Document Management System:A Notion Towards Paperless Office

Mahendra K. Ugale
Dept. of Computer Science & Engineering
MGM's Jawaharlal Nehru Engineering College,
Aurangabad, (M.S), India.
Email-mahendraugale@jnec.ac.in

Shweta J. Patil
Dept. of Computer Science & Engineering
MGM's Jawaharlal Nehru Engineering College,
Aurangabad, (M.S), India.
Email-patilshweta10@gmail.com

Dr. Vijaya B. Musande
Dept. of Computer Science & Engineering
MGM's Jawaharlal Nehru Engineering College,
Aurangabad, (M.S), India.
Email- vijaya.musande@gmail.com

Abstract— **Paperless Document Management System is used to eliminate the losses that businesses suffer because of physical paper files and filing systems. This Paper addresses some of the technologies that are helping professionals shift toward a paperless business world.**

A DMS based on organizing digital documents to search and store documents and to reduce paper. Most of the workplace consists a variety of documents having a mixture of handwritten and printed text. The detection of such documents is a crucial task for OCR developers. This paper describes different steps for processing different documents using scanning, tagging, and indexing for effective data retrieval with OCR and Indexing techniques.

Keywords: Paperless, Document Management System, OCR, Tagging, Indexing, Information Retrieval.

I. INTRODUCTION

The physical Paper process used handling of a physical document, from a shelf or filing cabinet or cupboard. Filing systems require a large amount of physical space and having inefficiencies in searching for papers. Cost required maintaining such type of physical documents is also large. Organizations that use paper-based processes also having security risks. A paperless office should be considered as an important project within organizations and establish an initiative with good environmental practices as a contribution to sustainable development [1].

Document Management System

Document management systems are generally filing cabinets that provide architecture for organizing all digital documents. These systems work mainly with scanners,

Which convert paper documents into digital formats. Through search engines, document management systems allow for quick retrieval to any document or file.

Document management system is commenced as a notion to lower handling of paper. EDMS is capable to handle document handling processes and to manage the flow of information. EDMS has different parts of capacity in terms of storage, archiving, administration, flow control of DMS reports to encourage work processes. In any case, intricacy and cost of operation of EDMS still force issues for being moderate for SME's [2].

A DMS includes five components:

I. Input

A decent scanner will make putting paper documents into your PC easily. Likewise, documents can be input utilizing photo, which is considered as softcopy.

II. Storage

The storage system provides a capacity to store documents. A decent storage system will oblige changing documents, expanding data volumes.

III. Indexing

The index framework will make a composed documenting framework and makes future searching or retrieval successfully. A decent indexing framework will make existing procedures and systems more successful.

978-1-5090-4264-7/17/$31.00 ©2017 IEEE

IV. Retrieval

The retrieval framework utilizes fundamental data about the documents along with an index and related content to recover pictures stored in a system.A productive retrieval system will make finding the required archives speedier and simpler.output should be Eco-friendly.

Documents are viewable to only authenticated personnel, whether in the workplace, in various areas, or over the Internet.

V. Retention

In record accessing system document retention schedule must be characterized. This permits the documents that came to end of their maintenance calendar to be distinguished and decimated. By overwriting distinctive records number of times the reports are successfully "destroyed."

Architecture of Document Management System

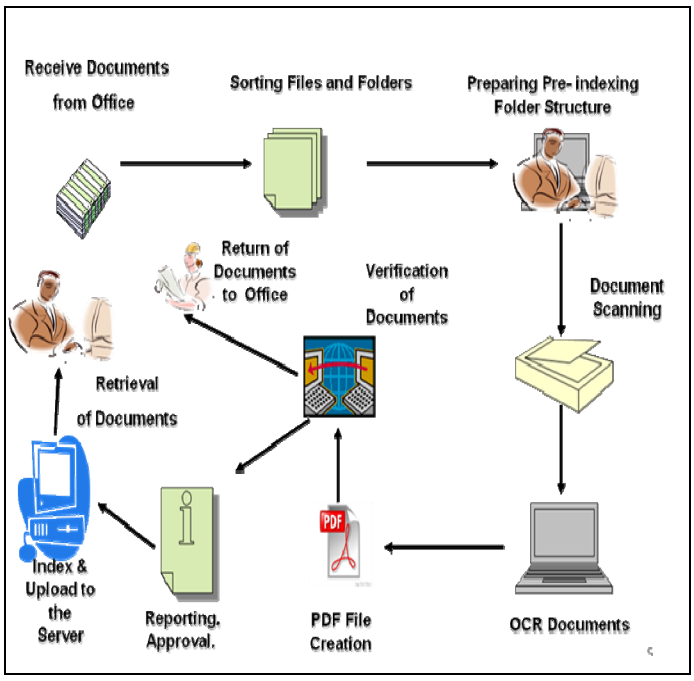

Fig-1 Document Management System

Electronic documents such as PDF's are becoming increasingly popular as we move further towards the notion of the document management system [13].For a paperless office, it is necessary to reform administrative processes by the organization [6].

Different Phases

- Scanning of Documents-Normally scanning at 300 dpi is recommended and Maximum dpi limit can be up to 600**.**
- Tagging of PDF Documents (Scanned Documents)-Read page images, Analyze page images, Recognize the contents of the image.
- Indexing of PDF Documents-Collect all PDF documents to be indexed into one or more folders.
- Searching of PDF Documents-Searches can be done by specifying single string, or multiple search strings or by using patterns.

In this paper, an emphasis is given on last three phases i.e. tagging of PDF documents, indexing of PDF documents and searching of PDF documents. The aim is to improve information retrieval from PDF documents by reducing time using effective tagging and indexing.

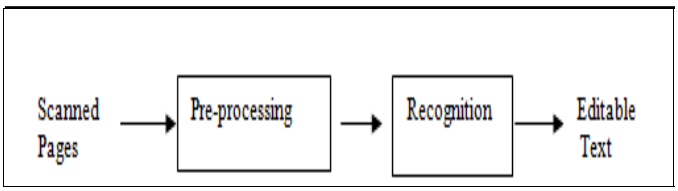

Fig-2 Tagging of Documents

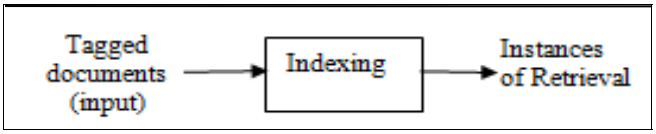

Fig-3 Indexing of Documents.

II. Related Work

A. Overview of different OCR Software

- Nuance Omni Page Ultimate

Initially used OCR software is Nuance OmniPage Ultimate. When its "OCR & train" process starts, OmniPage recognizes an entire page and then brings up a window with a grid of all doubtful characters. The user then goes through and corrects any character that OmniPage has incorrectly identified, after which the program OCRs the page and returns results. This approach may work well for texts in English (or one of the other nine languages that OmniPage supports), but it fails for language like Devanagari [9].

Testing Observation-

- Intermediate file (.opd) is working inefficiently, It takes time to load input file and to save output file, It doesn't support large file,Working behavior of software is dynamic.
  The recovery system is not appropriate, Retrieval of data is not satisfactory, and Re-editing of the document is tedious.

- Abbyy Finerader

ABBYY FineReader is an OCR framework that examined document picture records including advanced photographs i.e. digital photos into an editable format. [10].

Testing Observation-

- Simple GUI .
- Faster and Accurate Recognition.
- Supports Most Of The World's Languages.

- SimpleOCR

This is freeware OCR application that gives flexible accuracy for those who just want to convert a few no of pages using OCR. Developers can use SimpleOCR with their custom applications.

- Simple Index

Simple Index is designed to enable to quickly organize multiple documents in batches, extract data from text and barcodes with OCR, and then use that data to organize the files automatically.

Functionalities of Tested OCR Softwares

Table I- Functionalities of Tested OCR Softwares

Criterion	Simple OCR	ABBYY Fine-Reader	Nuance Omni-page Ultimate	Simple Index
Scanner Driver Supported	TWAIN	TWAIN	TWAIN	TWAIN
Table/ Spreadsheet Recognition	✘	✓	✓	✘
Searchable PDF Output	✘	✓	✓	✘
PDF Password Support	✘	✓	✓	✘
Vertical Text Recognition	✘	✓	✓	✘
Barcode Recognition	✘	✓	✓	✘
Image Pre-processing	✘	✓	✓	✘
Hot Folder	✘	✓	✘	✘
Language Supported	3	189	137	179
Installation	Desktop	Desktop/ Network	Desktop	Desktop

Based on above comparison of tested OCR software it is observed that Abbyy Fine reader OCR software is competitively better.

B. Different Indexing Software

- Nuance PDF Converter Professional
  - It provides an effective document conversion to PDF. It also offers
  - one-click scanning to PDF, Advanced PDF search capabilities,
  - Accurate PDF to Microsoft Excel conversion,
  - Enhanced multimedia support,
  - Better graphics management,
  - document flattening .[11]

Testing Observation-

- Difficult User Interface
- It missed out some instances while searching.
- Retrieval of data is not satisfactory.
- Accuracy for Information retrieval is less.

- Adobe Reader

A document that consists of scanned images of document is a consisting of scanned images of text which is inaccessible because the content of the document is a graphic representation not searchable text.

Scanned images of text must be converted into reachable and readable text using OCR. [12]

Testing Observation-

- Simple User Interface.
- It gives all correct instances while searching.
- Retrieval of data is satisfactory.
- Accuracy for content retrieval is more.

Table II- Functionalities of Tested Indexing Softwares

Criterion	Nuance PDF Converter Professional	Adobe Reader
User friendly	No	Yes
Retrieval	Not satisfactory	Good
Indexing Time	Less	More
Accuracy	Not satisfactory	Good

Based on above comparison of indexing software it is observed that Adobe Reader software is better for indexing PDF documents.

III. System Implementation

For Tagging of PDF documents, the proposed framework depends on the implementation of OCR with help of Asprise SDK. For Indexing of PDF archives, the proposed framework depends on implementation with the help of Hoot SDK.

The input for the OCR system is scanned document image. To perform character recognition , the application needs to go Through three steps-

- **Optical Scanner**

Character images can be acquired from a scanner or any other electronic source. It will be stored in any image format and that image may be a color, gray scale, but the actual processing will take place on binary images [3].

- **Preprocessing**

The resulting image from the scanning process may contain noise. Depending on the resolution of the scanner, the characters may be broken. These defects cause poor recognition rates that can be eliminated by using a preprocessing

- **Segmentation**

Given an input image, identify individual glyphs (basic units representing one or more characters, usually contiguous).

- **Feature Extraction:**

From each glyph image, extract features to be used as input of ANN. This is the most critical part of this approach. [3]

- **Classification:**

There are no of classification or recognition approaches that are as follows-

  - Template Matching,
  - Neural networks,
  - Decision tree based
  - Bayesian-based,
  - Support Vector Machines SVM[3].

This paper utilized the toolkits of Hoot that used for information retrieval.

- **Preprocessing Module:**

Before utilizing Hoot we have to preprocess the available text documents. The principle part of preprocessing is to change full-width characters into half-width characters. So as to better show the utilization of Hoot, this will divide the large documents into small documents and then assign a unique ID number to each document.

- **Indexing Module:**

In the wake of Preprocessing you can utilize Hoot to process significant data.

1. Index Creation for handling documents
2. Build a query object.
3. Pursuit or search in an index.

- **Searching Module:**

After indexing process, framework will build up a search class which gives two methodologies, Index search approach is utilized to search indexing which is constructed by Hoot. Nonetheless, string search approach used to seek information. To begin with, give the pursuit path then parse the string and create query object to search data.

A. *Tagging of PDF Documents using Asprise.*

Asprise OCR SDK is a commercial optical character recognition and barcode identification recognition SDK library that gives an API to perceive the text and barcodes from document photo and yield in formats like plain text, XML and accessible PDF [7].

Asprise C# .NET OCR (optical character recognition) and barcode recognition SDK offers an elite API library for you to develop your C# .NET applications (Windows applications, Silverlight, ASP.NET web service applications, ActiveX controls, and so on.) with the usefulness of extracting content and barcode information from examined reports.

Highlights

- **High Accuracy**

Low resolution documents can be easily recognize by Asprise OCR.

- **Format retention:**

Input document's Text layouts are preserved effectively.

- **High Speed:**

High Speed to perform recognition effectively in short time

- **Ease of utilization:**

Complex are expelled from Asprise OCR SDK

- **Barcode Recognition:**

  Alongside letters and numbers Asprise OCR can perceive practically every sort of barcode.

  You can choose to recognize characters or barcodes or both.

Algorithm of tagging of documents-

This flowchart explains different steps for tagging of documents.

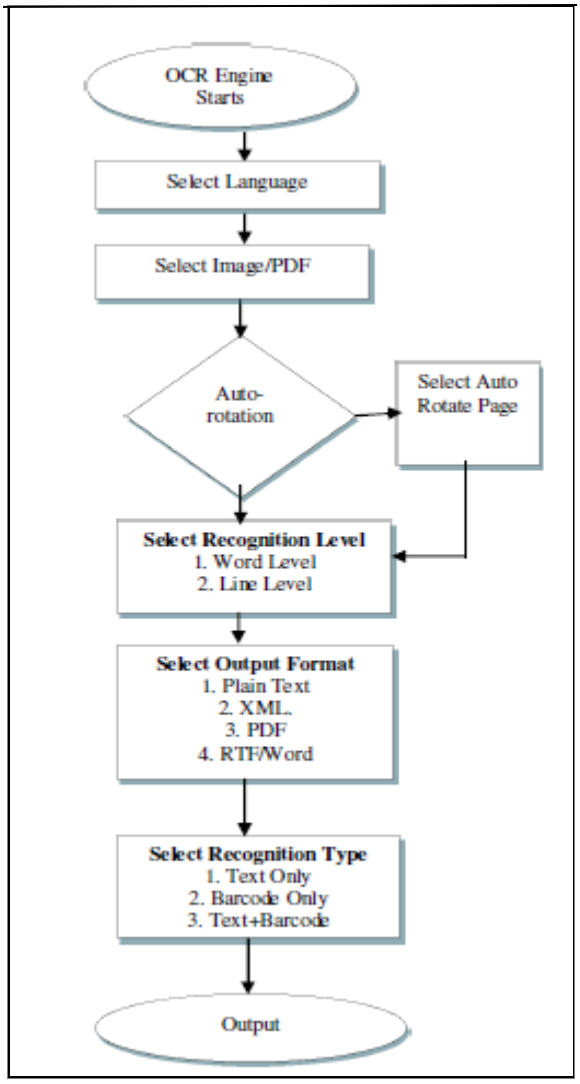

Fig 4 – Algorithm of Tagging of Documents

B. *Indexing of PDF Documents using Hoot.*

Hoot is a framework that considered as

- Part of the analysis engine,
- It can likewise give a straightforward however intense interface with the goal that individuals can helpfully and rapidly build up the crawler.
- Hoot is the littlest utilization of information retrieval using highly compact storage inverted WAH bitmap index and data recovery is the way toward hunting down words in a piece of content.
- A Text indexing engine
- A Query engine.

Features-

- Quick operating speed.
- Incredibly little code estimate.
- Uses WAH compacted BitArrays to store data.
- Multi-threaded execution.
- Tiny estimate just 38kb DLL
- Highly optimized storage.

Information Retrieval in HOOT-

The Important part of indexing module is to the speedup of recovery; the scanning module is principally utilized for cooperating with users[5].

An index is made out of various sections, a section is a mix of documents, a document is a combination of fields.
a Once documents are built and analyzed, next step is to index them so that this document can be retrieved based on certain keys instead of whole contents of the document.
Indexing process is similar to indexes which are at the end of a book where common words are shown with their page numbers so that these words can be tracked quickly instead of searching the complete book.

Working of HOOT:

- Set the index storage, it will store all the index information.
- Select Specific location or folder where you need to index for content.
- Load Hoot: It will stack Hoot so you can inquiry a current record using existing index.
- Begin Indexing : It will stack Hoot and begin indexing the folder you indicated.
- Stop Indexing : It will come dynamic after you have begun indexing so you can stop the procedure.
- Count Words : It will demonstrate the quantity of words in Hoots lexicon .
- Save Index : It will save index in memory to disk.
- Free memory : It will call the interior free memory method on Hoot
- You can look for content; it will demonstrate the count and time taken..
- To open the document simply double clicks on the record way in the listbox.

Indexing Technique-

There are number of popular information retrieval (IR) indexing techniques are available; the technique which used by Hoot is called Inverted Index.

- **Inverted index**

An inverted index is a special index which stores the words to bitmap conversion data.

In an ordinary index you would store what words are in a specific document, an inverted index is the inverse given a word 'x' what records have this word is stored.

A bitmap index is an index based record numbers to the archive. Bitmap indexes were initially additionally utilized as a part of Information Retrieval (IR), however, are today mostly replaced by inverted index.

WAH [15] is one of a few presented compression techniques. Despite the fact that there are plans with more minimized indexes, WAH supports effective query preparing.

This joined with the way that FastBit is transparently accessible persuades the utilization of WAH-compacted bitmap lists.

An inverted index comprises of a dictionary of the distinct values of the attribute, with pointers to inverted lists that reference tuples with the given value through tuple identifiers (TIDs). To decrease both space use and the I/O prerequisites in query processing, the inverted lists are frequently compacted [14].

IV. Experimental Results

There is training data of organization containing Approximately 15 Lakhs pages which need to be tagged by using Abbyy Fine Reader. The technique is applied in a sequence of reading page image, analyze page image and preprocess page images to recognize page contents correctly.Initially, we considered 1181 scanned pdf documents i.e Handwritten, Typed and Printed for tagging by considering some fixed keyword using Abbyy Fine Reader and Omnipage Ultimate and obtain the following result.

Table III- Result Analysis of Tagging Softwares

Tools Used	Nuance Omnipage Ultimate	Abbyy Fine Reader
Keyword Recognition Rate (Out of 200 tagged keyword)	150	190
Accuracy of Searching (Manually and Automatically) (%)	75%	95%
Size Before Tagging	349 MB (1181 Pages)	349 MB (1181 Pages)
Size After Tagging	55 MB	60 MB
Time Required To Load Pages	More Than 2 Hours	Less Than 1 Minute
Time to Perform OCR (Automatically)	4 Hours	3 Hours

Table IV- Accuracy of Correctly recognized Pages for Automated and Manual tagging

Criteria	No of Tested Pages (Printed and Handwritten)	Accuracy of Correctly Recognized Pages (%)	
		Nuance Omni-page Ultimate	Abby Fine-reader
Automated Tagging	**100**	**80%**	**85%**
Manual Tagging		**85%**	**95%**

The dataset we considered tagged pages for indexing considering some fixed keyword. Here we consider 20 keywords in a single page for searching after indexing in Nuance PDF Converter Professional and Adobe Reader. The result is presented for four cases having 20 keywords per page and obtain the following result.

Table V- Result Analysis of Indexing Softwares

Software Used	Nuance PDF Converter Professional	Adobe Reader	Retrieval Accuracy of Nuance PDF Converter Professional (%)	Retrieval Accuracy of Adobe Reader (%)
Keyword Searching Rate For Page 1	16 Keywords	19 Keywords	80%	95%
Keyword Searching Rate For Page 2	16 Keywords	18 Keywords	80%	90%
Keyword Searching Rate For Page 3	17 Keywords	19 Keywords	85%	95%
Keyword Searching Rate For Page 4	14 Keywords	16 Keywords	70%	80%
Average Retrieval Accuracy (%)			78.75%	90%

For testing purpose of tagging using Asprise OCR different cases are considered-

Case 1- Keyword recognition rate for plain text page with 300 dpi resolution and plain text page with 150 dpi resolution.
Case 2- Keyword recognition rate from Tabular data document.
Case 3- Barcode Recognition rate from document.

Following Table V shows experimental results of Tagging using Asprise OCR by considering various criteria.

Following Table VI shows experimental results of Indexing using Hoot for effective retrieval.

Table VI-Experimental Results of Tagging

Cases	Criteria	Nature of Page	Asprise OCR SDK
Case-1	Keyword Recognition Rate	Plain Text Page with 300 dpi resolution	356/358
	Accuracy of Recognized Keyword (%)		99%
	Size Before Tagging		1.12 MB
	Size After Tagging		2kb (rtf file)
	Time Required To Recognize 358 Keywords		10 Sec
	Keyword Recognition Rate	Plain Text Page with 150 dpi resolution	153/358
	Accuracy of Recognized Keyword (%)		59.60%
Case-2	Recognition of 2 rows and 3 columns with Keywords	Tabular data document.	All keywords are recognized with layout.
Case-3	Barcode Recognition rate	Barcode Data Document	Correctly Recognized contents

Table VII Experimental Results of Indexing

Software Used	Hoot	Retrieval Accuracy of Hoot (%)
Keyword Searching Rate For Page 1	19 Keywords	95%
Keyword Searching Rate For Page 2	19 Keywords	95%
Keyword Searching Rate For Page 3	17 Keywords	85%
Keyword Searching Rate For Page 4	18 Keywords	90%
Average Retrieval Accuracy (%)		91.25%

V. CONCLUSION

Maximum features are available in Abbyy Finereader than OmniPage Ultimate. Working performance of Abbyy is better than Omnipage because of an easy interface. Keyword Recognition rate of Handwritten and printed document is better in Abbyy Finereader. The accuracy of correctly searched documents by Abbyy FineReader is 95%.Keyword Recognition rate is better of Printed or Typed Document if Automated Tagging is used. Keyword Recognition rate is better for Handwritten Document if Manual Tagging is used. (Testing is done using an evaluation copy of OmniPage and Abbyy Finereader.)

In this paper implementation of tagging of a document is based on Asprise OCR SDK. Keyword Recognition Rate from plain text page having 358 Keywords with good Resolution is greater than keyword recognition rate from plain text page having 358 Keywords with low Resolution. The accuracy of Recognized Keyword from Plain Text Page with 300 dpi resolution is 99% in Asprise OCR SDK. Plain Text Page with 150 dpi resolution i.e. low resolution is 59.60%. Content formats or Text layouts on input documents are saved in asprise OCR SDK. Adjacent to Characters (letters and numbers), Asprise OCR can perceive each sort of barcode.

This paper also proposes an effective information retrieval indexing method using the evaluation copy of Adobe Reader. After indexing the documents, retrieval is much faster than simple tagged documents. Though PDF documents tagged with identifying keywords to enhance searchabilty, indexing gives a better solution to retrieve information efficiently. It reduces a time for information retrieval. Thus indexing needs to be carried out after tagging of documents to increase retrieval accuracy.

The work extensively has tested the same using Nuance PDF Converter Professional also. Retrieval accuracy is also more in Adobe Reader. The accuracy of correctly retrieved documents by Adobe Reader is 90%.
Hoot is a full-text indexing engine toolkit written in VB.net, multi-user support access, rapidly visit indexing time. This paper provides detail analysis of Hoot, indexing and searching. The accuracy of correctly retrieved documents using Hoot is 91.25%. The experimental results show that Hoot is efficient for information retrieval.

ACKNOWLEDGMENT

We would like to acknowledge the support and input from our Project Guides Dr. Vijaya B Musande for their constant encouragement and guidance during completion of work.

REFERENCES

[1] Sandra-Dinora ORANTES-JIMÉNEZ, Alejandro ZAVALA GALINDO, Graciela VÁZQUEZ-ÁLVAREZ, Paperless Office: a new proposal for organizations, ISSN: 1690-4524 SYSTEMICS, CYBERNETICS AND INFORMATICS VOLUME 13 - NUMBER 3 - YEAR 2015

[2] Hang Thu Pho, Torben Tambo, "Integrated Management Systems and Workflow-Based Electronic Document Management: An Empirical Study" , Journal of Industrial Engineering and Management, pp. 194-217, January 2014.

[3] M Swamy Das,Ram Mohan Rao Kovvur, "Evaluation of Neural Based Feature Extraction Methods for Printed Telugu OCR System", Advances in Computer Science and Information Technology (ACSIT) ,Volume 2, Number 11,pp. 85-90,April- June, 2015.

[4] Y. C. Li and H. F. Ding, "Research and Application of Full-Text Search Engine Based on Lucene," Computer Technology and Development, Vol. 20, No. 2, 2010, pp.4-56.

[5] FreeOCR Url [Online].Available: http://www.paperfile.net/

[6] Asprise OCR Url [Online]
Available: https://en.wikipedia.org/wiki/Asprise_OCR

[9] The Abbyy Finereader Website. [Online].
Available:https://www.abbyy.com/finereader/

[10] The Nuance Omnipage Ultimate Website. [Online]. Available:
https://www.nuance.com.

[11] The The Omnipage Reader website. [Online]. Available
http://www.nuance.com/

[12] The Adobe Reader website. [Online]. Available:
http://www.adobe.com/

[13] Jennifer Pearson,"Supporting Effective User Navigation in Digital Documents", CHI-2010,Atlanta,GA,USA,April 10- 15,2010.

[14] Truls A. Bjorklund and Nils Grimsmo, Johannes Gehrke , Oystein Torbjornsen, Inverted Indexes vs. Bitmap Indexes in Decision Support Systems, ACM, CIKM'09, November 2–6, 2009, Hong Kong, China.

[15] K. Wu, E. J. Otoo, and A. Shoshani. Optimizing bitmap indices with effcient compression, ACM Trans. Database Syst., 2006.