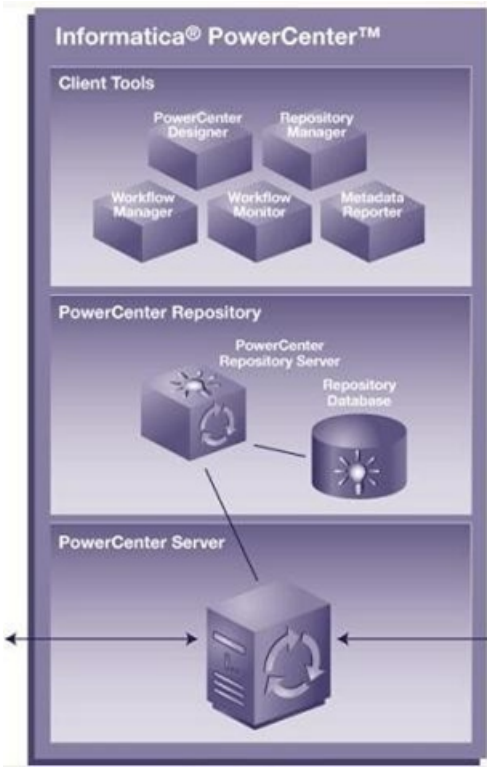
2. COMPONENTS OF INFORMATICA

Informatica provides the following integrated components:

Informatica repository. The Informatica repository is at the center of the Informatica suite. You create a set of metadata tables within the repository database that the Informatica applications and tools access. The Informatica Client and Server access the repository to save and retrieve metadata. The PowerCenter repository resides on a relational database. The repository database tables contain the instructions required to extract, transform, and load data. PowerCenter Client applications access the repository database tables through the Repository Server. You add metadata to the repository tables when you perform tasks in the PowerCenter Client application, such as creating users, analyzing sources, developing mappings or mapplets, or creating workflows. The PowerCenter Server reads metadata created in the Client application when you run a workflow. The PowerCenter Server also creates metadata, such as start and finish times of a session or session status[2].

You can develop global and local repositories to share metadata:

- **Global repository.** The global repository is the hub of the domain. Use the global repository to store common objects that multiple developers can use through shortcuts. These objects may include operational or Application source definitions, reusable transformations, mapplets, and mappings.
- **Local repositories.** A local repository is within a domain that is not the global repository. Use local repositories for development. From a local repository, you can create shortcuts to objects in shared folders in the global repository. These objects typically include source definitions, common dimensions and lookups, and enterprise standard transformations. You can also create copies of objects in non-shared folders.



- Version control. A versioned repository can store multiple copies, or versions, of an object. Each version is a separate object with unique properties. PowerCenter version control features allow you to efficiently develop, test, and deploy metadata into production.

You can connect to a repository, back up, delete, or restore repositories using pmrep, a command line program.

You can view much of the metadata in the Repository Manager. The Informatica Metadata Exchange (MX) provides a set of relational views that allow easy SQL access to the Informatica metadata repository.

Repository Server

The Repository Server manages repository connection requests from client applications. For each repository database registered with the Repository Server, it configures and manages a Repository Agent process. The Repository Server also monitors the status of running.

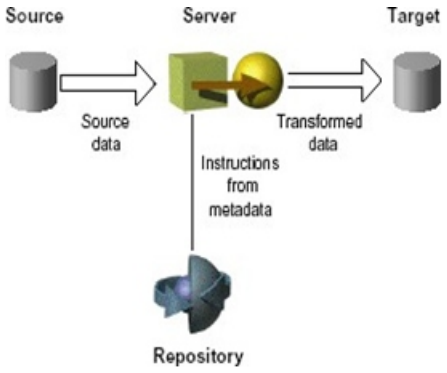
Repository Agents, and sends repository object notification messages to client applications. Informatica Client. Use the Informatica Client to manage users, define sources and targets, build mappings and mapplets with the transformation logic, and create sessions to run the mapping logic. The Informatica Client has three client applications: Repository Manager, Designer, and Workflow

Manager.

- Repository Server Administration Console. Use the Repository Server Administration console to administer the Repository Servers and repositories.
- Repository Manager. Use the Repository Manager to administer the metadata repository. You can create repository users and groups, assign privileges and permissions, and manage folders and locks.
- Designer. Use the Designer to create mappings that contain transformation instructions for the PowerCenter Server. Before you can create mappings, you must add source and target definitions to the repository. The Designer has five tools that you use to analyze sources, design target schemas, and build source-to-target mappings:

- **Workflow Manager.** Use the Workflow Manager to create, schedule, and run workflows. A workflow is a set of instructions that describes how and when to run tasks related to extracting, transforming, and loading data. The PowerCenter Server runs workflow tasks according to the links connecting the tasks. You can run a task by placing it in a workflow.

- **Workflow Monitor.** Use the Workflow Monitor to monitor scheduled and running workflows for each PowerCenter Server. You can choose a Gantt chart or Task view. You can also access details about those workflow runs.



InformaticaServer: The Informatica Server extracts the source data, performs the data transformation, and loads the transformed data into the targets.

The PowerCenter Server reads mapping and session information from the repository. It extracts data from the mapping sources and stores the data in memory while it applies the transformation rules that you configure in the mapping. The PowerCenter Server loads the transformed data into the mapping targets.

The PowerCenter Server can achieve high performance using symmetric multi-processing systems. The PowerCenter Server can start and run multiple workflows concurrently. It can also concurrently process partitions within a single session. When you create multiplepartitions within a session, the PowerCenter Server creates multiple database connections to a single source and extracts a separate range of data for each connection, according to the properties you configure.

3. INFORMATICA PRODUCT LINE

Informatica is a powerful ETL tool from Informatica Corporation, a leading provider of enterprise data integration software and ETL softwares.

The important products provided by Informatica Corporation is provided below:

- Power Center
- Power Mart
- Power Exchange
- Power Center Connect
- Power Channel
- Metadata Exchange
- Power Analyzer
- Super Glue

Power Center & Power Mart: Power Mart is a departmental version of Informatica for building, deploying, and managing data warehouses and data marts. Power center is used for corporate enterprise data warehouse and power mart is used for departmental data warehouses like data marts. Power Center supports global repositories and networked repositories and it can be connected to several sources. Power Mart supports single repository and it can be connected to fewer sources when compared to Power Center. Power Mart can extensibily grow to an enterprise implementation and it is easy for developer productivity through a codeless environment.

Power Exchange: Informatica Power Exchange as a stand-alone service or along with Power Center, helps organizations leverage data by avoiding manual coding of data extraction programs. Power Exchange supports batch, real time and changed data capture options in main frame(DB2, VSAM, IMS etc.,), mid-range (AS400 DB2 etc.,), and for relational databases (oracle, sql server, db2 etc) and flat files in unix, linux and windows systems.

Power Center Connect: This is adding on to Informatica Power Center. It helps to extract data and metadata from ERP systems like IBM's MQSeries, Peoplesoft, SAP, Siebel etc. and other third party applications.

Power Channel: This helps to transfer large amount of encrypted and compressed data over LAN, WAN, through Firewalls, tranfer files over FTP, etc.

Meta Data Exchange: Metadata Exchange enables organizations to take advantage of the time and effort already invested in defining data structures within their IT environment when used with Power Center. For example, an organization may be using data modeling tools, such as Erwin, Embarcadero, Oracle designer, Sybase Power Designer etc for developing data models. Functional and technical team should have spent much time and effort in creating the data model's data structures(tables, columns, data types, procedures, functions, triggers etc). By using meta deta exchange, these data structures can be imported into power center to identifiy source and target mappings which leverages time and effort. There is no need for informatica developer to create these data structures once again.

Power Analyzer: Power Analyzer provides organizations with reporting facilities. PowerAnalyzer makes accessing, analyzing, and sharing enterprise data simple and easily available to decision makers. PowerAnalyzer enables to gain insight into business processes and develop business intelligence. With PowerAnalyzer, an organization can extract, filter, format, and analyze corporate information from data stored in a data warehouse, data mart, operational data store, or other data storage models. PowerAnalyzer is best with a dimensional data warehouse in a relational database. It can also run reports on data in any table in a relational database that do not conform to the dimensional model.

Super Glue: Superglue is used for loading metadata in a centralized place from several sources. Reports can be run against this superglue to analyze meta data.

4. TYPES OF INFORMATICA PARTITIONS

Informatica provides you the option of enhancing the performance of the Informatica session by the The PowerCenter® Partitioning Option. After tuning all the performance bottlenecks we can further improve the performance by addition partitions[3].We can either go for Dynamic partitioning (number of partition passed as parameter) or Non- dynamic partition (number of partition are fixed while coding). Apart from used for optimizing the session, Informatica partition become useful in situations where we need to load huge volume of data or when we are using Informatica source which already has partitions defined, and using those partitions will allow to improve the session performance.

The partition attributes include setting the partition point, the number of partitions, and the partition types.

Partition Point: There can be one or more pipelines inside a mapping. Adding a partition point will divide this pipeline into many pipeline stages. Informatica will create one partition by default for every pipeline stage. As we increase the partition points it increases the number of threads. Informatica has mainly three types of threads –Reader, Writer and Transformation Thread. The number of partitions can be set at any partition point. We can define up to 64 partitions at any partition point in a pipeline. When you increase the number of partitions, you increase the number of processing threads, which can improve session performance. However, if you create a large number of partitions or partition points in a session that processes large amounts of data, you can overload the system.

You cannot create partition points for the following transformations:

- Source definition
- Sequence Generator
- XML Parser
- XML target
- Unconnected transformations

The partition type controls how the Integration Service distributes data among partitions at partition points. The Integration Service creates a default partition type at each partition point.

Types of partitions are:

1. Database partitioning
2. Hash auto-keys
3. Hash user keys
4. Key range
5. Pass-through
6. Round-robin

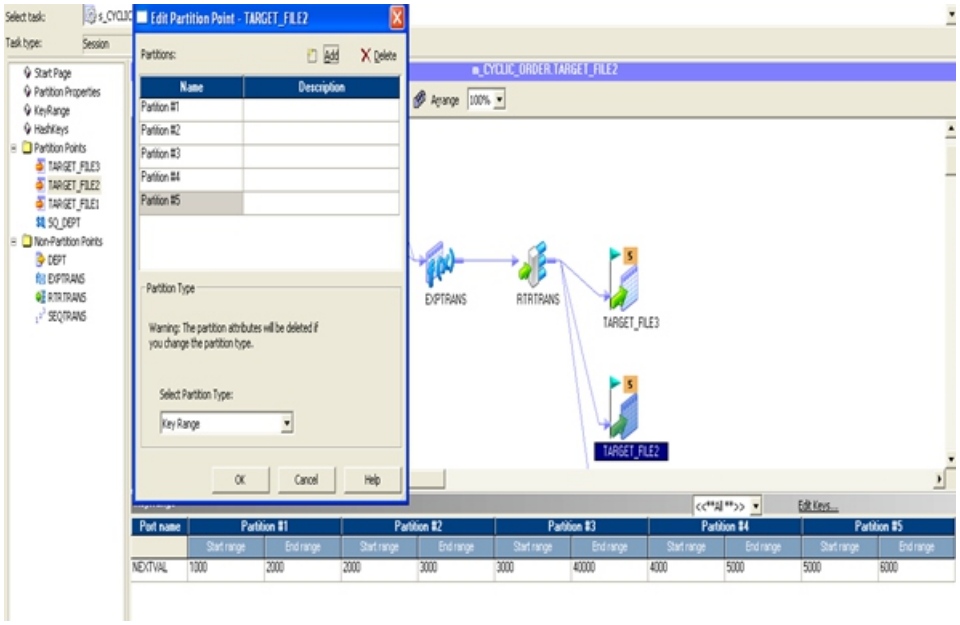
Database Partitioning: For Source Database Partitioning, Informatica will check the database system for the partition information if any and fetches data from corresponding node in the database into the session partitions. When you use Target database partitioning, the Integration Service loads data into corresponding database partition nodes.

Use database partitioning for Oracle and IBM DB2 sources and IBM DB2 targets.

Pass through: Using Pass through partition will not affect the distribution of data across partitions instead it will run in single pipeline which is by default for all your sessions. The Integration Service processes data without redistributing rows among partitions. Hence all rows in a single partition stay in

the partition after crossing a pass-through partition point.

Key range: Used when we want to partition the data based on upper and lower limit. The Integration Service will distribute the rows of data based on a port or set of ports that we define as the partition key. For each port, we define a range of values. Based on the range that we define the rows are send to different partitions.



Round robin partition is used to when we want to distributes rows of data evenly to all partitions Hash auto-keys: The Integration Service uses a hash function to group rows of data among partitions. The Integration Service groups the data based on a partition key.

Hash user keys: The Integration Service uses a hash function to group rows of data among partitions. We define the number of ports to generate the partition key.

5. TRANSFORMATIONS

Informatica Transformations: A transformation is a repository object that generates, modifies, or passes data. The Designer provides a set of transformations that performspecific functions. A transformation is a repository object that generates, modifies, or passes dataThe Designer provides a set of transformations that perform specific functions.Data passes into and out of transformations through ports that you connect in a mapping.

Transformations can be of two types:

Active Transformation: An active transformation can change the number of rows that pass through the transformation, change the transaction boundary, can change the row type. For example, Filter, Transaction Control and Update Strategy are active transformations. The key point is to note that Designer does not allow you to connect multiple active transformations or an active and a passive transformation to the same downstream transformation or transformation input group because the Integration Service may not be able to concatenate the rows passed by active transformations However, Sequence Generator transformation (SGT) is an exception to this rule[4]. A SGT does not receive data. It generates unique numeric values. As a result, the Integration Service does not encounter problems concatenating rows passed by a SGT and an active transformation.

Aggregator	performs aggregate calculations
Filter	serves as a conditional filter
Router	serves as a conditional filter (more than one filters)
Joiner	allows for heterogeneous joins
Source qualifier	represents all data queried from the source

Passive Transformation: A passive transformation does not change the number of rows that pass through it, maintains the transaction boundary, and maintains the row type. The key point is to note that Designer allows you to connect multiple transformations to the same downstream transformation or transformation input group only if all transformations in the upstream branches are passive.

Expression	performs simple calculations
Lookup	looks up values and passes to other objects
Sequence generator	generates unique ID values
Stored procedure	calls a stored procedure and captures return values
Update strategy	allows for logic to insert, update, delete, or reject data

6.1 TYPES OF TRANSFORMATION

6.1. Expression Transformation:

You can use the Expression transformation to calculate values in a single row before you write to the target. For example, you might need to adjust employee salaries, concatenate first and last names, or convert strings to numbers. You can use the Expression transformation to perform any non-aggregate calculations. You can also use the Expression transformation to test conditional statements before you output the results to target tables or other transformations.

Calculating Values: To use the Expression transformation to calculate values for a single row, you must include the following ports:

- Input or input/output ports for each value used in the calculation. For example, when calculating the total price for an order, determined by multiplying the unit price by the quantity ordered, the input or input/output ports. One port provides the unit price and the other provides the quantity ordered.
- Output port for the expression. You enter the expression as a configuration option for the output port. The return value for the output port needs to match the return value of the expression. For information on entering expressions, see "Transformations" in the Designer Guide. Expressions use the transformation language, which includes SQL-like functions, to perform calculations

You can enter multiple expressions in a single Expression transformation. As long as you enter only one expression for each output port, you can create any number of output ports in the transformation. In this way, you can use one Expression transformation rather than creating separate transformations for each calculation that requires the same set of data.

2. Joiner Transformation:

You can use the Joiner transformation to join source data from two related heterogeneous sources residing in different locations or file systems. Or, you can join data from the same source. The Joiner transformation joins two sources with at least one matching port. The Joiner transformation uses a condition that matches one or more pairs of ports between the two sources. If you need to join more than two sources, you can add more Joiner transformations to the mapping.The Joiner transformation requires input from two separate pipelines or two branches from one pipeline.

The Joiner transformation accepts input from most transformations. However, there are some limitations on the pipelines you connect to the Joiner transformation. You cannot use a Joiner transformation in the following situations:

- Either input pipeline contains an Update Strategy transformation.
- You connect a Sequence Generator transformation directly before the Joiner transformation

The join condition contains ports from both input sources that must match for the PowerCenter Server to join two rows. Depending on the type of join selected, the Joiner transformation either adds the row to the result set or discards the row. The Joiner produces result sets based on the join type, condition, and input data sources. Before you define a join condition, verify that the master and detail sources are set for optimal performance. During a session, the PowerCenter Server compares each row of the master source against the detail source. The fewer unique rows in the master, the fewer iterations of the join

comparison occur, which speeds the join process. To improve performance, designate the source with the smallest count of distinct values as the master. You can improve session performance by configuring the Joiner transformation to use sorted input. When you configure the Joiner transformation to use sorted data, the PowerCenter Server improves performance by minimizing disk input and output. You see the greatest performance improvement when you work with large data sets. When you use a Joiner transformation in a mapping, you must configure the mapping according to the number of pipelines and sources you intend to use. You can configure a mapping to join the following types of data:

- **Data from multiple sources.** When you want to join more than two pipelines, you must configure the mapping using multiple Joiner transformations.
- **Data from the same source.** When you want to join data from the same source, you must configure the mapping to use the same source

Perform joins in a database when possible.

Performing a join in a database is faster than performing a join in the session. In some cases, this is not possible, such as joining tables from two different databases or flat file systems. If you want to perform a join in a database, you can use the following options:

- Create a pre-session stored procedure to join the tables in a database.
- Use the Source Qualifier transformation to perform the join.

Join sorted data when possible.

You can improve session performance by configuring the Joiner transformation to use sorted input. When you configure the Joiner transformation to use sorted data, the PowerCenter Server improves performance by minimizing disk input and output. You see the greatest performance improvement when you work with large data sets.

For an unsorted Joiner transformation, designate as the master source the source with fewer rows. For optimal performance and disk storage, designate the master source as the source with the fewer rows. During a session, the Joiner transformation compares each row of the master source against the detail source.

3. Rank Transformation:

The Rank transformation allows you to select only the top or bottom rank of data. You can use a Rank transformation to return the largest or smallest numeric value in a port or group. You can also use a Rank transformation to return the strings at the top or the bottom of a session sort order. During the session, the PowerCenter Server caches input data until it can perform the rank calculations. You connect all ports representing the same row set to the transformation. Only the rows that fall within that rank, based on

some measure you set when you configure the transformation, pass through the Rank transformation.
You can also write expressions to transform data or perform calculations. As an active transformation, the Rank transformation might change the number of rows passed through it. You might pass 100 rows to the Rank transformation, but select to rank only the top 10 rows, which pass from the Rank transformation to another transformation.
Rank Caches

During a session, the PowerCenter Server compares an input row with rows in the data cache. If the input row out-ranks a cached row, the PowerCenter Server replaces the cached row with the input row. If you configure the Rank transformation to rank across multiple groups, the PowerCenter Server ranks incrementally for each group it finds.
Rank Transformation Properties:

- Enter a cache directory.
- Select the top or bottom rank.Select the input/output port that contains values used to determine the rank. You can select only one port to define a rank.
- Select the number of rows falling within a rank.
- Define groups for ranks, such as the 10 least expensive products for each manufacturer.

The Rank transformation changes the number of rows in two different ways. By filtering all but the rows falling within a top or bottom rank, you reduce the number of rows that pass through the transformation. By defining groups, you create one set of ranked rows for each group.

4. Router Transformation:

A Router transformation is similar to a Filter transformation because both transformations allow you to use a condition to test data. A Filter transformation tests data for one condition and drops the rows of data that do not meet the condition. However, a Router transformation tests data for one or more conditions and gives you the option to route rows of data that do not meet any of the conditions to a default output group. If you need to test the same input data based on multiple conditions, use a Router transformation in a mapping instead of creating multiple Filter transformations to perform the same task. The Router transformation is more efficient. For example, to test data based on three conditions, you only need one Router transformation instead of three filter transformations to perform this task. Likewise, when you use a Router transformation in a mapping, the PowerCenter Server processes the incoming data only once

5. Lookup Transformation:

Use a Lookup transformation in a mapping to look up data in a flat file or a relational table, view, or synonym. You can import a lookup definition from any flat file or relational database to which both the PowerCenter Client and Server can connect[6]. You can use multiple Lookup transformations in a mapping. It compares Lookup transformation port values to lookup source column values based on the lookup condition. Pass the result of the lookup to other transformations and a target.

You can use the Lookup transformation to perform many tasks, including:

- Get a related value. For example, your source includes employee ID, but you want to include the employee name in your target table to make your summary data easier to read.
- Perform a calculation. Many normalized tables include values used in a calculation, such as gross sales per invoice or sales tax, but not the calculated value (such as net sales).
- Update slowly changing dimension tables. You can use a Lookup transformation to determine whether rows already exist in the target.
- You can configure the Lookup transformation to perform the following types of lookups:
- Connected or unconnected. Connected and unconnected transformations receive input and send output in different ways.
- Relational or flat file lookup. When you create a Lookup transformation, you can choose to perform a lookup on a flat file or a relational table.
- Cached or uncached. Sometimes you can improve session performance by caching the lookup table. If you cache the lookup, you can choose to use a dynamic or static cache. By default, the lookup cache remains static and does not change during the session. With a dynamic cache, the PowerCenter Server inserts or updates rows in the cache during the session. When you cache the target table as the lookup, you can look up values in the target and insert them if they do not exist, or update them if they do.

Note: If you use a flat file lookup, you must use a static cache.

Using Sorted Input

When you configure a flat file Lookup transformation for sorted input, the condition columns must be grouped. If the condition columns are not grouped, the PowerCenter Server cannot cache the lookup and fails the session. For best caching performance, sort the condition columns. The Lookup transformation also enables an associated ports property that you configure when you use a dynamic cache.

6. ADVANTAGES OF INFORMATICA

A comprehensive integration platform that promotes code standardization, unifies collaboration

between business and IT roles, and provides capabilities that handle the highvolume and wide variety of today's business data. The Informatica Platform has eight distinct technologies designed to be a true industrial strength ETL solution. These include:

- Messaging
- Complex Event Processing
- B2B Data Exchange
- Cloud Data Integration
- Enterprise Data Integration
- Application Lifecycle Management
- Data Quality
- Master Data Management

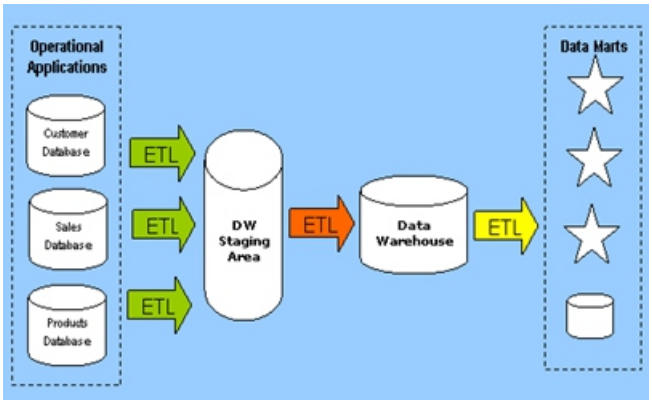
- **Improve network speed.** Slow network connections can slow session performance. Have your system administrator determine if your network runs at an optimal speed. Decrease the number of network hops between the PowerCenter Server and databases.
- **Use multiple PowerCenter Servers.** Using multiple PowerCenter Servers on separate systems might double or triple session performance.
- **Use a server grid.** Use a collection of PowerCenter Servers to distribute and process the workload of a workflow. For information on server grids.
- **Improve CPU performance.** Run the PowerCenter Server and related machines on high performance CPUs, or configure your system to use additional CPUs.
- **Configure the PowerCenter Server for ASCII data movement mode.** When all character data processed by the PowerCenter Server is 7-bit ASCII or EBCDIC, configure the PowerCenter Server for ASCII data movement mode[5].
- **Check hard disks on related machines.** Slow disk access on source and target databases, source and target file systems, as well as the PowerCenter Server and repository machines can slow session performance. Have your system administrator evaluate the hard disks on your machines.

- **Reduce paging**. When an operating system runs out of physical memory, it starts paging to disk to free physical memory. Configure the physical memory for the PowerCenter Server machine to minimize paging to disk.
- **Use processor binding.** In a multi-processor UNIX environment, the PowerCenter Server may use a large amount of system resources. Use processor binding to control processor usage by the PowerCenter Server

7. DATA STAGING

The data staging area is the data warehouse workbench. It is the place where raw data is brought in, cleaned, combined, archived, and eventually exported to one or more data marts. The purpose of the data staging area is to get data ready for loading into a presentation server (a relational DBMS or an OLAP engine). A staging area, or landing zone, is an intermediate storage area used for data processing during the extract, transform and load (ETL) process. The data staging area sits between the data source(s) and the data target(s), which are often data warehouses, data marts, or other data repositories [1]

Data staging areas are often transient in nature, with their contents being erased prior to running an ETL process or immediately following successful completion of an ETL process. There are staging area architectures, however, which are designed to hold data for extended periods of time for archival or troubleshooting purposes. Staging areas can be implemented in the form of tables in relational databases, text-based flat files (or XML files) stored in file systems or proprietary formatted binary files stored in file systems.[2] Staging area architectures range in complexity from a set of simple relational tables in a target database to self-contained database instances or file systems.[3] Though the source systems and target systems supported by ETL processes are often relational databases, the staging areas that sit between data sources and targets need not also be relational databases.[4]



The Data Warehouse Staging Area is temporary location where data from source systems is copied. A staging area is mainly required in a Data Warehousing Architecture for timing reasons. In short, all required data must be available before data can be integrated into the Data Warehouse.[1]
The staging area in Business Intelligence is a key concept. The role of this area is to have a secure place to store the source systems data for further transformations and cleanings. Why do we do that?
Because:

- It minimizes the impact on the source systems (you don't want to re-extract everything from the source systems if your ETL failed).
- It can be used for auditing purposes (we store the data that we process).
- It eases the development process (you don't need to be bound to the operational servers).

The data staging area of the data warehouse is both a storage area and a set of processes commonly referred to as extract-transformation-load (ETL).[2] The data staging area is everything between the operational source systems and the data presentation area. It is somewhat analogous to the kitchen of a restaurant, where raw food products are transformed into a fine meal. In the data warehouse, raw operational data is transformed into a warehouse deliverable fit for user query and consumption. Similar to the restaurant's kitchen, the backroom data staging area is accessible only to skilled professionals. The data warehouse kitchen staff is busy preparing meals and simultaneously cannot be responding to customer inquiries. Customers aren't invited to eat in the kitchen. [3]

It certainly isn't safe for customers to wander into the kitchen. We wouldn't want our data warehouse customers to be injured by the dangerous equipment, hot surfaces, and sharp knifes they may encounter in the kitchen, so we prohibit them from accessing the staging area. Besides, things happen in the kitchen that customers just shouldn't be privy to. The key architectural requirement for the data staging area is that it is off-limits to business users and does not provide query and presentation services.

Once the data is extracted to the staging area, there are numerous potential transformations, such as cleansing the data (correcting misspellings, resolving domain conflicts, dealing with missing elements, or parsing into standard formats), combining data from multiple sources, deduplicating data, and assigning warehouse keys. These transformations are all precursors to loading the data into the data warehouse presentation area.[1] The data staging area is dominated by the simple activities of sorting and sequential processing. In many cases, the data staging area is not based on relational technology but instead may consist of a system of flat files. After you validate your data for conformance with the defined one-to-one and many-to-one business rules, it may be pointless to take the final step of building a fullblown third-normal-form physical database. However, there are cases where the data arrives at the

doorstep of the data staging area in a third-normal-form relational format. In these situations, the managers of the data staging area simply may be more comfortable performingthe cleansing and transformation tasks using a set of normalized structures.

8. DATA STAGING PROCESS

A staging area, or landing zone, is an intermediate storage area used for data processing during the extract, transform and load (ETL) process. The data staging area sits between the data source(s) and the data target(s), which are often data warehouses, data marts, or other data repositories.[1]The data staging process imports data either as streams or files, transforms it, produces integrated, cleaned data and stages it for loading into data warehouses, data marts, or Operational Data Stores.[2]

First, Kimball distinguishes two data staging scenarios.

In (1) a data staging tool is available, and the data is already in a database. The data flow is set up so that it comes out of the source system, moves through the transformation engine, and into a staging database. The flow is illustrated in Figure One.

Figure One -- First Data Staging Scenario

In the second scenario, begin with a mainframe legacy system. Then extract the sought after data into a flat file, move the file to a staging server, transform its contents, and load transformed data into the staging database.

Figure Two illustrates this scenario.

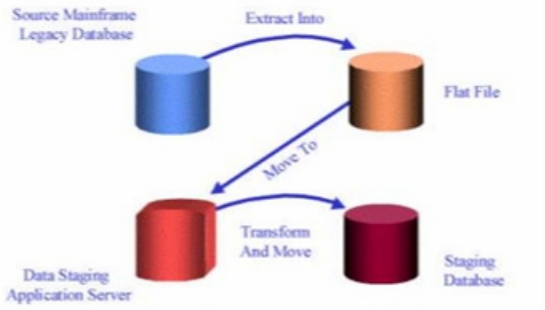


Figure Two -- Second Data Staging Scenario

We assume that the data staging area is not a query service. In other words, any database that is used for querying is assumed to be physically downstream from the data staging area. If the legacy data is already available in a relational database, then it may make sense to perform all the processing steps within the relational framework, especially if the source relational database and the eventual target presentation database are from the same vendor.[3] This makes even more sense when the source database and the target database are on the same physical machine, or when there is a convenient high-speed link between them. However, there are many variations on this theme, and in many cases it may not make sense to load the source data into a relational database. In the detailed descriptions of the processing steps, we will see that almost all the processing consists of sorting, followed by a single, sequential pass through either one or two tables.[1] This simple processing paradigm does not need the power of a relational DBMS. In fact, in some cases, it may be a

serious mistake to divert resources into loading the data into a relational database when what is needed is sequential flat-file processing. Similarly, we will see that if the raw data is not in a normalized entity-relationship (ER) format, in many cases it does not pay to load it into an ER physical model simply to check data relationships. The most important data integrity steps involving the enforcement of one-to-one and one-to-many relationships can be performed, once again, with simple sorting and sequential processing. It is acceptable to create a normalized database to support the staging processes; however, this is not the end goal. The normalized structures must be off-limits to user queries because they defeat understandability and performance. As soon as a database supports query and presentation services, it must be considered part of the data warehouse presentation area. By default, normalized databases are excluded from the presentation area, which should be strictly dimensionally structured. Regardless of whether we're working with a series of flat files or a normalized data structure in the staging area, the final step of the ETL process is the loading of data. Loading in the data warehouse environment usually takes the form of presenting the quality-assured dimensional tables to the bulk loading facilities of each data mart.[2] The target data mart must then index the newly arrived data for query performance. When each data mart has been freshly loaded, indexed, supplied with appropriate aggregates, and further quality assured, the user community is notified that the new data has been published.

The data staging area of the data warehouse is both a storage area and a set of processes commonly referred to as extract-transformation-load (ETL). The data staging area is everything between the operational source systems and the data presentation area. The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading.[3] It is a fundamental phenomenon in a data warehouse . Whenever DML (data manipulation language) operations such as INSERT, UPDATE OR DELETE are issued on the

the source database, data extraction occurs. After data extraction and transformation have taken place, data are loaded into the data warehouse.

Extraction: The first part of an ETL process is to extract the data from the source systems. Most data warehousing projects consolidate data from different source systems. Each separate system may also use a different data organization format. Common data source formats are relational databases and flat files, but may include non-relational database structures such as IMS or other data structures[4] .Extraction converts the data into a format for transformation processing. The amount of data is reduced by omitting any non- relevant data sets. Extraction must not negatively affect the performance of productive systems. Extracting means reading and understanding the source data and copying the data needed for the data warehouse into the staging area for further manipulation

Transformation: Any transformation needed to provide data that can be interpreted in business terms is done in the second step. Data sets are cleaned with regard to their data quality. Eventually, they are converted to the scheme of the target database and consolidated. The transform stage applies a series of rules or functions to the extracted data to derive the data to be loaded. Some data sources will require very little manipulation of data. Data transformations are often the most complex and, in terms of processing time, the most costly part of the ETL process. They can range from simple data conversions to extremely complex data scrubbing techniques.[2]

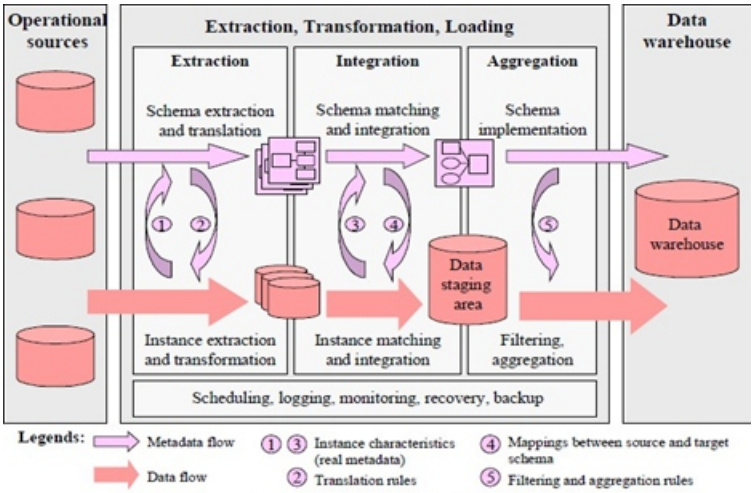


Figure 1. Steps of building a data warehouse: the ETL process

Loading: Finally, the actual loading of data into the data warehouse has to be done. The Initial Load which generally is not time-critical is distinguished from the Incremental Load. Whereas the first phase affected productive systems, loading can have an immense effect on the data warehouse. This especially has to be taken into consideration with regard to the complex task of updating currently stored data sets. In general, incremental loading is a critical task. ETL processes can either be run in batch mode or real

time. Batch jobs typically are run periodically. If intervals become as short as hours or even minutes only, these processes are called near real time. The load phase loads the data into the data warehouse. Depending on the requirements of the organization, this process ranges widely. Some data warehouses merely overwrite old information with new data.[3]

Unfortunately, there is still considerable industry consternation about whether the data that supports or results from this process should be instantiated in physical normalized structures prior to loading into the presentation area for querying and reporting. These normalized structures sometimes are referred to in the industry as the enterprise data warehouse; however, we believe that this terminology is a misnomer because the warehouse is actually much more encompassing than this set of normalized tables. The enterprise's data warehouse more accurately refers to the conglomeration of an organization's data warehouse staging and presentation areas. [4]

A normalized database for data staging storage is acceptable. However, we continue to have some reservations about this approach. The creation of both normalized structures for staging and dimensional structures for presentation means that the data is extracted, transformed, and loaded twice—once into the normalized database and then again when we load the dimensional model. Obviously, this two-step process requires more time and resources for the development effort, more time for the periodic loading or updating of data, and more capacity to store the multiple copies of the data.[1] At the bottom line, this typically translates into the need for larger development, ongoing support, and hardware platform budgets. Unfortunately, some data warehouse project teams have failed miserably because they focused all their energy and resources on constructing the normalized structures rather than allocating time to development of a presentation area that supports improved business decision making. While we believe that enterprise-wide data consistency is a fundamental goal of the data warehouse environment, there are equally effective and less costly approaches than physically creating a normalized set of tables in your staging area, if these structures don’t already exist.

9. CONCLUSION

The Informatica solution for enterprise data warehousing is proven to help IT departments implement data marts and departmental data warehouses and readily scale them up to enterprise data warehousing environments. This solution serves as the foundation for all data warehousing and enterprise data warehousing projects. It accelerates their deployment, minimizing costs and risks, by ensuring that enterprise data warehouses are populated and maintained with trustworthy, actionable, and authoritative data.

Data is and has been from the beginning created, stored, and retrieved by disparate, incompatible systems. Between 30% and 35% of all the data in the industry is still on mainframes, in languages and data structures that are archaic and generally unavailable.

The wave of specialty applications—HR, sales, accounting, ERP, manufacturing—have all contributed their share to the chaos. Informatica PowerCenter is the ETL tool that empowers an IT organization to implement highly scalable, high-performance data processing maps using a graphical interface that generates proprietary intermediate code[9]. This mapping code, when coupled with a defined data workflow plan can then be scheduled for a variety of execution strategies. Widely accepted as an industry front runner, Informatica boasts high productivity and low cost. In truth, this may be just the opposite. Perhaps high productivity at a cost is more accurate; in terms of experienced developers, administrators, and license fees. Companies who choose to use Informatica usually have very large IT teams and budgets.

REFERENCES

[1] http://dwhlaureate.blogspot.in/2012/07/informaticaetl-extract-transform-and.html

[2] http://www.dbbest.com/blog/extract-transform-load-etl-technologies-part-2/

[3] http://www.informatica.com/us/#fbid=rHNS5hx2rKv

[4] http://www.ventanaresearch.com/blog/commentblog.aspx?id=3524

[5] http://www.slideshare.net

[6] http://www.rpi.edu/datawarehouse/docs/ETL-Tool-Selection.pdf

[7] http://en.wikipedia.org/wiki/Informatica

[8] http://www.etltool.com/etl-vendors/powercenter-informatica/

[9] http://www.techopedia.com/definition/25983/informatica-powercenter

[10] www.ijsrp.org/research-paper-0214/ijsrp-p2688.pdf

Sentiment Analysis Of User's Views Using Machine Learning

[1]Rupinder, [2]Kaur, Ashok
[1] M.tech Research Scholar, Ambala College of Engineering and Applied Research, Devsthali, Ambala
[2] Assistant Prof., Department of Computer Science & Engineering, Ambala College of Engineering and Applied Research, Devsthali, Ambala

ABSTRACT

Sentiment Analysis or Opinion Mining is an important concept in today's world and due to the increased use of media it has become a huge source of database. Since everybody in the modern era is involved with some social media platform, the public mood is hugely reflected in the social media platform today. This study proposes to utilize this source of information and predict the all sentiments of public towards the food price in India expressed over twitter and twitter API is used for extracting live tweets. Oauth is used as handler and tweets are filtered for specific keywords and location using latitude and longitude data. The tweets are saved into a database. They first preprocessed for elimination of stop word, special characters, short words etc, after that stemming and tokenization steps are applied and TF-IDF score is calculated for all the keywords. A term document matrix (TDM) is created which is fed into the classifiers for classification. KNN and Naïve Baye's has been analyzed in this study and Hybrid algorithm using them was designed. The results of KNN and Naïve Baye's classifier in sentiment classification were found to be significant while the hybrid-KNN outperforms the Naïve Baye's Classifiers in terms of accuracy

Keywords -Opinion Mining, Sentiment Analysis, KNN, Naïve Baye's Classifier, Food price.

I. INTRODUCTION

Human life is filled with emotions and opinions. One cannot be imagined without emotions and opinions. They play a vital role in nearly all human actions and lead the human life by influencing the way they think, what they do and how they act. The recipients of the information do not only consume the available contents on web, but also can change this content and generate new data of information. In today's world of social media users can comment on already existing information, can book mark pages. They can also share their plans, news and knowledge with online communication. In this way, the entire community becomes a writer, in addition to being a reader. Internet users can port their data, give opinions and get feedback of other users through different medium like blogs, forums and social networks etc. The increasing popularity of different personal publishing services is increasing day by

day and in coming future this increase is expected to continue. Thus the opinionated information on web will become so large that it cannot be handled manually so need of automated Sentiment Classification of online opinions, reviews of information is desire of current and future scenario. Recently, most of researchers have focused on this area [1]. They fetch opinionated information to analyze and summarize the opinions expressed on web by different automatically with computers. Until now, all researchers have evolved several techniques to the solution of the problem. Current-day Information Retrieval (IR) and Natural Language Processing (NLP) is the crossroad for the Opinion Mining and Sentiment Analysis and share some characteristics with other disciplines such as text mining.

Use of Social Media

The popularity of social media can be accessed by the fact that almost 90% of internet users use it for some context or other. Some are active member of social networking; some use it for blogging and micro blogging. Other usage includes online video sharing, ecommerce etc [14]. Day by day, the number of internet users is also increasing. With increase of the user involvement on web their contribution to online data is also enhanced. One of contribution of this trend is providing online reviews in social networking sites. These reviews help users to take better decisions about the product for which reviews were placed. [12]. Hence to provide automation, we are studying sentiment analysis. Sentiment analysis is the modern method which helps to analyze huge amount of data to extract sentiments associated with the data.

Type of Social Media Applications

There are many social media platforms like Face book, Twitter, LinkedIn, You tube etc. In today's world they tend to be an integral part of almost every one's life. Figure 1.1 shows usage of different social media platform by the users of different age group.

Need for Analyzing Social Media Data

The use of social media is increasing day by day. Increasing growth of social media users over internet has also increased their participation in various discussions and activities simultaneously. In case of a product, users are tries to express the rating views about the product [12].These reviews are equally important for buyers and companies.

Table 1: User pattern of each Social Media

Social Network Used by U.S Internet Users, by Age, % of respondents in each group				
	14-17	18-34	35-54	TOTAL
Face book	63.80%	83.40%	74.20%	76.80%
YouTube	81.90%	77.20%	54.20%	66.40%
Twitter	31.00%	38.70%	28.30%	32.80%
LinkedIn	1.50%	15.90%	20.00%	16.60%
WhatsApp	8.00%	9.80%	4.00%	6.80%

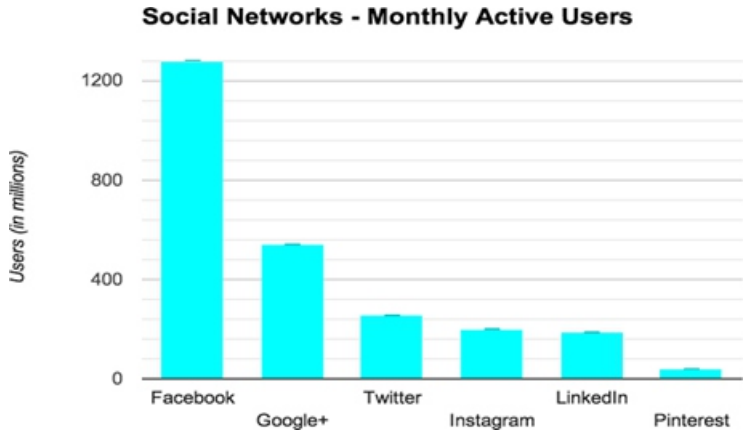


Figure 1: Depicting the Users of each type of social media platform

Sentiment Analysis

Sentiment analysis is a text classification problem which deals with extracting useful information present within the web documents. This extracted data can be then further classified according to its polarity as positive, negative or neutral. Sentiment analysis is widely used in business and government intelligence, Named Entity Recognition.

Sentiment analysis is also about finding subjectivity or objectivity of the opinion. Subjectivity is about someone's personal review whereas objectivity is the opinion given by an expert. For example: doctor's opinion about the patient on the basis of observed symptoms comes under the objectivity.

Definition of Sentiment Analysis

It can be defined as a computational task of extracting sentiments from the opinion. Some opinions represent sentiments and some opinions do not represent any sentiment. Sentiment analysis is a natural language processing and information extraction task. This aims to extract writer's feelings expressed in comments or reviews expressed over web.

Levels of Sentiment Analysis

Sentiment analysis is defined as a classification process. There are 3 main classification levels in sentiment analysis:

- **Document Level**

Identify if the document have (product reviews, forum posts) expressed opinions and whether opinions are positive negative or neutral.

- **Sentence Level**

The task at this level goes to the sentences and found whether each sentence expressed positive, negative or neutral opinions.

- **Attribute Level**

Extract object attribute (e.g. image quality, zoom size.) that are subject of opinion and opinion orientations (positive, negative or neutral).

Machine learning

Machine learning is technique by which a device modifies its own behavior due to the result of its past experience. This is systematic way which design algorithms and permit machine to evolve behaviors based on experimental data. Machine learning approaches can be divided into two categories:

- Supervised Learning
- Unsupervised Learning

II. LITERATURE SURVEY

Alexander Trilla [15] Used text classification scheme based on Multinomial Naïve Bayes to deals with Twitter messages. The effectiveness of this technique was evaluated using TASS- SEPLN twitter data sets and it achieved maximum macro averaged F1 measure rate of 36.28%. The accurate results provided by TASS-SEPLN organizers indicate that the proposal based on MNB was rather effective.

Aisopos and Fotis [4] have studied some serious challenges associated with respect to Micro blog content. Some of these are the applicability of sentiment analysis over past and different classification methods caused by their inherent characteristics of content. To resolve them, author introduced a method that relies on two orthogonal and complementary sources of evidence: context-based method captured by polarity ratio and content-based features acquired by n-gram graphs. Both the methods are language-neutral and tolerant to noise; guarantee high robustness and effectiveness in the manner author are considering. To ensure this approach can be applied into practical applications with large amount of data, aim should be enhancing its time efficiency. Thus author propose alternative sets of features having low extraction cost, explore dimensionality reduction techniques and discretization

techniques and also experiment with multiple different classification

Ortigosa et. al [1] proposed a novel method for sentiment analysis in social site giant Face book that, starting from the messages of its users, supports: (i) to extract useful information about the Face book users' sentiment polarity, which reflected from the user's messages; and (ii) to model the users' normal sentiment polarity and to analyze significant emotional changes in user. Author has implemented this method in Sent Buk which is a Face book application also presented in the paper. In general, in order to take decisions based on information and emotions of the users, it is necessary for a system to get and store this information. One of the most reliable procedures to fetch information about user's emotion consists of asking them directly to fill in questionnaires which helps to detect their option. However, for a particular user this task can be too time-consuming and tedious.

Agarwal and Apoorv [3] worked with micro blog data named as Twitter and manufacture models to classification of the "tweets" into positive and negative sentiment or they can be neutral. Author build novel models for two classification: first one is a binary task of classifying sentiment of users into positive and negative classes and secondly is a 3-way task of sentiment classification of users into positive, negative and neutral. Author experimented with two types of models: (1) unigram model which is a feature based model (2) a tree 30 kernel based model. They build a new tweet representation, for the tree kernel based model. They take a unigram model, which work well for sentiment analysis for Twitter data in the past. Result indicates that a unigram model is really a hard baseline. Feature based model that used 100 features gives similar accuracy as compared to the unigram model thatused about 10,000 features. Tree kernel based model gives improvement outperforming both these models by a significant margin.

Balahur and Alexandr [5], identified that the major difference between subjective texts type (like movie or product reviews) is that their target is unique and clearly stated across the text. Following various efforts of annotation and the analysis of the issues encountered, it was realized that news opinion mining is different from that of other text types. They identified 3 subtasks that need to be addressed: defining the target; separating the bad and good news content from the sentiment expressed which is good and bad; and finally analysis of clearly mentioned opinion that is expressed not ambiguously, not needing understanding or the utilization of world knowledge. Furthermore, they distinguish 3 not similar views on newspaper articles, which have to be handling differently while analyzing sentiment.

Horakova and Marketa [8] presented a model which collects tweets from social networking sites and thus provide a view of business intelligence. In the framework, there are two layers in the sentiment

analysis tool, the layer of data processing and the layer of sentiment analysis. Data processing layer deals with data collection and data mining, while sentiment analysis layer use a application to present the result of data mining.

III. PROPOSED METHODOLOGY

This section describes the various techniques were applied for the fulfillment of following objectives.

- To study existing algorithms for Sentiment Analysis.
- Analysis and Classification of User's views about products expressed over the web using machine learning techniques

The various text mining algorithm and streaming of twitter API are discussed in this section. The process starts with the extraction of tweets followed by preprocessing of the extracted tweets. Then Classifier algorithm has to be applied on it to identify the polarity.

Data extraction: The twitter API was used for tweet extraction. The major steps involved in development of the framework for live streaming of tweets begin with setting up an account on twitter.

- Set up your account on twitter
- Go to site of dev.twitter.com
- Create a new app and register for it
- Change access level to Read, write and access messages
- Generate security id and secret number
- Generate access token id and secret token number
- Save them to be utilized for streaming

Auth handler was used for streaming the tweets. Filters are applied on it using the track filter. The tweets were filtered by two ways.

- Filter by content
- Filter by location

Due to the policies of twitter the filtering is not absolutely correct and there might be a similar tweet which doesn't lie in the filtered bandwidth. The content filtering is done using the following keywords:

- Mehngai
- Food cost
- Inflation

- Food Security
- Prices of vegetables
- Price Rise

The location filtering was performed by using a 'location' filter available with tweepy. The location filter works on the basis of latitude and longitude of the place. A bounding box has to be formed in which the location filter works. Any tweets sent from that bounding box is streamed.
This has utilized the following settings:
South West Longitude=73 degrees
South West Latitude=15 degrees
North East longitude=85 degrees
North East Latitude=27 degrees
Using these settings the tweets are extracted and saved in a database.
For further processing Text mining was applied on the filtered tweets

Pre-processing: Tokenization, stop word removal and stemming are the main steps of preprocessing. Tokenization divides textual description into tokens by removing punctuation marks. Then stop words are performed that remove unnecessary information (conjunctions, interjections and articles) from datasets. Stemming on reduced datasets is performed to reduced terms into their root terms. Porter's stemming algorithm is used to perform stemming.

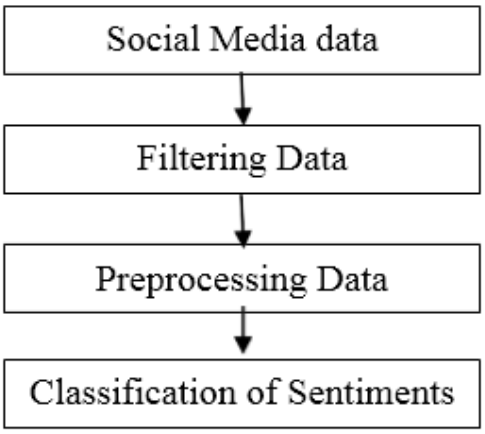


Figure 2: Sentiment Analysis process

Step 1: Preprocessing: To distill unstructured data to structured format this step is used. There are different preprocessing steps used in Text mining such as tokenization, stop word removal and stemming. These algorithms are discussed below.

I. Tokenization: There are two types of tokenization i.e. partial and full tokenization. For removing commas, full stop, hyphen and brackets we use this step. It divides the whole text into separate tokens to explore the words in document.

ii. Stop word removal: The purpose of this process is used to reduce conjunction, prepositions, articles and other frequent words such as adverbs, verbs and adjectives from data. Thus it reduces textual data and system performance is improved.

iii. Stemming: For reduction of words into their root word e.g. words like ”Processing”, ”processed” has it root word “process” stemming process is used. The purpose of stemming is to represent the words to only terms in their document. There are various tools to perform stemming such as Lovins Stemmer, Porters Stemmer, Paice/Husk Stemmer, Dawson Stemmer, HMM Stemmer.

iv. Weighting Factor: - Features are extracted from overloaded large datasets.TF-IDF (Term frequency- Inverse document frequency) [7] score is generally is used to giveweight to each term. TF-IDF is multiply of term frequency and inverse document frequency.

$$\text{TF-IDE} = n_w^d \log_2\left(\frac{N}{N_w}\right)$$

Where n_w^d = frequency of word w in document d.

N= total document and N_w = document containing word w.

v. Term-document matrix – After all steps of the preprocessing Term- document matrix is created from the text available in documents. Rows in matrix represents document in which word appears and columns represent the words that are extracted from documents. The TF-IDF score filled in the cell of matrix.

Machine Learning Approaches

There are different methods to design machine learning algorithms. The purpose of ML algorithms is to use observations as an input and this can be a data, information and past experience. To improve the performance of instances we used MI algorithms, which can be done by any classifier by trying to classify the input pattern into set of categories or to cluster unknown instances. There are two categories supervised and unsupervised approach [9].

KNN Algorithm

The K means algorithm takes the centroid of a cluster as mean value of the points within cluster. It randomly selects k, number of objects in dataset, each of which initially represents a cluster mean. For each of remaining objects, an object is assigned to cluster to which it is most similar, based on Euclidean distance. The algorithm iteratively improves within cluster variations. The iterations continue until assignment become stable.

KNN is lazy learning type of algorithm. In this learning the function is accurately local and all computation is deferred until classification. It is simplest of all machine learning algorithms. In KNN classification, the class membership is output. An object is classified by majority votes of its neighbors by the object being assigned to class which is most common among its k nearest neighbor (k is positive small integer). The nearest neighbor is calculated by using similarity measure usually distance functions are user. There is some distance function used by KNN [50].

Euclidean Distance Function $\sqrt{\sum_{i=1}^{N}(a_i - b_i)^2}$

Manhattan Distance Function $\sum_{i=}^{N}|a_i - b_i|$

Where $\{(a_1, b_1), (a_2,b_2), (a_3,b_3) \ldots\ldots (a_N, b_N)\}$ are training datasets.

In KNN algorithm all the distance from testing data point to training data point are computed. Then all testing points are sorted according to the ascending order. Then class labels are added for each K nearest neighbors and sign of sum are used for checking prediction. Finding value of K in K-nearest neighbor is more challenging task.

As choosing smaller value of k. e.g. by choosing K=1 may take lead to risk of over fitting and choosing larger value of K e.g. K=N may take lead to under fitting problem. Therefore optimal value of K has been taken between the values 3-10, which gives better result.

Strengths of KNN

1. It is relatively efficient and scalable in processing on large number of dataset.
2. It is often terminate at local optimum.

Weakness

1. Applicable only when mean is defined.
2. Unable to handle noisy data.
3. No guarantee to converge to global optimum

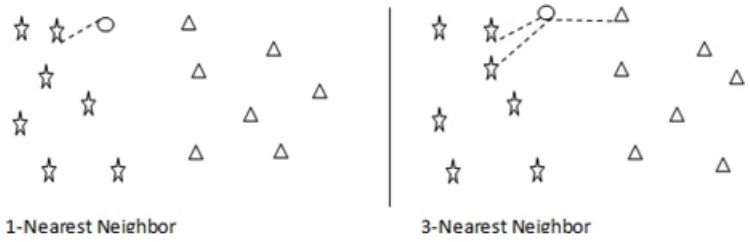


Figure 3: Working of KNN Algorithm

Naïve Bayes Algorithm

Bayesian classifier is statistical classifier. It can predict class membership probabilities such as probability that a given tuple belongs to a particular class. Bayesian classifier has a minimum error rate in comparisons to other classifiers.The algorithm is named after popular statistician Thomas Bayes who proposed Bayesian theorem. The Naïve bayes algorithm is also based on Bayesian theorem. This theorem supposes that all the attributes are conditionally independent to each other. This assumption is also called class conditional independence. In this algorithm, conditional probability for every attribute with respect to certain class level is calculated. The new document is classified using sum of probabilities for each class [12] . The classifier is easy to build and useful when there is large datasets.The classification framework is briefly discussed as follows:

Suppose we have D set of tuples and each tuple has attribute vector X(x1, x2, x3 , xn) of n dimensions. Let there are k number of classes C1, C2, C3... Ck. The classifier predicts X belongs to Ci if

$$P(\frac{C_i}{X}) = P\frac{C_j}{X}$$

for 1<= j <= k, j < > i

Posterior probability is calculated as

$$P\left(\frac{C_i}{X}\right) = \frac{P\left(\frac{X}{C_i}\right) P(c_i)}{P(X)}$$

Problems with Existing Approach:

Problems with KNN

- A shortcoming of the k-NN approach is that it is sensitive to the local structure of the data.
- KNN doesn't know which attributes are more important.
- Doesn't handle missing data gracefully.

Problems with Naïve Bayes

- Most important disadvantage of Naive Bayes is that it has strong feature of independence assumptions.
- In classification tasks you need a big dataset. You can use Naïve Bayes classification algorithm with a small data set but precision and recall will keep very slow.

Proposed KNN Hybrid algorithm

A KNN Hybrid algorithm is composition of KNN algorithm and Naïve Bayes algorithm. Here Naïve Bayes algorithm is embedded in KNN algorithm to take advantages of both algorithms. By this approach we will able to handle noisy database and missing values also. Hence the capabilities of KNN are enhanced by embedded it with Naïve Bayes.

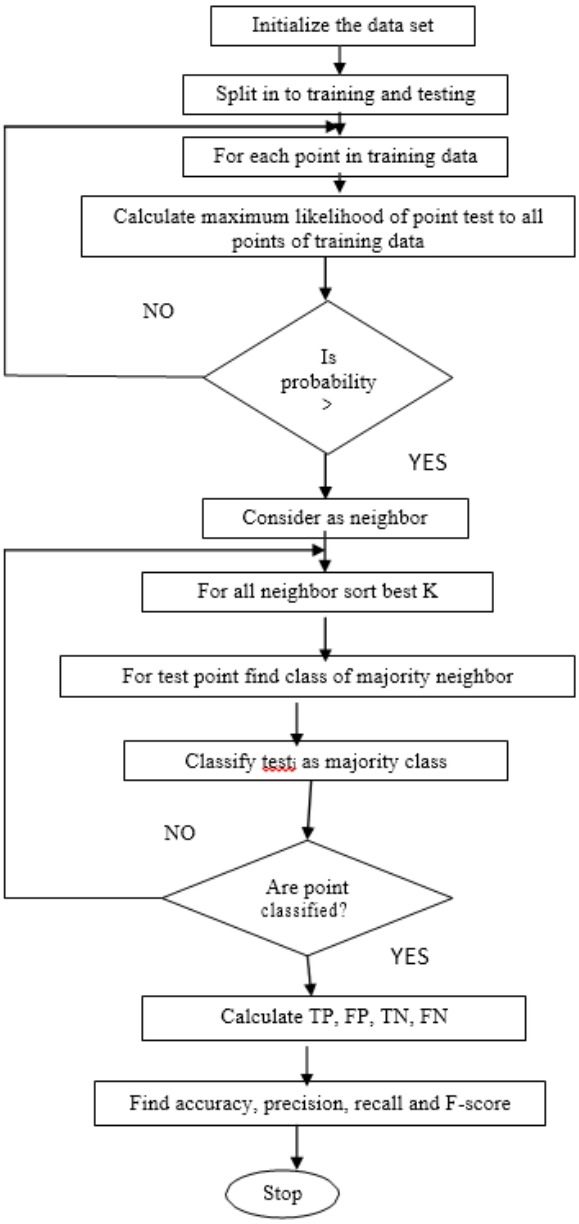


Figure 4: KNN Hybrid Algorithm

IV. RESULTS AND DISCUSSION

This Section presents the results obtained by various methodologies applied on the dataset discussed in the previous section. The results are verified by running the simulations for repeated number of times. The opinions are mined and analyzed for public response.

An app named 'rupinder kaur' was created. This was utilized for streaming.

A consumer Key is generated. Next figure shows the Keys generated which will be used for streaming.

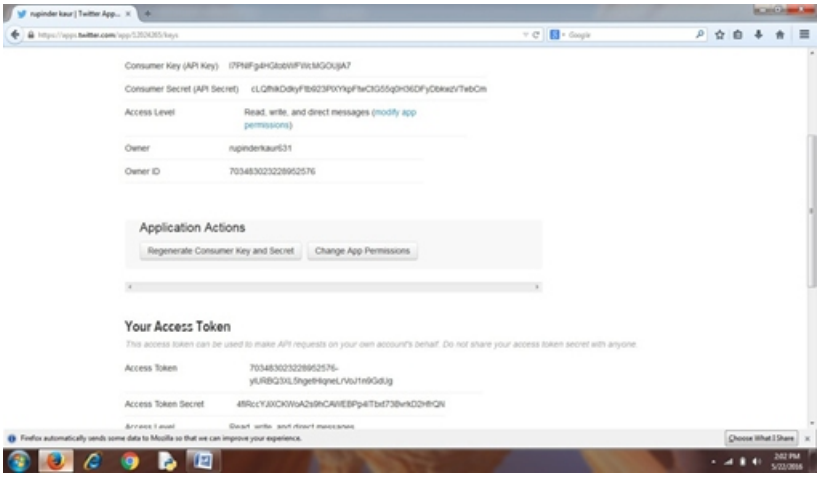


Figure 5: Key Used for Streaming

An array of the tweets is created and term document matrix is created using TFIDF score as shown below.

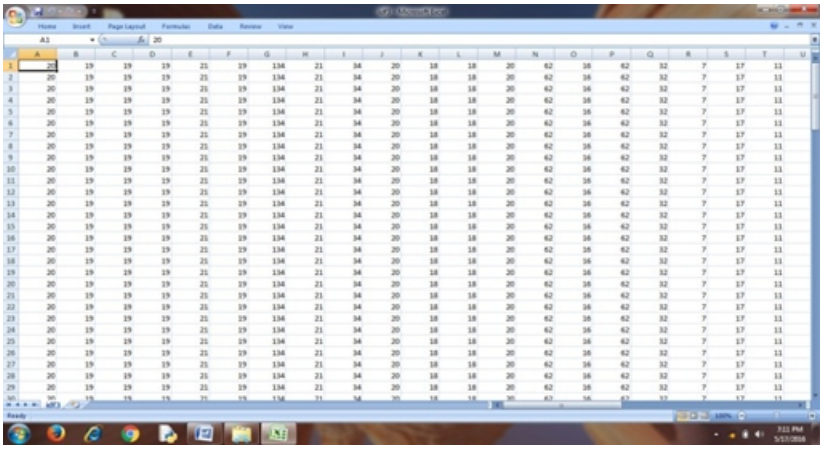


Figure 6: Creation of Term Document Matrix using TFIDF

- **Naïve Bayes**

When Naïve Bayes algorithm was applied on data of different sizes max accuracy achieved was 82%. The result of Naïve Bayes classifiers are shown in table 2.

Table 2:Result for Naïve Bayes Algorithm for different dataset

Total Test Record	Total Correct Records	Accuracy	Precision	Recall
42	28	82%	82%	74%
160	136	79%	80%	70%
312	159	81%	84%	76%

- **KNN**

When K- Nearest Neighbors algorithm was applied on data of different sizes then maximum accuracy of 83% was achieved. The result of KNN classifiers are shown in table 3.

Table 3: Result for KNN Algorithm for different dataset

Total Test Records	Total Correct Records	Accuracy	Precision	Recall
38	29	83%	78%	84%
436	324	81%	75%	84%
931	702	81%	75%	84%

- **KNN Hybrid**

Two classifiers have been analyzed in this: KNN and Naïve Baye's and a hybrid have been made using them. Hybrid scheme is the combination of two or more types, so we can design it in such a way that strengths of both types are maximized

When KNN Hybrid algorithm was applied on data of different sizes then it was observed that accuracy was enhanced as to Naïve Bayes and KNN. Precision of hybrid algorithm was found better than other two algorithms. The result of KNN classifiers are shown in table 4.

Table4: Result for Hybrid KNN Algo for

Total Test Record	Total Correct Records	Accuracy	Precision	Recall
42	33	84%	85%	76%
470	370	82%	80%	70%
1037	803	82%	84%	76%

V. CONCLUSION

A methodology for the classification of sentiments was developed in this study for food price data. Twitter API was used for streaming of tweets. The streamed tweets was filtered for relevant content and

content and stored in a database. The several steps of pre-processing were applied on it and the tweets were removed from special characters, stop word, tokenized, etc. Stemming was done to all words in order to extract the root words.

TF-IDF score based approach was utilized and the score was calculated for each tweets. The extracted features form a term document matrix which is utilized in the classification algorithm.

The results are found to be satisfactory and when comparative analysis was done between them it is found that KNN Hybrid outperforms Naïve Baye's and KNN Algorithm in terms of accuracy and precision. Thus an automated system is designed for opinion mining related to food price data.

REFERENCES

[1] Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. "Sentiment analysis in Facebook and its application to e-learning." Computers in Human Behavior 31 (2014): 527-541.

[2] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.

[3] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, 2011.

[4] Aisopos, Fotis, et al. "Content vs. context for sentiment analysis: a comparative analysis over microblogs." Proceedings of the 23rd ACM conference on Hypertext and social media. ACM, 2012.

[5] Balahur, Alexandra, et al. "Sentiment analysis in the news." arXiv preprint arXiv:1309.6202 (2013).

[6] Jebaseeli, A. Nisha, and E. Kirubakaran. "A Survey on Sentiment Analysis of (Product) Reviews." International Journal of Computer Applications 47.11 (2012).

[7] Scholar, P. G. "Big-SoSA: Social Sentiment Analysis and Data Visualization on Big Data."

[8] Horakova, Marketa. "Sentiment Analysis Tool using Machine Learning." Global Journal on Technology (2015).

[9] Gupta, Aditi, et al. "Sentiment analysis for social media." International Journal of Advanced Research in Computer Science and Software Engineering 3.7 (2013): 216-221.

[10] B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1 - 2):1{135, 2008.

[11] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79{86, 2002.

[12] Twitter Sentiment Classification using Distant Supervision by Alec Go, Richa Bhayani, and Lei Huang.

[13] Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Association for Computational Linguistics, 2005.

[14] K.Nigam, J. Lafferty, and A. Mccallum. Using maximum (2016) entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61{67, 1999.

[15] Alexandre Trilla. ”Sentiment Analysis of Twitter messages based on Multinomial Naïve Bayes” (2012).

.

Energy Minimization Techniques Over Multicore Processing System: A Review

K. Nagalakshmi[1], N. Gomathi[2]
[1] Department of Computer Science and Engineering, Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India
[2] Department of Computer Science and Engineering, Vel Tech Dr. RR & Dr. SR Technical University, Chennai, Tamilnadu, India

A B S T R A C T

Energy efficiency and processor performance have become key metrics in the designing of multicore computational systems. Due to breaking down of Moore's law, increasing energy savings without compromising raw performance is considered as a major limiting factor in multicore architecture. Recent technological advances in energy minimizing methods of multicore system substantially meet the contradictory demands of low power, low cost, small area and outstanding performance. This paper aims at ascertaining more competent energy-minimizing techniques for managing energy consumption of multicore processor through investigations. We highlight the necessity of the energy savings techniques and study several novel technologies to focus their pros and cons. This paper is intended to serve the researchers and architects of multicore processors in accumulating ideas about the energy savings techniques and to incorporate it in near future for more effective fabrications.

Keywords - Clock Gating; DVFS; Energy Efficient Design; Multicore Processor; Task Scheduling

I. INTRODUCTION

At present, multicore architectures (MCA) are becoming dominant design paradigm which assimilates two or more processing elements (cores) in a single die for higher performance computing. Proliferation of heavy computational requirements of real time applications in MCAs leads severe energy efficiency and performance constraint to provide quality of service (QoS) to the users. Thus, designing the MCAs to resolve power-performance tradeoff is a challenging endeavor. The research and design community have invested significant efforts in exploring several energy efficient technologies to scale their performance and make sure reliability, prolonged existence and acceptance in wide range of applications.

As stated in the Moore's law, the number of transistors fabricated in a single chip approximately doubles

every 18 to 24 months [1], resulting in an exponential increase in transistor density. This indicates that the speed (clock frequency) of the processor will also double in every 18 months. However we cannot enjoy this exponential growth continuously due to its increasing power density on chip which prevents all the cores to be switched on simultaneously. This utilization barrier is called as Dark silicon, is driving the emergence of heterogeneous MCAs [2]. Multicore processors can be categorized into three types: Homogeneous [3, 6, 8], Heterogeneous [10, 11] and Dynamic reconfigurable processors [12]. Conventionally, most of the general purpose MCAs are built with identical cores. All these cores consist of same micro-architectural innovations (i.e., cache memory, out-of- order execution, speculation, pipeline, branch prediction configuration, etc.) and are able to operate under same instruction set architecture (ISA). This type of architecture is called as a homogeneous or symmetric multicore architecture (SMP). It is easy to design and implement as we just need to duplicate the core.

INTEL CORE i7 [3], AMD PHENOM [4] and SUN NIAGARA [5] multicore architectures are general purpose SMPs with large cache memories. These processors are designed for general purpose desktop and server applications where energy efficiency is not a primary concern. In contrast, the homogeneous architectures XMOS-XS1 [6] and ARM-CORTEX [7] are specially designed for mobile devices, where energy efficiency is an increasing concern. Some of the MCAs are developed for high performance computing. Therefore, they employ larger number of cores. For example, AMD RADEON 700 GPU [8] contains 160 cores while NVIDIA G200 [9] contains 240 cores. Though homogeneous cores are simple to design, easy to implement and provide regular software environments; they cannot deliver required performance and energy efficiency for different real time applications. Real time applications with different QoS encourage the computer architects and software designers to exploit architecture innovations and design heterogeneous multicore processors.

Asymmetric Multicore Processor (AMP) or Heterogeneous processor implements mixture of non-identical processing elements that are asymmetric in their underlying principles and performance. The cores are varying in size and complexity, but they are designed to cooperate with each other to increase the performance of the system. These designs provide an efficient solution for dark silicon era and also increase reliability, performance Asymmetric Multicore Processor (AMP) or Heterogeneous processor implements mixture of non-identical processing elements that are asymmetric in their underlying principles and performance. The cores are varying in size and complexity, but they are designed to cooperate with each other to increase the performance of the system. These designs provide an efficient solution for dark silicon era and also increase reliability, performance and energy efficiency of the applications. A typical AMP will integrate small (slow) cores to process simple tasks in an energy efficient way, and complex (fast) cores to provide higher performance.

ARM big.LITTLE [10] and Cell BE [11] are renowned examples for AMP architectures. Cell BE is extensively used in gaming devices and computing platforms aiming high performance computing. ARM big.LITTLE is designed for mobile platforms where complex, performance driven, quad core Cortex A15 assembly is combined with simple, power-optimized, quad core Cortex A7 assembly to provide peak performance.

Dynamic heterogeneous multicore architectures are able to reconfigure itself at run time to regulate their performance, speed and complexity level based on application requirements. It has the ability to resolve the power-performance tradeoff by integrating efficient hardware with flexible software. By introducing adaptability and hardware flexibility, this dynamic architectures can achieve high performance within the energy budget and therefore to meet the QoS requirements of real time applications. Flexible heterogeneous Multicore processor (FMC) is an eminent example of dynamic reconfigurable architecture that can deliver both an increased throughput for uniformly distributed parallel workloads and outstanding performance for fluctuating real time tasks *12]. Depending upon the application requirements, the FMC can scale up and down its computing resources such as memory engines (ME), functional units, and pipelines that are anticipated to improved performance.

The evolution of multicore designs opens up a new space of research called energy saving techniques. Reducing the peak and average energy consumption could have a positive impact on performance of multicore processors, starting from circuit level to system level. In the last decade, several researches have been keen to explore various energy saving techniques in multicore regime. The remaining part of our survey is structured as follows: Section II presents the basics of power dissipation in multicore domain and also highlights the necessity of energy saving techniques. Section III provides an overview and taxonomy of classic power management techniques and explores some of these techniques in detail. Finally, Section IV is arranged to provide the conclusions of our investigation.

II. BACKGROUND

A. Basics of Power dissipation

Nowadays, multicore is ubiquitous both in general purpose and application specific computing systems starting from smartphones to commercial servers. Power dissipation is an important constraint because it increases the temperature and cooling costs, reduces reliability, and degrades performance. Power consumed by CMOS devices can be resolved into three parts [13] as shown in equation (1).

$$P_T = P_{st} + P_{dy} + P_{sc} \quad (1)$$

Where P_T is total power dissipation, P_{st} is static or leakage power due to leakage of a transistor's bias currents. P_{dy} is dynamic power due to switching of transistors. P_{sc} is power dissipation related to concurrent conduction of p type and n type transistors.

$$P_{st} = V.I_L \quad (2)$$

Where V denotes source voltage, I_L is the transistor leakage current. Usually, leakage power contributes 20 to 40 % of the total power dissipation [14]. Dynamic power is a prevalent factor as compared to other two components in equation (1). It can be denoted as follows

$$P_{dy} = \beta C_L . V^2.f \quad (3)$$

Where β is activity factor, C_L indicates the effective load capacitance and f is the switching speed. To simplify equation (3), it is assumed that the clock speed is linearly proportional to the source voltage. If we apply this notion to the above equation (3) then, the reduction in source voltage and switching speed lowers dynamic power cubically.

$$P_{dy} \alpha V^3 \quad (4)$$

Short circuit power is calculated by the following equation

$$P_{sc} = V.I_{sc} \quad (5)$$

Here I_{sc} represents the short circuit current flowing from supply to ground. Short circuit power is comparatively trivial for static CMOS circuits.

B. Failure of Dennard scaling

For almost 30 years, the computing community has realized a stable performance evolution in uniprocessor, motivated by Moore's Law *1] and Classical (Dennard) scaling [15]. But now this curve is slowed down and came to halt due to memory wall, power wall and Instruction Level Parallelism (ILP) wall [16]. Under Dennard scaling, the power requirements per uni tspace can remain constant across semiconductor generations. According to this principle [15], with a linear dimension scaling ratio of 0.707, the transistor count could doubles (Moore's Law), frequency increases by 40%, but the power consumed per transistor is reduced by half keeping the total chip power constant in every two years [17]. From equation (3), the power density in a chip area A is measured as follows

$$\text{Dynamic Power density} = (\beta C_{Load} . V^2.f)/A \quad (6)$$

As we move towards the next generation of IC manufacturing technology, the linear size of an IC gets scaled by 0.707. The same scaling ratio is applied for load capacitance and supply voltage while clock speed is scaled by 1⁄0.707. So the area of the chip is now 0.707^2A. If we calculate a new power per unit space, we have $0.707CX0.707V^2Xf/(0.707^3XA)$. Hence, the power per unit area becomes unchanged. But unfortunately, in 65nm technology and below, this law ceased because of exponential growth in leakage current and reduction in supply voltage decreases the speed of the processor. Nonetheless, the high performance demand is continued and this stimulates a shift from the single core to a multicore paradigm.

C. Need for Energy Saving Techniques

Today's innovations in semiconductor technology lead to not only an increase in the number of cores on a die, but also an increase in power density and concomitant heat dissipation. Increasing power dissipation leads many negative impacts on power delivery, performance/watt (PPW) ratio, packaging and cooling costs, reliability, availability and overall performance of the processors. So energy consumption issues occasionally more important than speed of the processor.

Energy efficiency is essential in mobile electronics where devices are battery powered. For last few decades, processor performance has been accelerating at a rate faster than the evolutions in battery technologies. This has led to a considerable drop of the battery life in mobile devices. At the same time, modern computational intensive applications demand very high performance. These two conflicting requirements, the need to conserve energy and the demand to deliver outstanding performance lead new approaches to resolve it. Existing researches to achieve optimal energy budget have two significant guidelines. First is in what way to increase the processor's efficiency within a specified energy limit. Second is in what way to decrease the energy consumption of computing devices without compromising processor performance.

III. OVERVIEW OF CLASSIC ENERGY SAVING TECHNIQUES

In this section, we delve into the up-to-date techniques in energy saving of MCAs. As of now, several hardware and software approaches have been adopted for alleviating power and energy costs. Through this investigation, we aim to demonstrate how the research community is trying to achieve an outstanding performance and energy efficiency. We can classify the power management techniques into three broad categories: Hardware techniques, Hardware-enabled middleware techniques and Software techniques

A. Hardware techniques

Several energy saving techniques with dedicated controllers are embedded into the modern processor architectures to provide energy efficiency. Applications running in a multicore domain need a carefully tailored computing architecture to meet their QoS within the power budget. The architectural innovations in designing of core, memory and interconnection networks improve the energy efficiency significantly.

1) Core Layout

Puttaswamy et al. propose a 3D microarchitecture with Thermal Herding techniques, which provides outstanding PPW ratio [18]. Compared to a conventional planar processor the proposed architecture can achieve 15% to 30% of active power reduction depending on the characteristics of the application. But 3D architecture incurs augmented power per unit area and associated temperature issues. This problem is resolved by Fazal Hameed et al. They present a thermal-aware 3D microarchitecture that effectively integrates the potential gains of dynamic architectural adaptation, fail-safe DVFS, and global migration [19]. Research shows that thermal-aware 3D architecture can achieve significant power reduction over 3D multicore processors [18] because it can reduce the active and the leakage powers simultaneously.

Kontorinis et al. introduce an adaptive processor with peak power guarantees which can reduce peak power by table-driven reconfiguration [20]. Most of the functional units (e.g. ALU, L1 cache, register files, load-store units and so forth) of the processor are dynamically organized for power conservation and maximum performance whereas peak power constraints are assured. Adaptive processor can reduce peak power about 25% with a smaller amount of performance cost.

Rodrigues et al. propose Dynamic Core Morphing (DCM) architecture for heterogeneous multicores [21]. The resources of the cores are morphed at runtime based on varying performance requirements. Depending on the computational need of the current workload, two cores may swap the execution units to maximize the PPW ratio.

The Scalable stochastic processor developed by Narayanan et al. has been demonstrated as an auspicious way to tackle power dissipation problem for error-tolerant applications such as audio or video. Improved scalability is realized by substituting or augmenting conventional computational units by gracefully degrading functional units [22]. The scalability leads power savings range between 20% and 60% in the well-known H.264 video encoder.

2) Memory Design

Many researches are carried out to bring innovations in memory organization in order to minimize the power dissipation. Smart caching [23] emphases on power saving computing techniques and implements way-predicting caches with reduced leakage designing techniques. Flaunter et al. [24] explore the use of instruction pre-fetch algorithms combined with the drowsy caches, where cache lines are periodically put into a low power mode without considering their access histories. Implementation of a drowsy cache in a 0.07μm CMOS process can reduce 50% to 75% of the total energy consumption in the caches. Cai and Lu present a joint venture for saving energy in system memory and hard disk unanimously [25]. This technique periodically reconfigures the size of physical memory by adding or freeing up the allotted memory pages and uses a timeout policy for shutting down the hard disk. The suitable memory size and timeout are selected according to their proportionality with the average power consumption. This technique achieves energy savings higher than 50% over a fixed-timeout scheme.

3) Interconnection network

Several researches show that the design choices for interconnect fabric have significant impact on the power budget [20, 23]. The interconnect network itself is a power consuming resource. The power consumption of the interconnection network for a 16 core processor is more than the combined power consumption of two cores. Rakesh Kumar et al. [26] demonstrates the need for careful co-design of interconnect network and memory hierarchy. The power consumption of the core increases super linearly with the number of connected units and average length of wire. So a power-optimized architectural design needs compact length of interconnection wire segments and appropriate routing algorithms [23].

B. Hardware-Enabled Middleware techniques

The following techniques are employed as middleware and partially implemented in hardware. The hardware enables middleware to shut down or slow down the functional units according to the operating temperature. Hardware-enabled middleware techniques including Stop-and-Go [27], Dynamic Voltage and Frequency Scaling [29], Advanced Configuration and Power Interface [50] and different Gating Techniques [44,46] have attracted a great deal of attention.

1) Stop-and-Go

The stop-and-go is the simplest form of dynamic power management (DPM) technique [79]. The DPM techniques reduce the power consumption by shutting down or lowering the performance of idle cores. Stop-and-go can be realized on both global and local scale. In global approach, if one of the cores reaches its specified threshold temperature, this scheme shut down the whole chip until its non-critical

level has been recovered. If stop-and- go is realized locally, only the overheating core will be halted until it has cooled down. Global stop-and-go mechanism provides a smaller amount of control and less efficiency as a particular overheating core leads to unwanted delaying of all other non-critical cores.

Donald et al. [27] implement 12 combinations of local and global stop-and-go policies with other power management techniques (i.e. DVFS and Task migration) for managing the temperature of multicore processors. They investigate the pros and cons of each combination by comparing their performance. Whenever peak temperature of the processor reaches 84.2° C the stop-and-go controller shuts down the cores for 30msec to allow the cores to cool down. Their implementation results show that local stop-and-go can outperform global schemes. Chaparro et al. [28] propose a stop-and-go mechanism with clock gating mechanism. Whenever the core reaches its critical temperature, this combined technique halts the core, stores its current state information, and then shuts the core off completely. There is no dynamic as well as leakage power consumption in this technique. Soit allows the overheating core to cool down quicker. As the current state of the core is saved before shutting down, this method would be the ideal choice of the options.

2) Dynamic voltage and frequency scaling (DVFS) [29]

DVFS is the prevailing and powerful DPM technique, used to regulate the power consumption of the processor by dynamically scaling the level of supply voltage and clock frequency [29, 30, 31, 32]. The sub-threshold leakage current and gate-oxide tunneling leakage can be reduced by reducing supply voltage [33]. Reducing the clock frequency reduces the supply voltage linearly and decreases power consumption quadratically [34]. DVFS is widely used for memory bound workloads and can be employed in two ways:

1. The Local (per-core) DVFS allow us to scale the voltage of individual cores, so that the overheating core can cool down faster [28, 35, 36].
2. The Global DVFS allow us to adjust the voltages and frequencies of all cores uniformly and simultaneously. Similar to stop-and-go, a single hotspot on one of the cores could result to unnecessary performance penalty on all cores [28, 35, 37].

Using Local DVFS introduces more flexibility as each core can select its own voltage– frequency pair individually. However, that suffers from a large number of expensive inherent voltage regulators. The global DVFS can solve the thermal issues faster but the efficiency of DVFS is affected by limited flexibility to determining a single optimal voltage to all cores.

Weiser et al. present the first paper to suggest an interval based DVFS for reducing the power dissipation in computing devices. Their work focuses on three scheduling algorithms: Unbounded-delay perfect-future (OPT), Bounded-delay limited-future (FUTURE), and Bounded-delay limited-past (PAST) [38]. The deployment of each algorithm controls the clock frequency and makes the scheduling decisions simultaneously. The PAST scheduling algorithm with a 50msec adjustment interval can achieve power conservation of 50% to 70% based on circuit conditions.

Wonyoung et al. develop a fast, per-core DVFS mechanism with on-chip integrated voltage regulators [32]. This mechanism uses the potential benefit of both per-core voltage regulation and very fine-grained voltage switching. The in-built regulators can increase the energy efficiency opportunities of DVFS and result in 21% of energy savings over conventional global DVFS with off-chip regulators.

Many researchers have unified DVFS technique with thread migration polices to reduce energy consumption. Cai et al. develop a new thread shuffling algorithm, which integrates thread migration and DVFS techniques on a MCA supporting simultaneous multithreading (SMT) [39]. Thread shuffling dynamically migrates slower threads with same criticality degrees to a particular core and implements DVFS for other cores having fast threads. The proposed scheme realizes energy savings around 56% with no performance degradation. Quan Chen et al. propose an Energy-Efficient Workload Aware (EEWA) task scheduler, that consists of a work load aware frequency adjuster and a preference-based task-stealing scheduler [40]. With the help of DVFS, the workload-aware frequency adjuster can accurately configure the frequencies of the cores in an efficient fashion based on the profiled workload statistics. The preference-based task-stealing scheduler can successfully distribute the tasks across various cores at runtime according to the preference list. The EEWA can save energy about 28.6% with only 0.9% of performance loss.

All the previous works cited above, simply fail to consider the static power that has turn into a substantial portion of the total power consumption, unfortunately. LeSueur and Heiser [41+ assess the factors influencing the efficiency of DVFS on AMD Opteron processors, using an extremely memory-bound benchmark. They illustrate that the ability of DVFS is retreating in modern digital systems due to escalating leakage power. Furthermore, their investigation reveals that switching-off idle cores will facilitate greater energy savings. To reduce leakage power, Awan et al. [42] propose an enhanced race-to-halt (ERTH) approach. By integrating DVFS and slack management policies, ERTH can improve energy efficiency considerably.

When global DVFS is realized in MCA, determination of optimal voltage that satisfies all cores is a challenging endeavor; some applications will suffer from performance penalty or overheads. This issue exacerbates as the running applications and number of cores in next generation processors. From a hardware implementation point of view, local DVFS is more expensive than global DVFS, because of its costly inherent voltage regulators and phase- locked loops. However, the per-core DVFS provides better tradeoff between performance and power.

3) Gating Techniques

Clock Gating [44] and Power Gating [46] are very useful methods for decreasing dynamic and static power correspondingly [43]. Gating techniques are realized by insertion of anadditional logic between the clock source and clock input of the processor's circuitry. It diminishes power consumption by logically turning off (gating) the power to the portions of core that are not useful to the current workload.

a) Clock Gating (CG) Techniques

The clock gating techniques employed in the Hexagon™ Digital Signal Processors (DSP) are analyzed by Bassett et al. [44]. The proposed four levels of clock gating and spine-based clock distribution allow switching off the power to the different regions, from single logic cell to entire chip. Further power reduction is achieved through a structured clock tree by distributing the clock signal across the chip with low skew and delay. This technique provides reduction in power consumption by 8% for active mode and over 35% for sleep mode.

Hai et al. describe a deterministic clock gating (DCG) technique, which hinge on the advance knowledge about at what time the functional block will be idle in the upcoming cycles [45]. With this advance information DCG can switch-off the idle blocks that maximize the energy efficiency. By exploiting DCG to various functional units, the proposed technique achieves 19.9% of average diminution in power without any performance cost. However, for all these techniques, the effectiveness of gating is restricted by the granularity of components that can be gated, the failure to change the overall size and complexity of the processor. Also, these designs are still vulnerable to leakage inefficiencies.

b) Power Gating (PG) Techniques

Power Gating (PG) is a circuit-level technique to reduce leakage power consumption by effectively turning off the source voltage to the idle elements. PG can be applied either at the core-level [46] or at the unit-level of the processor such as cache banks, ALUs, pipeline branches etc. [47, 48]. Recently, Intel Core i7 processors use power gating transistors to turn off its idle cores [49].

Hu et al. develop a parameterized model based on analytical equations, which decides the breakeven point used for proper gating. They evaluate the dynamic power gating ability of the fpu (floating-point units) and fxu (fixed-point units) of POWER4 processor by three techniques namely ideal, time-based, and branch-misprediction-guided [47]. The implementation of these techniques in various execution units shows that a considerable decrease in static power consumption can be realized through power gating.

Lungu et al. propose a Success Monitor Switch (SMS) and a Token Counting Guard Mechanisms (TCGM) for applying predictive power gating technique in POWER6 processor [48]. By employing SMS, the control logic enables or disables the PG depends on the success of policy. By implementing work for TCGM, this predictive power gating achieves a guarantee on the worst case execution of the policy. Leverich et al. [46] propose a Per-Core Power Gating (PCPG) technique with DVFS for datacenter workloads. This combined technique can save up to almost 60% of energy consumption.

4) Advanced Configuration and Power Interface (ACPI) [50]

ACPI is an industry standard for efficient handling of power management in computing devices. It is developed by the collaborative effort of Intel, Hewlett-Packard, Phoenix, Microsoft, and Toshiba [50]. ACPI provides platform-independent interfaces for power management and monitoring. These interfaces have the potential to work with existing DPM techniques [50]. ACPI is relay on operating system-directed configuration and Power Management (OSPM), which defines four switchable C-states (CPU idle states) C0, C1, C2, and C3 and n P-states (CPU-performance states) P0 to Pn for active power management. ACPI allows the processor to achieve fine tuning of the power consumption by moving idle devices into lower power states (sleeping state). Bircher and John [51] point out the implicit and explicit performance impacts of various CPU-idle states and Performance states of AMD quad-core processors. They verify their results for both compute-bound and memory-bound applications with fixed and OS scheduling. They develop an enhanced hardware and operating system configurations that decreases average active power by 30% with 3% of performance loss.

C. Software techniques

The performance per watt ratio of a MCA is depends on efficient built-in hardware and the ability of software to effectively control the hardware. Many up-to-date processors exploit software level power management techniques for energy efficiency. Recently, researchers have paid greater attention to the software power management policies because it can gain the power disparity statistics of processing threads on the fly with low cost. Software techniques can achieve predictable performance through transferring or scheduling tasks to minimize thermal gradients and hot spots. Software-based

approaches include data forwarding [52, 53] and task scheduling [54].

1) Data forwarding

Most modern processors use large size of on-chip L1 caches with multiple ports. Such a cache consumes a substantial part of the overall power owing to its larger size and high frequency access rate. Researches reveal that L1 data cache contributes 15% of the overall energy consumption of the processor [52]. Thus it is essential to develop tactics for precluding large power consumption in cache. Data forwarding is one of the appropriate solutions to reduce the energy consumption of L1 data cache.

Carazo et al. [53] propose a data cache filtering technique with forwarding predictor to reduce the power consumption of L1 data cache (DL1). This mechanism exploits an effective utilization of load-store queue (LSQ), which is responsible to provide the right data to load instructions by data forwarding method. Their experiments exhibit that the proposed cache filtering technique can achieve an average power savings up to 36% with a 0.1% of performance degradation. To reduce the access rate of DL1, Nicolaesu et al. propose a cached LSQ (CLSQ) to maintain load and store instructions after their execution [52]. Hitting in the CLSQ is faster and wastes not as much of energy as a DL1 access. Thus the significant savings in the frequency of accesses leads to 40% of energy reduction without any additional hardware complexity and performance penalty.

2) Task scheduling

Task scheduling is another breakthrough technique in power management arises from software approach. These algorithms are designed to solve temperature issues by distributing tasks among different cores. There have been wide ranges of literature put out on scheduling algorithms to achieve more processor utilization, better power conservation and more uniform power density without degrading the processor throughout. These algorithms schedule the tasks across cores based on predetermined temperature threshold. Work proposed by Hsin-Hao Chu and Yu-Chon Kao is a perfect example of how an adaptive thermal-aware multi-core task scheduling algorithm with multiple run-time controllers can mitigate the inter-core thermal costs and dynamic variations of task execution [54]. Implementations of run-time controllers increase the system complexity. To resolve this problem, the temperature-aware task scheduling algorithm, called Low Thermal Early Deadline First (LTEDF) is suggested by Wu et al. The LTEDF allocates tasks based on a novel History Coolest Neighborhood First allocation algorithm [55]. Simulation of the LTEDF algorithm demonstrates that it can satisfy the timing constraints for soft real time tasks and minimize the thermal consequences simultaneously.

Power aware task scheduling algorithms for MCAs can be classified into three types [56]: the global (dynamic binding) approaches [57], the partitioned (static binding) approaches [58] and the semi partitioned approaches [59, 60]. In the global approach, any core in the multicore system may execute any task. Global scheduling saves tasks in single priority- ordered queue, shared by all processors [57]. At every moment, the global scheduler chooses the highest-priority task for operation and the tasks are permitted to migrate between the processors. There are three types of priority assignment schemes for MCA: fixed-priority, [61, 58], dynamic priority [64] and Proportionate Fair (PFair) priority [65].

Fisher et al. develop a thermal-aware global scheduling algorithm for sporadic real-time tasks based on two priority assignment schemes, namely the global earliest-deadline-first (EDF) and the global deadline-monotonic (DM) [62]. The suggested schemes can substantially lessen the peak temperature around 30°C to 70°C as compared to load- balancing strategies.

Wang et al. develop a scheduling approach for hard real-time systems. They execute delay analysis for generic task arrivals using First-In-First-Out (FIFO) scheduling and Static-Priority (SP) scheduling with reactive speed control techniques [62]. But Andersson and Baruah prove that the fixed priority scheduling algorithms cannot achieve a utilization bound greater than 50% [58]. Some researchers handle this deficiency by PFair priority assignment. Baker addresses the aforementioned problem and demonstrates a schedulability test for preemptive deadline scheduling of periodic or sporadic real-time tasks [64]. Baruah and Shun-Shii Lin propose a new Pinfair algorithm that is very efficient in terms of runtime complexity and has a superior density threshold for a very large subclass of generalized pinwheel task systems [65]. Levin et al. propose a deadline partitioning algorithm, called DP- WRAP algorithm to handle sporadic task sets with arbitrary deadlines [66].

In the partitioned approaches, task set is partitioned and statically allocated to a designated processor. These task set are executed by existing scheduling algorithms and migration across core is not permitted. Fan et al. present a Partitioned Scheduling algorithm with Enhanced RBound (PSER) that exploits a flexible task set scaling technique and enhanced utilization bound for fixed-priority periodic real-time tasks [67]. This algorithm effectively improves the schedulability of the system. They combine PSER with Harmonic Aware Partition Scheduling (HAPS) in [68], which converts the complete task set into harmonic set and takes the benefit of harmonic relationship between tasks to achieve increased utilization bound up to 100% [69].

Andersson demonstrates global PFair and partitioned static-priority scheduling on multiprocessors [70]. Guan et al. develop two separate fixed-priority scheduling algorithms for light tasks and heavy

tasks. The algorithm RM-TS/light (Rate monotonic-task set for light loads) can execute light task sets with sustainable parametric utilization bound and the RM- TS algorithm can perform any task set, whereas the utilization bound is lower than a specified limit [71].

Recently, a significant portion of semi-partitioned approaches [59, 60, 72, 73, 74, 75, 76] have been proposed to minimize energy expenditure in multicores. Semi- partitioned algorithms allocate most of the tasks to one particular processor. But, limited tasks (i.e. less than number of cores – 1) are partitioned into many subtasks and are allocated to various cores under some constraints.

Lakshmanan et al. introduce a Highest-priority task-splitting (HPTS) algorithm to enhance the utilization bound of partitioned deadline-monotonic scheduling algorithms (PDMS) from 50% to 60% on implicit deadline task sets [75]. They can obtain 88% of average utilization with very low migration overhead for randomly generated implicit-deadline task sets by extend this algorithm, which assigns the tasks in the decreasing order in terms of their size. Kato et al. [74] and Andersson et al [60] present real-time scheduling algorithms with high schedulability. Similar to partitioned scheduling, the proposed algorithms assign each task to a specific processor but can divide a task into two processors if there is not sufficient capacity remaining on a processor. The semi-partitioned approaches are more efficient as compared to the conventional global and partitioned approaches theoretically [75, 76, 77] and also suitable for practical implementations [73]. Zhang et al. show that the implementation complexity of semi-partitioned scheduling algorithm is relatively low. They investigate semi-partitioned approaches in the Linux OS and demonstrate their results on an Intel Core-i7 processor [78].

IV. CONCLUSIONS

Current innovations in semiconductor technology lead to not only an increase in the number of cores on a die, but also an increase in power density and concomitant heat dissipation. This adversely affects the system reliability and availability. Achieving high performance with low power consumption is imposing a new challenge on IC fabrication technology. This article presents various effective techniques for alleviating power dissipation of MCA and its classification based on their attributes. We accept as true that our review will help the researchers and architects to acquire ideas into the next-generation multicore processors and encourage them to endorse new energy efficient elucidations for fabricating competent architectures.

REFERENCES

1. Gordon E. Moore. Cramming more components onto integrated circuits. vol. 86(1); pp. 82 – 85, IEEE (1998)

2. Christian Martin.: Post-Dennard Scaling and the final years of Moore's Law consequences for the evolution of multicore-architectures (2014)

3. Intel Corp.: Intel core i7-940 processor. Intel Product Information, 2009 [Online]. Available: http://ark.intel.com/cpu.aspx?groupId=37148

4. Advanced Micro Devices Inc.: Key architectural features—AMD Phenom II processors. AMD Product Information, 2008 [Online]. Available: http://www.amd.com/usen/Processors/ProductInformation/ 0,301181533115917% 5E159 19, 00.html

5. Johnson, T., Nawathe, U.: An 8-core, 64-thread, 64-bit power efficient SPARC SoC (niagara2). In Proceedings of 2007 International Symposium on Physical Design (ISPD '07), pp. 2 – 2, ACM, (2007)

6. May, D.: XMOS XS1 architecture. Micro IEEE 2012; vol. 32(6); pp. 28 – 37.

7. ARM Ltd.: The ARM Cortex-A9 Processors. ARM Ltd. White Paper, Sept.2007 [Online] Available: http://www.arm.com/pdfs/ARM CortexA-9Processors .pdf

8. Advanced Micro Devices Inc.: ATI Radeon HD 4850 & ATI Radeon HD 4870—GPU specifications. AMD Product Information, 2008. [Online]Available: http://ati.amd.com/products/radeonhd4800/ specs3. html

9. NVIDIA Corp.: NVIDIA CUDA: Compute unified device architecture. NVidia CUDA Documentation,J u n e 2 0 0 8 . [Online] Available: http://developer.download.nvidia.com/compute/cuda/2_0/docs/NVIDIA_CUDA_Pro gramming_ Guide_2.0.pdf

10. ARM Ltd.: ARM Development Tools, 2011. [Online] Available: http://www.arm.com/products/tools/development-boards/versatileexpr ess/index. php

11. Gschwind, M., Hofstee, H.P., Flachs, B., Hopkin, M., Watanabe, Y., Yamazaki, T.: Synergistic Processing in Cell's Multicore Architecture, vol. 26(2), pp.10 – 24, IEEE (2006)

12. Pericas, M., Cristal, A., Cazorla, F.J., Gonzalez, R., Jimenez, D.A., Valero, M.: A flexible heterogeneous multi-core architecture. In Proceedings of 16th International Conference on Parallel Architecture and Compilation Techniques, pp.13 – 24, IEEE (2007)

13. Zhuravlev, S., Saez, J.C., Blagodurov, S., Fedorova, A., Prieto, M.: Survey of Energy- Cognizant Scheduling Techniques, vol. 24(7), pp.1447 – 1464, IEEE (2013)

14. David Chinnery, Kurt Keutzer.: Overview of the Factors Affecting the Power Consumption. In proceeding of Tools and Techniques for Low Power Design, pp.11 – 53, Springer, (2007)

15. Dennard, R.H., Gaensslen, F. H., Yu, H., Rideout, V.L., Bassous , E., LeBlanc, A.R.: Design of ion-implanted MOSFET's with very small physical dimensions. In proceedings of IEEE Solid-State Circuits, vol. 12(1); pp. 38 – 50, IEEE (2007)

16. Hennessy, J.L., Pattersson, D.A.: Computer Architecture - A Quantitative Approach. Morgan Kaufmann, second edition (1996)

17. Hadi Esmaeilzadeh, Emily Blem, Renee St. Amant, Karthikeyan Sankaralingam, Doug Burger.: Dark silicon and the end of multicore scaling. In proceeding of 38th International Symposium on Computer Architecture (ISCA), pp. 365 – 376, IEEE (2011)

18. Puttaswamy, K., Loh.: Thermal Herding: Microarchitecture Techniques for controlling hotspots in high-performance 3D integrated processors. In proceeding of 13th
International Symposium on High Performance Computer Architecture, pp.193 – 204, IEEE (2007)

19. Fazal Hameed, Mohammad Abdullah Al Faruque, Jorg Henkel.: Dynamic thermal management in 3D multi-core architecture through run-time adaptation. In proceeding of Design, Automation & Test in Europe Conference & Exhibition, pp.1– 6, IEEE (2011)

20. Kontorinis, V., Shayan, A., Tullsen, D., Kumar, R.: Reducing Peak Power with a Table- Driven Adaptive Processor Core. In Proceedings of IEEE/ACM 42nd Annual International Symposium on Microarchitecture (MICRO-42), pp. 189 ─ 200, IEEE (2009)

21. Rodrigues, R., Annamalai, A., Koren, I., Kundu, S., Khan, O.: Performance per Watt Benefits of Dynamic Core Morphing in Asymmetric Multicores. In Proceedings of International Conference on Parallel Architectures and Compilation Techniques, pp. 121–13, ACM (2011)

22. Narayanan, S., Sartori, J., Kumar, R., Jones, D.: Scalable Stochastic Processors. In Proceedings of Design, Automation & Test in Europe Conference & Exhibition (DATE), pp.335 – 338, IEEE (2010)

23. Kornaros G.: Multi-core Embedded Systems. Taylor and Francis Group, CRC Press, (2010)
24. Flautner, K., Kim, N., Martin, S., Blaauw, D., Mudge, T.: Drowsy Caches: Simple Techniques for Reducing Leakage Power. In Proceedings of 29th Annual International Symposium on Computer Architecture, pp. 148 – 157, IEEE (2002)
25. Le Cai, Yung-Hsiang Lu.: Joint Power Management of Memory and Disk. In Proceedings of the conference on Design, Automation and Test in Europe, pp. 86 – 91, IEEE Computer Society (2005)
26. Kumar, R., Victor Zyuban, Dean M. Tullsen.: Interconnections in multi-core architectures: Understanding mechanisms, overheads and scaling. In Proceedings of 32nd International Symposium on Computer Architecture (ISCA), pp.408 – 419, IEEE (2005)
27. Donald, D., Martonosi, M.: Techniques for Multicore Thermal Management: Classification and New Exploration. In Proceedings of 33rd International symposium on Computer Architecture, pp.78 – 88, IEEE (2006)
28. Chaparro, P., Gonzalez, J., Magklis, G., Cai, Q. Gonzalez, A.: Understanding the Thermal Implications of Multicore Archtectures. IEEE Transactions on Parallel and Distributed Systems, vol.18 (8), pp. 1055 – 1065, IEEE (2007)
29. Isci, C., Buyuktosunoglu, A., C.-Y. Chen, Bose, P., Martonosi, M.: An Analysis of Efficient Multi-Core Global Power Management Policies: Maximizing Performance for a Given Power Budget. In Proceedings of 39th Annual IEEE/ACM International Symposium on MICRO-39, pp. 347 – 358, IEEE (2006)
30. Hanumaiah, V., Vrudhula, S.: Energy-efficient Operation of Multicore Processors by DVFS, Task Migration and Active Cooling. Computers, vol. 63, no. 2, pp. 349 – 360, IEEE (2012)
31. B. de Abreu Silva, Bonato V.: Power/performance optimization in FPGA-based asymmetric multi-core systems. In Proceedings of 22nd International Conference on Field Programmable Logic and Applications (FPL), pp. 473 – 474, IEEE (2012)
32. Wonyoung Kim, Gupta M S, Gu-Yeon Wei, Brooks D.: System level analysis of fast, per-core DVFS using on-chip switching regulators. In Proceedings of 14th International Symposium on High Performance Computer Architecture, pp.123 – 134, IEEE (2008)
33. Dongwoo Lee, Wesley Kwong, David Blaauw, Dennis Sylvester.: Simultaneous Subthreshold and Gate-Oxide Tunneling Leakage Current Analysis in Nanometer CMOS Design. In Proceedings of 4th International Symposium on Quality Electronic Design, pp. 287 – 292, IEEE (2003)
34. Burd, T., Pering, T., Stratakos, A., Brodersen, R.: A dynamic voltage scaled microprocessor system. In Proceedings of International Solid-State Circuits Conference, pp. 294 – 295, IEEE (2000)
35. Jayaseelan, R., Mitra, T.: A Hybrid Local-global Approach for Multi-core Thermal Management. In Proceedings of 2009 IEEE/ACM International Conference on Computer-Aided Design, pp. 314 – 320, IEEE (2009)
36. Pruhs, K., Van Stee, R., Uthaisombut, P.: Speed scaling of tasks with precedence constraints in approximation and online algorithms. In Proceedings of 3rd International conference on approximation and online algorithms, pp.307– 319, Springer (2006)
37. March, J.L., Sahuquillo, J., Hassan, H., Petit, S., Duato, J.: A new energy-aware dynamic task set partitioning algorithm for soft and hard embedded real-time systems. vol. 54, no. 8, pp. 1282 – 1294, The Computer Journal (2011)
38. Weiser, M., Welch, B., Demers, A., Shenker, S.: Scheduling for reduced CPU energy. In Proceedings of 1st USENIX conference on OSDI, pp. 13 – 23, ACM (1994)
39. Cai Qiong, Gonzalez, J., Magklis, G., Chaparro, P., Gonzalez, A.Q.:Thread shuffling: Combining DVFS and thread migration to reduce energy consumptions for multi-core systems. In Proceedings of 2011 International Symposium on Low Power Electronics and Design, pp. 379 – 384, IEEE (2011)
40. Quan Chen, Long Zheng, Minyi Guo, Zhiyi Huang.: EEWA: Energy-Efficient Workload- Aware Task Scheduling in Multi-core Architectures. In Proceedings of Parallel & Distributed Processing Symposium Workshops (IPDPSW), pp.642 – 651, IEEE (2014)
41. Le Sueur E., Heiser G.: Dynamic voltage and frequency scaling: The laws of diminishing return. In Proceedings of Hot Power: Workshop on Power aware computing and systems, pp. 1– 8 (2010)
42. Awan, M.A, Petters, S.M.: Enhanced Race-To-Halt: A Leakage-Aware Energy Management Approach for Dynamic Priority Systems. In Proceedings of 23rd EUROMICRO Conference on Real-Time Systems (ECRTS), pp.92 – 101, IEEE (2011)
43. Li Li, Ken Choi, Haiqing Nan.: Activity-driven fine-grained clock gating and run time power gating integration, vol. 21(8); pp. 1540 – 1544 IEEE (2013)
44. Bassett, P., Saint-Laurent M.: Energy efficient design techniques for a digital signal processor. In Proceedings of International Conference on IC Design & Technology, pp.1–4, IEEE (2012)

45. Hai Li, Swarup Bhunia, Yiran Chen, Vijaykumar T N, Roy K.: Deterministic Clock Gating for Microprocessor Power Reduction. In Proceedings of 9th International Symposium on High-Performance Computer Architecture, pp.113 – 122, IEEE (2003)

46. Leverich, J., Monchiero, M., Talwar, V., Ranganathan, P.: Power management of data center workloads using per-core power gating. vol. 8(2); pp. 48 – 51, IEEE (2009)

47. Hu, Z., Buyuktosunoglu, A., Srinivasan, V., Zyuban, V., Jacobson Bose, P.: Micro architectural Techniques for Power Gating of Execution Units. In Proceedings of 2004 International Symposium on Low Power Electronics and Design, pp.32 – 37, IEEE (2004)

48. Lungu, A., Bose, P., Buyuktosunoglu, A., Sorin, D.: Dynamic Power Gating with Quality Guarantees. In Proceedings of International Symposium on Low Power Electronics and Design (ISLPED), pp. 377– 382, IEEE (2009)

49. Rajesh Kumar, Glenn Hinton.: A Family of 45nm IA Processors. In Proceedings of IEEE International Solid-State Circuits Conference, pp. 58 – 59, IEEE (2009)

50. Hewlett-Packard, Intel, Microsoft, Phoenix Technologies, Toshiba: Advanced Configuration and Power Interface Specification, Revision 4.0a. April 2010, [Online]. Available: http://www.acpi.info/spec.html.

51. Lloyd Bircher, Lizy, W., John, K.: Analysis of Dynamic Power Management on Multi- Core Processors. In Proceedings of 22nd Annual International Conference on Supercomputing, pp. 327–338, ACM (2008)

52. Dan Nicolaescu, Alex Veidenbaum, Alex Nicolau: Reducing Data Cache Energy Consumption via Cached Load/Store Queue. In Proceedings of 2003 International Symposium on Low Power Electronics and Design, pp. 252 – 257, IEEE (2003)

53. Carazo, P., Apolloni, R., Castro, F., Chaver, D., Pinuel, L., Tirado F.: L1 Data cache power reduction using a forwarding predictor, vol. 6448; pp.116–125, Springer (2011)

54. Hsin-Hao Chu, Yu-Chon Kao, Ya-Shu Chen.: Adaptive thermal-aware task scheduling for multi-core systems. vol. 99; pp.155–174, ELSEVIER (2015)

55. Wu, G., Xu, Z., Xia, Q., Ren, J., Xia, F.: Task allocation and migration algorithm for temperature-constrained real-time multi-core systems. In Proceedings of 2010 IEEE/ACM International Conference on Green Computing and Communications, pp.189–196, IEEE (2010)

56. Carpenter, J., Funk, S., Holman, P., Srinivasan, A., Anderson, J., Baruah, S.: A Categorization of Real-time Multiprocessor Scheduling Problems and Algorithms. In Handbook on Scheduling Algorithms, Methods, and Models, pp. 30.1– 30.19 (2006)

57. Andersson, B.: Global Static-Priority Preemptive Multiprocessor Scheduling with Utilization Bound 38%. In Proceedings of ACM International Conference on Principles of Distributed Systems (OPODIS), vol. 5401; pp.73–88, ACM (2008)

58. Andersson, B., Baruah, S., Jonsson, J.: Static-Priority Scheduling on Multiprocessors. In Proceedings of 2nd Real-Time Systems Symposium, pp.193 – 202, IEEE (2001)

59. Kato, S., Yamasaki, N.: Semi-partitioned fixed-priority scheduling on multiprocessors. In Proceedings of 15th Real-Time and Embedded Technology and Applications Symposium, pp.23 – 32, IEEE (2009)

60. Andersson, B., Bletsas, K, Baruah, S.: Scheduling Arbitrary Deadline Sporadic Task Systems on Multiprocessors. In Proceedings of Real-Time Systems Symposium, pp. 385 – 394, IEEE (2008)

61. Lundberg, L.: Analyzing fixed-priority global multiprocessor scheduling. In Proceedings of 8th Real-Time and Embedded Technology Symposium, pp.145–153, IEEE (2002)

62. Fisher, N., J.-J. Chen, Wang, S. Thiele, L.: Thermal aware global real-time scheduling on multicore systems. In Proceedings of Real-Time and Embedded Technology and Applications Symposium, pp. 131–140, IEEE (2009)

63. Wang, S., Bettati, R.: Delay analysis in temperature constrained hard real-time systems with general task arrivals. In Proceedings of 27th IEEE International Real- Time Systems Symposium, pp. 323–334, IEEE (2006)

64. Baker, T.P.: An analysis of EDF schedulability on a multiprocessor. vol. 18(8); pp. 760 – 768, IEEE (2005)

65. Baruah, S.K, Shun-Shii Lin: Pfair scheduling of generalized pinwheel task systems. Transactions on Computers, vol.47 (7), pp. 812– 816, IEEE (1998)

66. Levin, G., Funk, S., Sadowski, C., Pye, I., Brandt, S.: DP-fair: A simple model for understanding optimal multiprocessor scheduling. In Proceedings of 22nd EUROMICRO Conference, pp. 313, IEEE (2010)

67. Ming Fan, Qiushi Han, Gang Quan, Shangping Ren: Multi-core partitioned scheduling for fixed-priority periodic real-time tasks with enhanced RBound. In Proceedings of 15th International Symposium on Quality Electronic Design, pp.284 – 291, IEEE (2014)

68. Ming Fan, Qiushi Han, Shuo Liu, Shaolei Ren, Gang Quan, Shangping Ren: Enhanced fixed-priority real-time scheduling on multi-core platforms by exploiting task period relationship. pp.85– 96, Elsevier (2014)

69. Liu, J.W.S.: Real-time systems. Prentice Hall (2000)

70. Andersson, B., Jonsson, J.: The utilization bounds of partitioned and pfair static priority scheduling on multiprocessors are 50%. In Proceedings of 15th EUROMICRO Conference on Real-time Systems, pp.33 – 40, IEEE (2003)

71. Guan, N., Martin Stigge, Wang Yi, Ge Yu: Parametric Utilization Bounds for Fixed- Priority Multiprocessor Scheduling. In Proceedings of 26th International Parallel and Distributed Processing Symposium, pp.261-272, IEEE (2012)

72. Anderson, J.H., Bud, V., Devi, U.C.: An EDF-Based Scheduling Algorithm for Multiprocessor Soft Real-Time Systems. In Proceedings of EUROMICRO Conference on Real-Time Systems (ECRTS), pp.199– 208, IEEE (2005)

73. Bastoni, A., Brandenburg, B.B, Anderson, J.H.: Is Semi-partitioned scheduling practical? In Proceedings of 23rd Conference on Real-Time Systems, pp. 125 – 135, IEEE, (2011)

74. Kato, S., Yamasaki, N.: Semi-partitioned fixed-priority scheduling on multiprocessors. In Proceedings of 15th Real-Time and Embedded Technology and Applications Symposium, pp. 23 – 32, IEEE (2009)

75. Lakshmanan, K., Rajkumar, R., Lehoczky, J.P.: Partitioned fixed priority preemptive scheduling for multi-core processors. In Proceedings of 21st EUROMICRO Conference on Real-Time Systems, pp. 239 –248, IEEE (2009)

76. Guan, N., Stigge, M., Yi, W., Yu G.: Fixed-Priority Multiprocessor Scheduling with Liu and Layland's Utilization Bound. In Proceedings of IEEE Real Time and Embedded Technology and Applications Symposium (RTAS), pp. 165 – 174, IEEE (2010)

77. Guan, N., Martin Stigge, Wang Yi, Ge Yu: Fixed-Priority Multiprocessor Scheduling: Beyond Liu and Layland's Utilization Bound. In Proceedings of WiP Real-Time Systems Symposium (RTSS), pp. 1594– 1601, IEEE, (2010)

78. Zhang, Y., Guan, N., Yi. W.: Towards the Implementation and Evaluation of Semi- Partitioned Multi-Core Scheduling. In Bringing Theory to Practice: Predictability and Performance in Embedded Systems. vol. 18; pp. 42– 46, In Open Access Series in Informatics (2011)

79. Young-Si Hwang, Ki-Seok Chung: Dynamic Power Management Technique for Multicore Based Embedded Mobile Devices. IEEE Transactions on Industrial Informatics, vol. 9(3), pp. 1601 – 1612 (2013)

Thwart The Capturing Of Videos And Images In Unauthorized Place Through Camblocker

S. Narmadha[1], S. Ashimabaanu[2], S. Umadevi Yasodhei[3]
[1] Final Year, CSE, IFET College of Engineering, Villupuram
[2] Final Year, CSE, IFET College of Engineering, Villupuram
[3] Associate Professor, CSE, IFET College of Engineering, Villupuram

ABSTRACT

Sensor network and mobile application are ubiquitous today. Mobiles are used for both legal and illegal process. Most of the people use the mobile in apposite manner but some of their used in bad comportment. For example, capture the videos in the theatre. We are proposing an automated system that will protect from video recording and capturing images in an authorized place such as theaters, temples, trial rooms etc., and also for women's safety. The proposed system consists of a sensor network and mobile application named as CamBlocker. Thisapp will be inbuilt in each person's mobile.If the person entersinto the range of the signal, then the sensor sense the CamBlocker application and activated the app. After the activation of CamBlocker app, that will automatically chunk the camera application in our mobile.The CamBlocker will protect the person to take unwanted images and videos. The person who tries to open camera inside the range of the signal the app will automatically close the camera and prevent from recording.
Keywords -Mobile Application, Sensor Network, CamBlocker.

I. INTRODUCTION

Nowadays, cameras are used frequently. Since 2004, Japan was used 75% of smart phones [1]. From the past 2015, more than 90% of people using smart phones and it is expected swell to 100% of users in future. 83% of all phones have cameras [2]. The cram says that 90% of all people take pictures only done on camera mobile. The statistics from 2007 to 2015 will be shown in the Figure 1.Most probably they used their smart phones for capturing the picture and recording the videos. The working process of camera is similar to the function of the human eye [3]. Camera mobiles are mostly used for capturing memories that means we can take a picture of our friends or our family's trip to the beach, museums, temple etc. if we want, we can print the picture in papers or we can just view them on a systems, laptops, etc. We can also use it as a scanner and it is easier to take a snapshot and upload the taken picture when we using the smart phone. In this paper, we proposed a new technology to block the camera application while recording in theatre, trial rooms, temples etc.

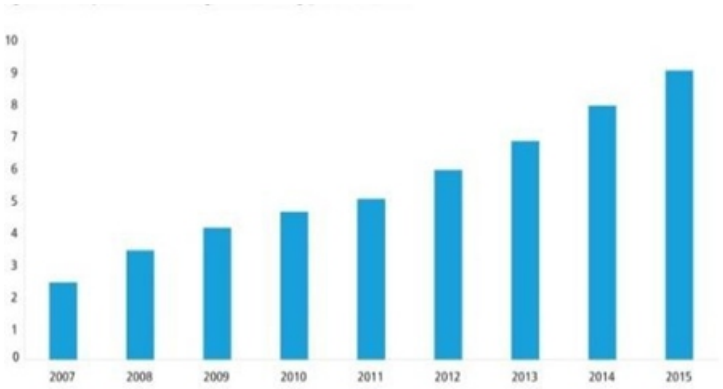


Figure1: Bar Chat

II. EXISTING AND PROPOSED SYSTEM

a) Existing System

In the existing system [1], they use two devices such as camera detector and camera neutralizer. All cameras contain CCD (charge coupled devices) lens and CMOS (Complementary Metal Oxide Semiconductor) sensor which are used to capture the picture and recording the videos. Only the CCD and CMOS sensor can operate the camera to snap the picture and videos. Camera detector is used to detect the CCD lens of the camera and the camera neutralizer used to neutralize the camera mode. Sothat the user can't able to taking the images and recording the videos.

b) Proposed System

There is a disadvantage in this existing system (i.e.) the cost of the devices is too high and it is only available in the aboard not in India. In thisproposed system, we use the automated system that contains Mobile Application and Sensor Network with Signal Radiator for blocking the camera application in our mobile phones. An application will be provided to each person's mobile by default. Whenever the person enters into the range of the signal the mobile will be automatically paired with the signal. The application will protect the person from opening the camera in the mobile. Whenever the person tries to open camerainside the theatre the application will automatically close the camera and prevent from recording. The comparison of existing technique with proposed technique will show in the Table 1.

Table 1: comparison of existing techniques with proposed techniques

Technique	Existing Technique	Proposed Technique
Blocking Camera	Camera Detector and Camera	Sensor Network with Signal Radiator and
	Neutralizer	Mobile Application

III. HARDWARE

The Kit contains Micro Controller, Universal Asynchronous Receiver and Transmitter (UART) and Liquid Crystal Display (LCD) and Sensor Network with Signal Radiator. The Micro Controller needs power supply, which contains Step down transformer, Bridge Rectifier, Filter Circuit and Voltage Regulator. The Sensor is going too paired with the Mobile application andblocks the camera application. The block diagram and sample kit of our proposed system will be shown in Figure 3 and Figure 4 respectively.

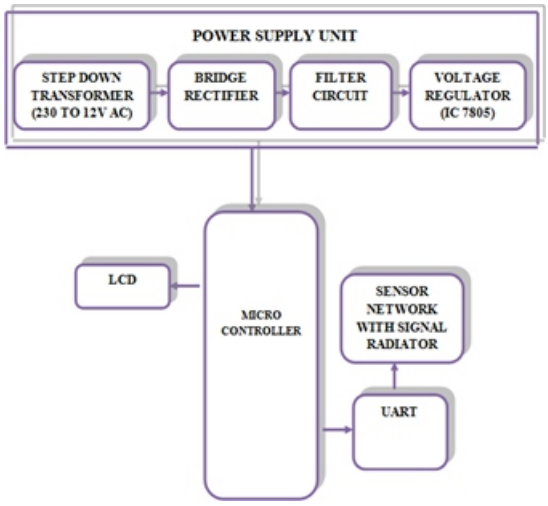


Figure 3: Block diagram

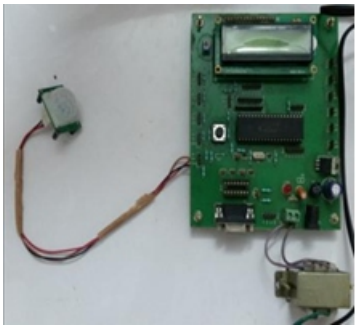


Figure 4: Sample kit

IV. Software

CamBlocker is the inbuilt application which is paired into the sensor network to block the camera application. To develop the mobile application, we used Android-Studio (SDK).

Figure 5: Smart Phone

III. HARDWARE

The Kit contains Micro Controller, Universal Asynchronous Receiver and Transmitter (UART) and Liquid Crystal Display (LCD) and Sensor Network with Signal Radiator. The Micro Controller needs power supply, which contains Step down transformer, Bridge Rectifier, Filter Circuit and Voltage Regulator. The Sensor is going too paired with the Mobile application andblocks the camera application. The block diagram and sample kit of our proposed system will be shown in Figure 3 and Figure 4 respectively.

V. Emulation

The Camera application will be blocked using CamBlockerApplication and Sensor Network with Signal Radiator. The block diagram of the Hardware is shown in the Figure 3. The Sensor Network can be worked with the support of Micro Controller. Micro controller is a small computer with an integrated circuit. It is used to controls the products and devices such as Automobile engine control system, Appliances, Remote system etc. There is an power supply for Micro Controller which consist of Step down transformer is used to convert Secondary voltage to primary voltage, it transforms 230 v to 12 v AC. Bridge Rectifier is used to converts the voltage AC to DC. Filter Circuit is used to remove unwanted frequency from the Bridge rectifier. Voltage Regulator is used to maintain a constant voltage level and also remove unwanted signal or noise. LCD is to display the status of the Micro Controller. UART is the universal Asynchronous receiver and transmitter, which is the computer interface to its attached serial devices. Sensor Network is activated with the help of Micro Controller and UART. The Sensor Network senses the mobile application which is inbuilt in our Mobile. Whenever the person enters into the range of the signal, the Mobile application is paired into the Sensor Network. Once it paired into the Sensor the camera application will be blocked. Consequently they do not take photos and videos. The working process of our proposal is shown in Figure 6.

a) Working process:

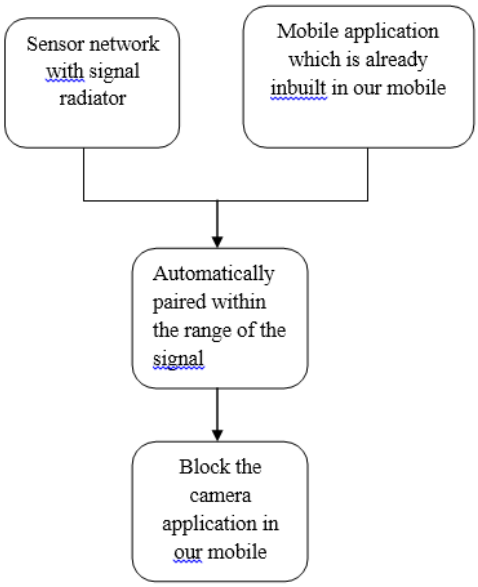


Figure 6: Working process

V. CONCLUSION

The raise of mobile phones including camera is more ubiquity. Some of the people use the camera in appalling manner [4][5][6]. We propose an automated system to prevent the pirating of capturing image and video. This system contains the sensor network and mobile application. The mobile application will automatically pair with the sensor when the person who having the mobile, enters into the range of the signal. This automated system will be 100% useful for avoid the recording of videos in theatres, trail rooms, etc. In this approach, the coverage area is squat. This technique only used for blocking the camera which is in the mobile, not all the cameras like hidden cameras. The future work will be, to increase the coverage area like large environment and to block the hidden cameras also. This technique mainly used for the women's safety and has to be more useful for entities, such as people.

REFERENCES

[1] Khai N. Truong, Shwetak N. Patel, Jay W. Summet and Gregory D. Abowd "Preventing Camera Recording by Designing a Capture-Resistant Environment"

[2] https://photofocus.com/2013/11/10/90-of-people-have-only-taken-a-photo-with-a- camera-phone-in-their-lifetime/

[3] https://en.wikipedia.org/wiki/Camera

[4] Art. 29 Data Protection Working Party. Opinion 4/2004 on the Processing of Personal Data by means of Video Surveillance. Document 11750/02/EN WP89, European Commission (2004). http://europa.eu.int/comm.

[5] Chung, J. Threat of Subway Photo Ban Riseth Again," Gothamist, 2004 November 30.

[6] Video Voyeurism Prevention Act of 2004. 18 USC 1801. December 2004.

Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

-

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).

2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);

3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);

4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);
2. Informative contribution (editorial, commentary, etc.);
3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.
The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.

Address of the Editorial Office:

Enriched Publications Pvt. Ltd.
S-9,IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR -5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-45525005