DESIGN AND IMPLEMENTATION OF STRUCTURED DOCUMENT RETRIEVAL SYSTEM

BY

UDIE CYNTHIA
ICT/2252051262

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE, SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY, AUCHI POLYTECHNIC, AUCHI
EDO STATE

IN PATRIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF HIGHER NATIONAL DIPLOMA (HND) IN COMPUTER SCIENCE, SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY.

NOVEMBER, 2022

CERTIFICATION

We, the undersigned hereby certify that this research work titled **“DESIGN AND IMPLEMENTATION OF STRUCTURED DOCUMENT RETRIEVAL SYSTEM, A CASE STUDY OF COMPUTER SCIENCE DEPARTMENT AUCHI POLYTECHNIC, AUCHI** was carried out by **UDIE CYNTHIA** with Matriculation **ICT/225205126** with of the department of Computer Science.

I further certify that this research project is adequate in scope and quality and meet partial fulfillment of the requirement for the Award of National Diploma (HND) in Computer Science Auchi Polytechnic Auchi, Edo State.

____________________ ________________

UDUIGUOMEN, U. C **Date**
(Project Supervisor)

______________________________ ________________

AKHETUAMEN, S. O. **Date**
(Head of Department)

DEDICATION

The research is dedicated to God Almighty, the keeper of my soul for granting me the courage and ability to make this work a success.

ACKNOWLEDGEMENT

In understanding this study, have inevitably become indebted to more people that I can mention here.

Be that as it may, my profound gratitude goes to my honorable Project **UDUIGUOMEN, U. C** who despite his scheduled painstaking effort to go through this work. I own him much appreciation for his enormous assistance right from shaping the topic to conclusion. Thanks also goes to my Head of Department **AKHETUAMEN, S.O** for his assistance during the course of this project.

Special appreciation goes to my love parents **MR & MRS UDIE** for their support, morally, spiritually and financial. May God richly bless you. Amen.

TABLE OF CONTENTS

ABSTRACT

Information retrieval (I.R) is a very important field in the Natural Language Processing (NLP). It facilitate in a reasonable time, the access to specific information in a set of documents. Several works were elaborated in order to implement a perfect information retrieval system. Information management has become a major strategic factor in companies' development. Digital document management system can help an organization succeed by; Saving time, Saving money, Increasing efficiency, Increasing productivity, Increasing inter-departmental and inter-organizational communication, and Enabling automation. It has become widely recognized that manual paper-based processing has inherent problems that disadvantage companies in critical ways, undermining productivity and impeding the flow of information. Ignoring the problem is proving less and less an option in the current economic and competitive climate. Looking at the Department of Computer Science, Auchi Polytechnic, Auchi as a case study,

CHAPTER ONE

1.0 INTRODUCTION

1.1 BACKGROUND TO THE STUDY

Information retrieval (I.R) is a very important field in the Natural Language Processing (NLP). It facilitate in a reasonable time, the access to specific information in a set of documents. Several works were elaborated in order to implement a perfect information retrieval system. The majority of these works, in particular for the Arabic language, are confronted with problems due to the establishment of a correspondence between required information and all documents of a collection. Among these problems, there are the different formulations of the same concept: A relevant document can contain terms partly semantically close to those of the request and partly different (synonyms, hyperonyme, term having a different morphological form, terminology, etc). This phenomenon causes a decrease of the recall of these systems which are unable to propose some interesting documents to the user. This problem comes to join that of the words polysemia. A hypernym is a word whose meaning includes the meaning of one or more other generic word. For example the flower is the hypernym of pink. This is the origin of a decline of accuracy of the systems since it involves potentially the recovery of non relevant documents (Tazzite, *et al.*, 2008).

For that purpose, we used terminological dictionaries for every keyword. The results obtained are very important on the level of the precision and of the recall. The only disadvantage is that this approach requires a prolonged time for its execution. In order to address this problem, we propose in this article another approach which takes into account the introduction of semantics at the indexing phase (Tazzite, *et al.*, 2008).

Information management has become a major strategic factor in companies' development. It is important to get the right information circulated to the right people, as efficiently as possible, yet still keep it secure. Document management provides a way for companies to organize their information, in all its forms, in one place. Streamlining business processes and increasing efficiency are fundamental concerns for any organization regardless of size or sector. In today’s ever increasingly strict regulating environment, compliantly managing documents and records of all types takes significant time and money that could better be spent on achieving mission critical objectives. By implementing a document management system, organizations can realize many benefits that noticeably improve organizational efficiency.

Digital document management system can help an organization succeed by; Saving time, Saving money, Increasing efficiency, Increasing productivity, Increasing inter-departmental and inter-organizational communication, and Enabling automation. A Document Management System (DMS) is a system based on computer programs used to store and access documents.

Electronic document management solutions are designed to organize business files and records digitally, whether they started out in paper form or were generated by software applications. Paper files are first converted to electronic format by scanning. This provides a more compact means of storage, universal access for retrieval, and higher levels of data security and privacy.

A company-wide document management system also controls digital files that are generated directly through applications -- such as those in the Microsoft Office suite (Word, Excel and PowerPoint), accounting software, CAD, email, and so on. Managing (rather than simply storing) documents enables quicker access to, and greater command over, business information.

1.2 STATEMENT OF THE PROBLEM

It has become widely recognized that manual paper-based processing has inherent problems that disadvantage companies in critical ways, undermining productivity and impeding the flow of information. Ignoring the problem is proving less and less an option in the current economic and competitive climate. Looking at the Department of Computer Science, Auchi Polytechnic, Auchi as a case study, students submit hard copies of document as a means of making requests, Memos are sent to the department in paper form which degenerates to bulk of files in the offices. This system of storing documents is tedious, complicated and time consuming.

Faced with the need to organize documents, the proposed system for management of documents is unique and totally innovative in its integrated approach. Its functionality of making documents available anytime, anywhere and enabling easy access, retrieval and storage of documents makes it called for.

The system to be developed makes use of rich internet technology to replace desktop application with web application running on a remote server. The system shares the advantage of both web application and desktop application, and removes the most disadvantages of both.

1.3 AIM AND OBJECTIVES OF THE STUDY

The aim of the project is to develop a document management system for the Department of Computer Science, Auchi Polytechnic, Auchi that is able to deliver access to anyone authorized anytime, anyplace, and on any device. The objectives of the study are to develop a system that should be able to;

1. Store documents properly
2. Archive and retrieve documents properly and efficiently
3. Ensure document security and availability.

1.4 SIGNIFICANCE OF THE STUDY

The proposed system will offer the following advantages to the Department of Computer Science, Auchi Polytechnic, Auchi;

1. Reduced Storage: The cost of commercial property and the need to store documentation for e.g. retrieval, regulatory compliance means that paper based document storage competes with people for space within an organization. Scanning documents and integrating them into a document management system can greatly reduce the amount of prime storage space required by paper. It also allows any documents that still have to be stored as paper to be stored in less expensive locations.

2. Flexible Indexing: Indexing paper in more than one way can be done, but it is awkward, costly and time-consuming. Images of documents stored within a document management system can be indexed in several different ways simultaneously.

3. Improved, faster and more flexible search: Document Management Systems can retrieve files by any word or phrase in the document - known as full text search - a capability that is impossible with paper.

4. Controlled and Improved Document distribution: Imaging makes it easy to share documents electronically with colleagues and clients over a network, by email or via the Web in a controlled manner. Paper documents usually require photocopying to be shared. This provides a cost saving by reducing the overheads associated with paper based document distribution, such as printing and postage and removes the typical delay associated with providing hard copy information.

5. Improved Security: A document management system can provide better, more flexible control over sensitive documents. Many document management system solutions allow access to documents to be controlled at the folder and/or document level for different groups and individuals. Paper documents stored in a traditional filing cabinet or filing room does not have the same level of security i.e. if you have access to the cabinet you have access to all items in it. A document management system also provides an audit trail of who viewed an item, when or who modified an item and when, which is difficult to maintain with paper based systems.

6. Disaster Recovery: A document management system provides an easy way to back-up documents for offsite storage and disaster recovery providing failsafe archives and an effective disaster recovery strategy. Paper is a bulky and expensive way to back-up records and is vulnerable to fire, flood, vandalism and theft.

7. No Lost Files: Lost documents can be expensive and time-consuming to replace. Within a Document Management System, imaged documents remain centrally stored when being viewed, so none are lost or misplaced. New documents are less likely to be incorrectly filed and even if incorrectly stored can be quickly and easily found and moved via the full-text searching mechanisms.

8. Digital Archiving: Keeping archival versions of documents in a document management system helps protect paper documents that still have to be retained, from over-handling.

1.5 SCOPE OF THE STUDY

The main challenge of document management is flexible storage and retrieval. This study's intended user is the Department of Computer Science, Auchi Polytechnic, Auchi. The system will be provided with facilities for easy storage, retrieval and security of documents.

1.6 LIMITATION OF THE STUDY

- Inadequate books and other reference material was a limitation of this study as a polytechnic library, which do not have most of the relevance book and several material needed for data collection.

- **Time factor:** A large detailed investigation could not be conducted within a period of time.

- **Financial Problem:** This is one of the outstanding problem that attracts this researcher was untenable to go out fetch more information that could have been relevant in this research work.

1.6 DEFINITION OF TERMS

a) **Document:** A document is a form of information. It could be electronic or in the form of paper.

b) **Management System:** A documented and step by step method aimed at smooth functioning through standard processes.

c) **Retrieval:** The process of accessing information from memory or other storage devices.

d) **Web Application:** An application program stored on a remote server and delivered over the internet through a browser interface.

e) **Web browser:** A software application used to locate, retrieve and display contents on a World Wide Web, including web pages, video and other files.

f) **Server:** A running instance of an application capable of accepting requests from the client and giving responses accordingly.

g) **Database:** A computerized record-keeping system. It is a repository for storing information.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 BRIEF HISTORY ON DOCUMENT RETRIEVAL

Document Retrieval first emerged as a field of both inquiry and application in the late 1940’s and early 1950’s at a time when the amount of scientific literature being published was growing at such a tremendous rate that it raised concerns as to how scientists would be able to stay informed of new developments as they were being reported in the scientific literature. Researchers, mainly computer scientists and library scientists, began to consider how computers might be programmed to accomplish some of the laborious human tasks of representing and retrieving relevant sources that were previously considered within the purview of Librarians and Library Science.

Initially, Document Retrieval was actually citation retrieval, that is, a search against bibliographic fields such as title, author, source, and subject-based keywords from a controlled vocabulary that were humanly assigned to documents. The documents themselves were neither actually stored nor accessible by computer. The appropriate analogy is closer to an automated library card catalog search than a full text search of the contents of actual documents. The results from a Document Retrieval search were simply citations that led the user to a hard copy document, book, or report. Over time, advances in the theoretical models underpinning Document Retrieval and decreases in the cost of computer storage permitted the full text of documents to be ingested and searched, not just a limited set of manually assigned bibliographic access points. While the increase in scientific literature may have been the practical motivator for the field of Document Retrieval, Vannevar Bush’s article in the Atlantic Monthly entitled “As We May Think” is considered by most historians of science as the intellectual forbearer of the field (Bush, 1945). In this piece, Bush presaged much of what we today consider to be the crux of the field, when he called for a solution for improved access to information that he referred to as "memex" a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility an enlarged supplement to his memory." As all other computer-automated applications at that time, Document Retrieval initially made use of punch cards which enabled the post-coordinate revolution – that is, rather than documents being assigned pre-coordinated descriptors, such as automatic information retrieval as would be typical in a card catalog or a printed index, post-coordinate indexing enabled single word descriptors to be assigned one to a card, such as automatic on one card describing the document, information on another, and retrieval on yet another card and then for the user to combine search terms as they best saw fit to describe their information need at search time (Webler, 2022).

The increasing availability of machine readable texts led to rapid, widespread growth in the usage of Document Retrieval systems as the collection against which users could search increased dramatically. Not surprisingly, this also led to the flourishing of the early commercial Document Retrieval systems such as Dialog and LexisNexis. These systems were sufficiently complex and non-intuitive that end users needed to have trained search intermediaries do their searches, because these intermediaries could understand the intricacies of the data records in the Document Retrieval System and were well- trained in constructing queries in Boolean logic. These early systems were not based on free text searching, but rather, required the intermediary to know the exact wording and syntax to use in searching, both in proper names and subject based descriptions – referred to respectively as authority files and controlled vocabulary. While the commercial realizations of Document Retrieval were seeing success in improved provision of citations or actual documents to researchers in the scientific disciplines, researchers in the growing field of Information Science were actively moving on to more complex models of retrieval and more natural representations of documents on which users themselves might search – namely the text as it occurred naturally in the documents. Of course this then obviated the need for intermediaries, lowered the cost of entry to the field, and opened the modes of access. This was a trend that began in the early days of Document Retrieval, but which really only came to full fruition with the tremendous growth in open availability of information on the Web. Interestingly, the Document Retrieval systems, with full-text searching, relative weighting of terms, and ranking of results, which were being developed and tested with increasing rigor by academic researchers, were not adopted by commercial providers of Document Retrieval services, due to these organizations' sizeable investments in Boolean-based retrieval (Webler, 2022).

2.2 INFORMATION

Information is an abstract concept that refers to that which has the power to inform. At the most fundamental level information pertains to the interpretation of that which may be sensed. Any natural process that is not completely random, and any observable pattern in any medium can be said to convey some amount of information. Whereas digital signals and other data use discrete signs to convey information, other phenomena and artifacts such as analog signals, poems, pictures, music or other sounds, and currents convey information in a more continuous form.[1] Information is not knowledge itself, but the meaning that may be derived from a representation through interpretation (Hubert, 2005).

According to Hubert (2005), information is often processed iteratively: Data available at one step are processed into information to be interpreted and processed at the next step. For example, in written text each symbol or letter conveys information relevant to the word it is part of, each word conveys information relevant to the phrase it is part of, each phrase conveys information relevant to the sentence it is part of, and so on until at the final step information is interpreted and becomes knowledge in a given domain. In a digital signal bits may be interpreted into the symbols, letters, numbers, or structures that convey the information available at the next level up. The key characteristic of information is that it is subject to interpretation and processing.

The concept of *information* is relevant in various contexts, including those of constraint, communication, control, data, form, education, knowledge, meaning, understanding, mental stimuli, pattern, perception, proposition, representation, and entropy (Luciano, 2010). The derivation of information from a signal or message may be thought of as the resolution of ambiguity or uncertainty that arises during the interpretation of patterns within the signal or message (Webler, 2022).

To the growing number of academic computer science departments and research labs that were becoming interested in Document Retrieval, the Cranfield experiments that ran from 1957 to 1967, (Cleverdon, 1967) were a major milestone as they established the basic evaluation paradigm, including metrics and experimental procedures which enabled true scientific experimentation to be conducted and the field to be recognized as a science. The test collection paradigm of gathering a stable collection of documents of interest to an identifiable user group, as well as possible questions from these users that might be expected to be answerable by documents in that collection, and relevance assessments on the retrieved documents (or the whole collection if possible) is still used today. The accepted metrics used in Document Retrieval experimentation and evaluation were also established, namely, Recall – the percent of known relevant documents in the collection that are retrieved; and Precision - the percent of retrieved documents that are relevant. Based on the Cranfield test collections and the new evaluation metrics, an era of great research activity and development ensued. However, unlike Cranfield's work which evaluated automatic searching of manually assigned index terms, the ensuing work utilized these same collections for experimentation on automatic indexing of the natural language of document abstracts first, and later, of the full text of the documents. The various models of Document Retrieval which are so well-known and experimented with today were introduced, including the vector space model of Salton and his group of researchers at Cornell (Salton, 1968); the probabilistic model (Robertson & Sparck Jones, 1976); and the inference model (Turtle & Croft, 1990). The most widely known and used of the experimental systems was SMART, developed by Salton (1971). It was used for extensive empirical runs which showed the value of the simple natural language processing techniques of stemming, deletion of stop words, phrase-based indexing, as well as many experiments investigating the most appropriate term-weighting formula.

Document Retrieval systems with empirically-tested ranking and weighting algorithms slowly spread into the operational world as increasing numbers of experimental results demonstrated that automatic indexing of the natural language of documents was as good as manually controlled vocabulary indexing. The range of retrieval models that were utilized commercially increased due to availability of increased computing power and lower cost of storage of indexes based on the full text of documents (Webler, 2022).

Large scale comparative testing of Document Retrieval systems began with the first of the Text REtrieval Conferences (TREC) at the National Institute of Standards and Technology in 1992 (Harman, 1993) and has continued and expanded each year. The goal of these annual events is to bring together researchers from around the world who test their systems on a common test collection of queries and documents and then share the details of their systems with the other participants. Twenty-five research groups participated in the first evaluation using a large, new test collection, a set of questions generated by a likely group of users, and their relevance assessments on the documents retrieved by the participating systems. Most member of the Document Retrieval community believe that TREC has been one of the most positive influences on both scientific advances and expansion of interest in Document Retrieval.

However, most would also agree that it is the Web that has made the field as prominent and exciting as it is today. Web searching is now omnipresent in most individuals' lives and little thought is given to it as being the original Document Retrieval that began back in the 1940s. And while there are some real differences between earlier methods and practices of Document Retrieval from pre-established databases as compared to the very dynamic and less controlled world of Web searching, the basics are the same, with the addition of link analysis, which utilizes a network structure of links among web pages as an additional source of information (in addition to document content) in retrieving and ranking potentially relevant documents or pages.

2.3 DOCUMENT

A document is a written, drawn, presented, or memorialized representation of thought, often the manifestation of non-fictional, as well as fictional, content. In the past, the word was usually used to denote written proof useful as evidence of a truth or fact. In the computer age, "document" usually denotes a primarily textual computer file, including its structure and format, e.g. fonts, colors, and images. Contemporarily, "document" is not defined by its transmission medium, e.g., paper, given the existence of electronic documents. "Documentation" is distinct because it has more denotations than "document". Documents are also distinguished from "realia", which are three-dimensional objects that would otherwise satisfy the definition of "document" because they memorialize or represent thought; documents are considered more as 2-dimensional representations. While documents can have large varieties of customization, all documents can be shared freely and have the right to do so, creativity can be represented by documents, also. History, events, examples, opinions, etc. all can be expressed in documents.

2.4 ELECTRONIC DOCUMENT

An electronic document is any electronic media content (other than computer programs or system files) that is intended to be used in either an electronic form or as printed output. Originally, any computer data were considered as something internal — the final data output was always on paper. However, the development of computer networks has made it so that in most cases it is much more convenient to distribute electronic documents than printed ones. The improvements in electronic visual display technologies made it possible to view documents on screen instead of printing them (thus saving paper and the space required to store the printed copies).

However, using electronic documents for final presentation instead of paper has created the problem of multiple incompatible file formats. Even plain text computer files are not free from this problem — e.g. under MS-DOS, most programs could not work correctly with UNIX-style text files (see newline), and for non-English speakers, the different code pages always have been a source of trouble.

Even more problems are connected with complex file formats of various word processors, spreadsheets, and graphics software. To alleviate the problem, many software companies distribute free file viewers for their proprietary file formats (one example is Adobe's Acrobat Reader). The other solution is the development of standardized non-proprietary file formats (such as HTML and OpenDocument), and electronic documents for specialized uses have specialized formats – the specialized electronic articles in physics use TeX or PostScript.

2.4.1 *Features of Electronic Document*

Following our post about finding the right electronic document management partner to work with, we're now turning our attention to what we believe are the most useful features within EDM. To truly take advantage of digitising the way your business accesses, secures and archives documents, you need a solution that comes with certain features as standard. Here, we list the ones that we know are most essential for our 41,000 users. The tools, functions and automations that really do save them time, money and effort.

2.4.2 Easy Accessibility

The modern work environment is one in which we access information in a multitude of ways. Whether it's from a mobile device on our way into the office, from a tablet after a client meeting or on our desktops at home or in the office. A system that only allows you to upload and access documentation when you're logged into a particular device at a specific network location, doesn't reflect today's working practices and won't be met with enthusiasm by potential users. Instead, look for an anywhere, anytime, any device solution.

2.4.3 Instant Search and Retrieval

Many of our customers have come to us because they are fed up of continually searching for documents and keeping complex filing systems organised. If that's a challenge you recognise, then it's essential that you find an EDM solution that comes with a comprehensive search facility. The easier to use the better. We all use search engines for locating what we need on the internet, and as such, you should expect a search function in your chosen EDM system that it just as easy and thorough to use.

One that can search for what you need even if you only know fragments of the document name, or a keyword such as a client name or address that is used in the document. Thumbnail views of the retrievable documents should also be expected –

as a quick visual of the front page can often be all that is needed to help you find the exact document you need.

2.4.4 Open Integration

Similar to the accessibility point above, a standalone system that doesn’t integrate with the other business essential tools you have (from email to practice management) will create more work, as colleagues are asked to log in and upload files to yet another system. Make sure your chosen system integrates and automates workflow with the other tools that the potential users in your business use, to offer them clear benefits. For example, automatically saving all relevant emails into a client’s folder, so all relevant staff can see communications – even if the original sender is on leave or no longer with the business.

2.4.5 Optimum Security

The benefits of easy accessibility and open integration do not need to come at the cost of security. In order to make sure that you are protecting yours and your client’s data, all data needs to be tightly encrypted as it passes through any solution that you choose. We recommend using extended validation SSL certification provided by VeriSign and encrytping files using AES-256. Check that your chosen system adheres to International Standard, ISO/TR 22957:2009.

2.4.6 Audit trails and Traceability

EDM really comes into its own, when it comes to making business processes more efficient. For example if you have to share documents with a client for approval before sending them elsewhere in the business (or externally), and need to keep an audit trail of where documents are, a good EDM solution will automate the whole process for you. To facilitate this even further, we recommend intelligent portal technology. By this we mean a secure site, accessible to invited parties only, where documents can be shared for electronic signatures and approval on the go. If a document has to go to several parties for signing, it can be done remotely, rather than waiting for busy schedules to allow for everyone to be in the same place at the same time.

2.5 DOCUMENT RETRIEVAL

Document retrieval is defined as the matching of some stated user query against a set of free-text records. These records could be any type of mainly unstructured text, such as newspaper articles, real estate records or paragraphs in a manual. User queries can range from multi-sentence full descriptions of an information need to a few words (Lin and Wilbur, 2007). Document retrieval is sometimes referred to as, or as a branch of, text retrieval. Text retrieval is a branch of information retrieval where the information is stored primarily in the form of text. Text databases became decentralized thanks to the personal computer. Text retrieval is a critical area of study today, since it is the fundamental basis of all internet search engines (Kim, *et al.,* 2011).

Retrieval of information by subjects from huge mass of documents requires that essential concepts are identified and organised in a searchable form. Indexing is a mechanism by which information contained in documents can be organised. But the problems lie with identifying and organising the concepts. In the documentary information, authors communicate in natural languages which are characterized by linguistic features. To overcome the problems of natural language, the need for an artificial language or indexing languages arises. It means that an indexing language is a language used for subject classification or indexing of documents. An Indexing language is defined as the set of terms used in an index to represent topics or features of documents, and the rules for combining or using those terms. The purpose of an indexing language is to express the concepts of documents in an artificial language so that users are able to get the required information (Kim, *et al.,* 2011).

The indexing language does this by depicting the relationships among the differently related concepts. There are three main types of indexing languages. **1. Natural indexing language -** Any term from the document in question can be used to describe the document.

2. Free indexing language - Any term (not only from the document) can be used to describe the document.

3. Controlled indexing language - Only approved terms can be used by the indexer to describe the document. In the following sections, you will be introduced, in brief, to natural, free and controlled indexing languages.

2.6 BASIC DOCUMENT RETRIEVAL SYSTEM

In Document Retrieval, some processes take place dynamically when the user inputs their query, while other processes take place off-line in advance and in batch mode and do not involve individual users. These static processes are run on the documents that will be made available in the retrieval system. These will be explained first. Then, the two dynamic processes, Query Processing and Matching, will be presented. Figure 1 provides a simple, but clear view of the relationship between these three processes (Kim, *et al.*, 2011).

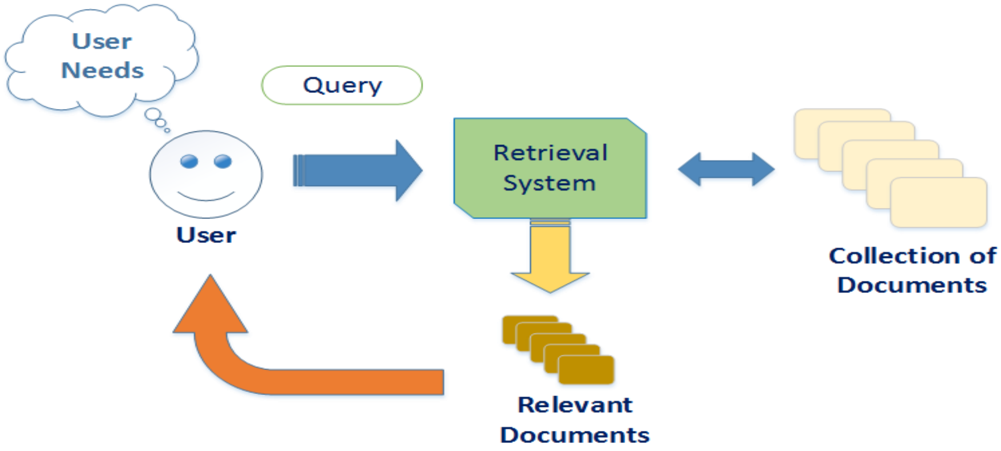


Fig 1: Basic Components of a Document Retrieval System (Kim, *et al.*, 2011).

Document Processing: The first two steps in the processing of documents are somewhat mundane, but necessary, and can be considered as batch pre-processing (Webler, 2022) These are:

1. Normalize document stream to a predefined format, whereby multiple external formats (e.g. newsfeeds, web pages, word processed documents) are standardized into a single consistent format. This is an essential step (much akin to data clean-up in data mining) as all downstream processes rely on receiving a common format they can recognize and process. Preprocessing is particularly vital for systems with more complex processing than simple 'characters between white spaces' indexing (Kim, *et al.*, 2011).
2. Break document stream into desired retrievable units, whether this is web page, chapter, full document, paragraph, etc. The pointers stored in the inverted file are to whatever unit size has been pre-determined. Therefore, document retrieval could in fact be paragraph retrieval, if the indexable unit was determined at this stage to be the paragraph. From this step forward, the system is performing the heart of the document indexing process.
3. Identify potential indexable elements in documents. This is a key decision point that dramatically affects the nature and quality of the retrieval performance. First, the important definition needs to be made as to what is a term. Is it any string of alpha- numeric characters between blank spaces or punctuation? If so, are non- compositional phrases or multi-word proper names, or inter-word symbols such as hyphens or apostrophes treated differently (e.g. are "small business men" and "small- business men" the same)? At this stage, the system requires a set of rules to be executed which control what actions are taken by the 'tokenizer' – the algorithm which recognizes 'indexable terms'. IR systems vary as to which of these processes they perform, but the most frequently used processes are:

a. Delete stop words via an algorithm that filters the document’s potential indexable elements against a Stop Word list to eliminate terms that are deemed to be insignificant in determining a document’s relevance to a user’s request. The original objective in using stop words was to save system resources by eliminating those terms that have little value for retrieval performance. Although these terms may comprise up to 40% of the tokens in a document set, index size is of far less importance today due to cheap memory, but their omnipresence renders them of little value to retrieval. The typical word classes that are marked as stop words include the function word classes and a few more (i.e. articles, conjunctions, interjections, prepositions, pronouns, and ‘to be’ verb forms).

b. Stem terms by removing suffixes. In this morphological step, some IR systems do just inflectional (‘weak’) stemming which only changes the subclass within a part-of-speech category, i.e. past tense to present tense, while others also do derivational (‘strong’) stemming which removes suffixes, sometimes recursively, that may actually change the part of speech of a word. Use of stemming will result in fewer entries in an index, each of which is likely to have higher frequency counts than if all morphological variants and their counts are used. The initial goal of stemming was to reduce the storage requirements of the inverted index file by reducing the number of unique words, but stemming has remained in use even today when storage is not an issue, because it improves recall of relevant documents. For example, if a query includes analyze, the user may well want documents which contain analysis, analyzing, analyzer, or analyzed. In order for the system to match on all these variants, it must stem both the query and the document terms toanaly-. Obviously, stemming may negatively impact precision (Webler, 2022).

c. Bracket noun phrases, usually by means of regular expressions which define the part-of-speech patterns which comprise a noun phrase (e.g. <ADJ NN> or <NN NN>). This is a step that can negatively affect recall of retrieval results by either excluding documents when the phrasal expression in the query is not exactly the same as the index entry of a document, or positively affect precision by retrieving only documents that include the terms in the desired phrasal expression.

4. Produce an inverted file containing a sorted array of all indexable terms (with term defined as referring to either a word or a phrase), along with the unique identification number of each document in the collection in which the term occurs, a link to each of these documents, weights for each term as determined by the IR model being implemented in the system [which will be described in the next section] and optionally, the within-document location of the term. More sophisticated systems may include further information in the inverted file, such as named entity category for Proper Names (i.e. PERSON, ORGANIZATION, GEO-LOCATION, etc) but the most common features are simply term, document ID, and weight. Query Processing: The system's internal representation of the user's question / search terms is typically referred to as the query. Most of the same processes that are run on the documents are also run to produce the query, but there are some unique processes as well. As distinct from document processing, all of the query processing is done in real time,

   while the user awaits their documents. These are:

a) Recognize query terms vs. special operators, such as "I need information about..." which do not convey the topic of the user's information need and will not be including the query representation.

b) **Tokenize query terms:** A process that requires similar decisions as were described on the document processing side – that is stop word deletion, stemming, and phrase recognition.

c) Create query representation, which typically follows stop word removal and stemming, and which may also include insertion of logical operators between / amongst terms requiring co-occurrence or simple presence of only one of the arguments.

d) Expand query terms to include variant terms that refer to or relate to the same concept. These may be synonymous terms that are found in an electronic thesaurus such as WordNet (Fellbaum, 1998) or terms that are highly associated with the query term, based on co-occurrence statistics preferably computed on the same or a similar document collection as the one on which the search is being conducted.

Query expansion relieves the user of needing to generate all conceptual variants of their search terms and is likely to improve recall, but may reduce precision when erroneous senses of the newly introduced terms retrieve irrelevant documents. The longer a query is, the less likelihood that erroneous senses of expanded terms will have a negative impact, but also the less likely that expansion will contribute much to the retrieval results (Webler, 2022).

e. **Compute query term weights**: This step is less commonly included in Document Retrieval systems, mainly because it is difficult both for users to know how to assign weights to query terms in a way that improves retrieval results, or for automatic weighting, since queries are frequently so short as to give little evidence of the relative importance of query terms as most terms only occur once in a single query. Some NLP-based systems have positive results from automatic determination of the ‘mandatory’ concept in a query which is then assigned a greater weight (Liddy *et al.,* 1995).

a) **Matching of Query to Documents:** Once the query representation is produced, the matching process begins. The process description below may be easier to follow if you conceive of both the query and the documents as vectors of terms, with frequency information or weights for each term in the vector. Search inverted file for documents that contain terms in the query. This is typically done using a standard binary search. Each document that contains any of the query terms becomes a candidate for retrieval.

b) Compute similarity score between query and each candidate document using the algorithm prescribed by one of the four Document Retrieval models being used. This score is referred to as the Similarity Coefficient. The scoring mechanism for each of the major Document Retrieval models will be detailed in the next section.

c) Rank order the documents in decreasing order based on the scores assigned them by the scoring algorithm. This may be either straightforward ranking based on the similarity Coefficient, or the system may utilize automatic relevance feedback whereby the system takes the top N-ranked terms from the top N-ranked documents as they are being shown to the user, and adds these terms to the query representation and reruns the search with the revised query to produce the continuation of the ranked list of relevant documents.

d) Provide list of perceived relevant documents to user ranked by similarity score between query and document. Systems that utilize other sources of evidence of value of a document to the query, such as number of links from the page/document to or from other pages/documents, would integrate this information and produce a potentially different ranked list.

e) Allow for query modification by the user if user-based relevance feedback is provided by the system. If so, typically, the user marks the documents they find relevant, either based on just the title and brief description shown them in the initial list or by actually reviewing the full document, which they can link to from the results page. Perform relevance feedback based on user's input. The algorithm for user-based relevance feedback is typically the same as that for automatic relevance feedback as described in Step 3 above. The system then re-runs the search with the revised, and hopefully improved, query and produces a revised ranked list of documents. The relevance feedback loop is iterative and can be performed as many times as the user wants.

2.7 NATURAL INDEXING LANGUAGE

Natural language refers to our language, which we normally use for communication. Whereas, languages that we design for a specific purpose or use in a specific sense or only for limited use are artificial languages. Natural indexing languages are thus 'natural language' or ordinary language of the document being indexed. Any term that appears in the document is a candidate for index terms. In practice, natural language indexing tends to rely upon the terms present in an abstract or the title of a document. Natural language indexing is based upon the full text of a document, depending on how it is archived. It may lead to very extensive indexing of each document or will involve establishing some mechanism for deciding which terms are the most important in relation to a particular document. In computerized indexing this will involve statistical analysis of the relative frequency of occurrence of terms (Webler, 2022). In human indexing some judgment would be required in selecting the terms. Many of these problems can be minimized by restricting indexing to titles and abstracts. Either, a computer or a person can execute natural language indexing. In computer indexing the computer may well use a list of terms deemed to be useful in indexing (example, a type of thesaurus) to identify appropriate terms. The use of natural language is depicted in a Figure 1.

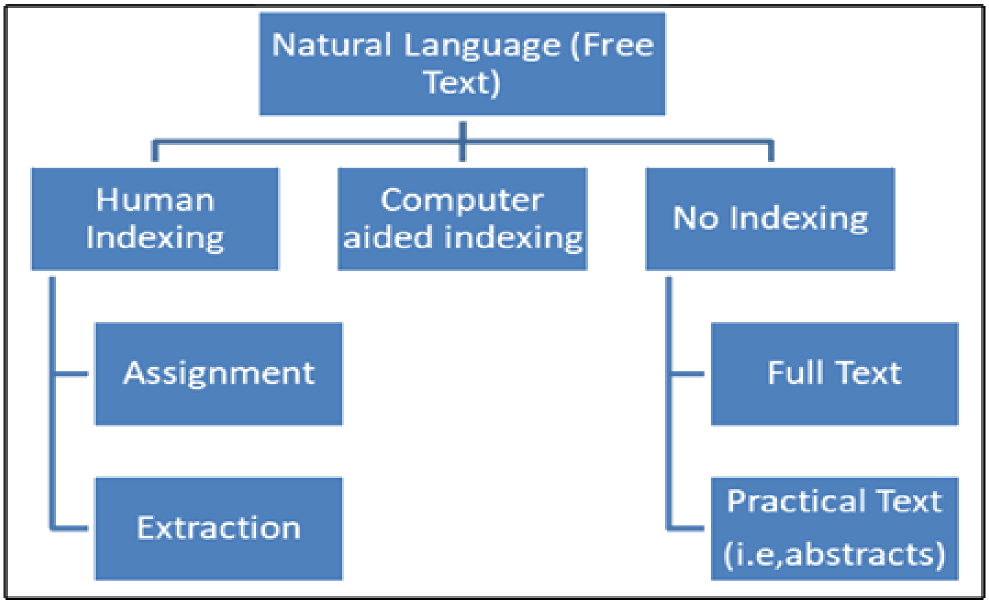


Fig. 2: Natural Indexing Language (Webler, 2022).

Document retrieval systems find information to given criteria by matching text records (*documents*) against user queries, as opposed to expert systems that answer questions by inferring over a logical knowledge database. A document retrieval system consists of a database of documents, a classification algorithm to build a full text index, and a user interface to access the database (Webler, 2022). A document retrieval system has two main tasks:

1. Find relevant documents to user queries
2. Evaluate the matching results and sort them according to relevance, using algorithms such as PageRank. Internet search engines are classical applications of document retrieval. The vast majority of retrieval systems currently in use range from simple Boolean systems through to systems using statistical or natural language processing techniques (Kim, *et al.*, 2011).

Document Retrieval (more commonly referred to as Information Retrieval by researchers in the field) is the computerized process of producing a list of documents that are relevant to an inquirer's request by comparing the user's request to an automatically produced index of the textual content of documents in the system. These documents can then be accessed for use within the same system. Nearly everyone today uses Document Retrieval systems, although they may not refer to them as such, but rather as Web-based search engines, e.g. Google, Yahoo, Alta Vista, etc.

Document Retrieval systems are based on different theoretical models, which determine how matching and ranking are conducted. The most prevalent models are Boolean, Vector Space, Probabilistic, and Language Modeling, each of which is explained below. Within the indexing aspect of each model, the system processes, represents, and weights the substantive content of documents and queries for matching. It is here in feature selection that one might expect to see linguistic theories and models used extensively, however, to date, most systems utilize only the morphological and lexical levels of language, with some notable exceptions where full Natural Language Processing is utilized (Liddy, 1998; Strzalkowski *et al.*, 2002). Following a brief history of the major mileposts and trends in the field of Document Retrieval, the basic components and processes of a generic Document Retrieval System are described in fairly non-technical language in order to provide context for understanding later, more technical distinctions

2.8 CHALLENGES OF DOCUMENT RETRIEVAL SYSTEM

Although the field of document retrieval has made much progress, many problems still exist. Those who provide information or manage it must take these problems into full consideration. Indexing and classification are the most commonly used tools to answer the user's need. Some advanced systems for better retrieval such as Boolean, Vector, and Fuzzy approaches are developed to cope with the problems.

But there is still doubt that these approaches and systems can highly promote the efficiency of the task. To evaluate the retrieval process, recall and precision are the most popular methods known at the present time. But some think that they do not work properly. While uncertainty is a major obstacle on the way to answer the user's need, the efforts of information providers are devoted mostly to the process of Information Technology (IT). Although Information technology is of high importance, it must be used totally to serve needs. Information system (IS) management not only should be regarded in the same way as information technology but we must assign it some priority. That is, if we allocate some money and energy for IT, we must allocate more for IS. It is critical to serve users with least investment in IT in order to get more benefit in information system management.

2.8.1 Recall and Precision

Although there are various methods to evaluate the retrieval process as well as classification activity, recall and precision are highly recommended by the authors (Jiri, 2002). The disputing opinions on this range from recall and precision being nonsense and completely rejected to nearly full acceptance (Arthur, 1975). Regardless, as mentioned before, the satisfaction of the user in the retrieval process is to be shown by relevance. And recall and precision are highly connected to relevance and non-relevance (Masse, 2001) argues that there is no advantage to using recall and precision. One of the major reasons for the inapplicability of recall, he says, is because we do not know the exact number of relevant items in the whole database. So recall which means relevant items retrieved in relation to the whole number of relevant items in the system, actually becomes impossible to calculate and unreachable. Precision, too, is defined as relevant items found in relation to the relevant items found plus the irrelevant items found by the user. Bloomfield argues that non-relevant items found by the system are not really counted as retrieval.

Retrieval, practically, means those relevant items, for which the user is looking. If the system retrieves some items that are not relevant, it is a defect of the system and a wasting of time for the user, and it is not effective retrieval. So here we do not see any advantage for precision except that it is equal to retrieval itself. As Maltby says: Recall depends on many factors including depth and accuracy of indexing, but attempts to achieve greater precision involve the use of controls of various kinds and these often are distinctly classificatory in character (Masse, 2001) state that recall depends on the system's ability to filter out unwanted items. Mention that these two are capable of being measured under controlled conditions, and they are used to express them by ratios (Webler, 2022). They count hits, miss, noise, and dodge for the system; in a good system one should minimize the noise and miss in order to get more hits. The indexing system and search software, they emphasize, are the means to maximize recall and precision. What is known from the statements of the above mentioned authors who still believe in recall and precision may be categorized as follows:

- Recall and precision are a traditional measure for retrieval qualification.
- They are one of the evaluation measures and may be the simplest one.
- There is a classificatory measure in them. This means that if we maximize the potential of the classification/indexing system we can be more hopeful of fulfilling our needs.
- They are ideally measured under controlled conditions.
- They are usually inversely related to each other. That is, by broadening the search we have improved the recall but at the cost of lower precision.

- By minimizing noise (retrieved irrelevant items) and miss (relevant items not being retrieved) we can maximize both recall and precision.

2.9 IMPORTANCE OF INFORMATION RETRIEVAL SYSTEM

- Information retrieval systems solve one of the biggest problems of Knowledge Management (KM): quickly finding useful information within massive data stores and ranking the results by relevance.
- Information retrieval can provide organizations with immediate value--while it's important to try to figure out ways to capture tacit knowledge, information retrieval provides a means to get at information that already exists in electronic formats.
- Information retrieval products are maturing beyond just searching, and now provide capabilities for Knowledge Management (KM) functions such as information dissemination.
- Information retrieval systems are taking advantage of Web technologies and providing browser-based front ends.
- Information retrieval systems are already established in the marketplace--vendors have existing install bases and established revenue streams, which put them in a better position than small startups that are introducing completely new concepts and technologies.

Access information from any data store

The information most organizations use is located in a variety of different data stores. Those valuable repositories may include file servers, groupware systems, relational databases, legacy systems and even external sources such as the Web. Text indexing and retrieval systems can index information in those data stores and allow users to search against it.

Thus, retrieval systems give users online access to information that they might not know about, and they don't have to know or care where the information is located. With a single search, users can query all information that the administrator has seen fit to index.

In that way, companies can essentially "capture" information in any location and give users access to it. The value of that becomes clear as you begin to move text retrieval from a departmental deployment to an enterprise deployment. It's nice to be able to search information within your department, but the real value comes when the information can be shared throughout the organization (Webler, 2022).

The ability to index diverse data stores is hardly trivial. The free Web search engines can only search HTML. Products that include bundled searching capabilities can only search within their own systems. File systems such as Windows NT and Novell can search their own file stores, but they do not index the content, so searching is slow. Only advanced information retrieval systems provide the ability to simultaneously search indexes that have been built from different types of repositories.

While most information retrieval products can index a few different data stores, Fulcrum seems to support the widest range of repository types. Excalibur is leading the way in indexing and searching files types such as images, video and other specialized media formats--for example, users can search an image repository for images that resemble a particular sample. Because so much information already exists in digital form, information retrieval can provide organizations with value right away. Granted, the KM goal includes sharing information that resides in people's heads by allowing users to actively contribute to the group memory--but that usually requires a change in business practices. In the meantime, there is value in giving a broad user base access to a broad information base. Of course, that technology can later be integrated into a more evolved Knowledge Management (KM) practice.

Information retrieval systems include technology designed to help reduce the noise and arrive at more precise results. Most systems provide advanced searching capabilities that allow users to create complex and sophisticated queries, and many systems provide behind-the-scenes functionality to improve precision. For example, Fulcrum and Verity can display document summaries in the result list, so users can quickly determine if the document meets their needs before downloading it. Verity offers the ability to cluster results into categories based on common themes, displaying documents in related groups.

2.9.1 *Moving Beyond Search*

As information retrieval systems mature, they are incorporating new features that address other areas of KM. No longer just limited to searching, they are addressing issues such as delivery, helping users find expertise instead of just data, and helping organizations learn from the types of queries that are being submitted by users.

For example, Verity and Fulcrum offer agent technology that allows users to create queries for information they're interested in, and to add the queries to their user profiles. The server executes the query on a schedule and automatically delivers the results to the user. Thus, users can receive information without having to actively search for it.

Dataware provides technology that helps users locate experts or specialists on particular subjects, instead of just locating information. The company's Knowledge Management Suite can store employee contact information or link to an existing LDAP directory, and can return contact information for particular employees as part of a query result list.

Microsoft's Site Server now includes the company's Index Server indexing and searching engine. In addition, Site Server can perform usage analysis on the queries that users have submitted. Administrators can understand how particular users employ the system by seeing the kinds of queries that users submit. Thus, indexes can be restructured or relationships created to help users get to the information they need more efficiently.

CHAPTER THREE

3.0 SYSTEM DESIGN AND ANALYSIS

3.1 INTRODUCTION

System analysis, according to “Joseph E. Ochiedu” in his book “introduction to system analysis and design” is the term used to describe the process of collecting and analyzing facts in respects of existing operations, procedures and systems in order to obtain a full appreciation of the situation prevailing so that an effective computerized system may be designed and implemented if proved feasible”. By this definition we can now say that system analysis, which is a process, is aimed at studying the network interaction within an organization and assisting in developing of a new and improved method for performing certain necessary task. Also, system analysis can be linked to the act of investigating, analysis, and designing, implementing and evaluating information.

While, system design according to Ochiedu (2013) is the requirement analysis that leads to a detailed system, making a series of specification for the computer based system”. These specifications must show the processing logic the database contents, the input required, the structure of the major programs and manual procedures associated with processing. However, system design in another tem entails, input design and medial files a preceding design and logic, the output design specification and of course in advance the reform analysis and design it contains storage specification. The use of charts, decision tables and graphical illustration are also encouraged in turn coded into a computer program using a suitable programming language. The use of system flow chart table allows for the simplicity of the coded program. Her it also include the program flowchart. However, the system specification for the purpose of a new system is drawn out in the format that will suit the conditions to the investigation and observation carried out.

3.2 ANALYSIS OF THE EXISTING SYSTEM

Coming back to the indexing system, and usually with the lack of information about the user's needs and behavior, some major problems exist with methods for retrieval. The main ones are: Uncertainty. The consent of the user is the major problem. Going through user's consent leads us to the study of his search behavior. And this, in turn, leads us to the uncertainty and probability of one's decision-making. Stating user's behavior and understanding his information needs highly affects the organization and operation of the information retrieval system and may help us in predicting his decision-making.

Search engines have tried to fill the void on the Internet, yet users became more frustrated with the thousands of so- called "hits" beyond their desired results
As the classifier/indexer is not in the same environment where the user may be, although one may try his best in this domain, the differences in the time and place may affect his work. We must accept that both classifier/indexer and user are decision-makers in their job and in the special environment they are in. For example, the cultural and scientific difference between classifier/indexer and user may economically affect the task. And from this angle, a new field of study has appeared, called the Economics of Information. The priority of classification and retrieval, as well as production and consumption, goes back to the debate of the priority of want and need. Because of too much production by some special groups of people, others must apply them. Marketing is ultimately a job which entails more clients for this production. Flaming up the wants, but not needs, generates the motive for more and new production. Therefore, the idea for too much production becomes superior to good consumption. For this reason, every project for IT is to satisfy humans' ambitious desire, while fewer attempts are made for IS to use them in a natural way according to general human needs. Meanwhile, the information acquired in this way

is sent to users to stimulate their appetites and desires. This sending of everything to everybody may cause information traffic or “information pollution

Some statistics mentioned by Kim, *et al.*, (2011) confirm this:

- A recent study of over 300 large companies shows that software or hardware developments fail at a rate of 65%.
- Half of IT projects become runaways while failing to deliver fully on their goals.
- Up to 75% of software projects are cancelled.
- Of approximately 17,500 projects costing more than $250 billion each year, 52.7% will overrun their initial cost estimates by 189%. Most of these projects are delivered with only 74% of original functionality.

The existing system is observed to have been associated with the following advantages and disadvantages, which are listed below”

3.2.1 Advantages of Document Retrieval System

a. Better Alternative

Document Management System has become a better if not the best alternative to traditional manual paper-works. With more and more companies embracing the digital world, the DM system has drastically reduced the usage of papers for documentation.

b. Centralized Storage

It has become more than easier to store all the documents in one place through the DM system. The centralized storage system makes the process of storing documents effortless and efficient.

c. Control and Regulation

With all information stored in a single site, the administrator of the DM system will have control of documents that flows in or out of the system and regulate it accordingly thereby enforcing the order.

d. Security and Privacy

The digitalization of documents has also put it at risk of getting hacked. With sophisticated security mode, companies can safely secure the documents and also install privacy settings in case certain documents are to be hidden from general users.

e. Disaster Management

It is only prudent to safeguard against unforeseen events to save the documents from being lost forever. The backup and recovery system efficiently saves all the documents to the cloud which can be recovered in case there's a crash in the DM system. Nothing's ever lost if they're backed-up regularly.

3.2.2 Disadvantages of Document Retrieval System

a. Security

The ever-evolving technology has given birth to several vices that companies and governments or even private users need to save their data from possible hacking and tampering. If the security controls aren't updated regularly, hackers will find holes in the system which would enable them to loot sensitive information. Even a minor lax in security will cost the company to a great extent.

b. Loss of Data

Improper disaster management will lead to loss of data. Documents need to be backed-up regularly, failure to do which will lead to loss of data.

The beneficial features of the Document Management System overpower its adverse attributes. In a digitally exclusive business set-up, the DM system can play a crucial role in the creation and distribution of data management and increase productivity.

c. **Dependency on Technology**

Digital era means relying heavily on technology. But what if one day, technology becomes a liability? With this in mind, organizations should also consider not having to depend too much on technology in case it becomes a problem in the future.

d. **Security**

With information sharing as one of its features, there’s always a possibility that the information handed might end up in the wrong hands. Other than that, the biggest threat in security is in the internet as business records are the types of information that hackers would love to get a hand on.

e. **Equipment Cost**

Whenever an organization decided to go paperless, a huge volume of data must be scanned. The hardware needed for this type of scanning service would need a substantial amount of money.

3.3 ANALYSIS OF THE PROPOSED SYSTEM

Technological forecast is a prediction of the future characteristics of useful machines, processes, or techniques. People make technological forecasts for many reasons including risk management, market analysis, demand planning, etc. Predicting the future is obviously an exceptionally difficult problem since we may only provide estimates under uncertainty. Since that technological forecasting is a domain dependent and data-driven kind of activity, it is necessary to provide a context that specifies particular data sources and requirements of a decision making person.

In various government and non-government scientific endowments, invited or employed experts are reviewing incoming grant proposals and deciding whether a given research project should be or should not be awarded with a grant or other kind of benefit. Opinions of such experts may be biased by some reasons: field multi disciplinarity, innovativeness factor, and so forth. Experts use various tools for clarifying their decisions: citation indexes, patent databases, electronic libraries, etc. These data are often closed source and can barely be used openly. Nevertheless, producing new informative features for facilitating the technological forecasting tools may reduce the rate of errors made by decision makers while relying on traditional information retrieval (IR) techniques for the rest of search process.

3.3.1 Advantages of the Proposed System

A successful implementation of the computerized crime tracking information system will greatly increase the efficiency of the document retrieval system and will help to ensure that document records are managed properly datas. It will as well ensure the following:

1) Reduction of redundancies and inconsistencies in information retrieval system.
2) Ensure user defined rules to ensure the integrity of data.
3) Enables data sharing across all applications.
4) Ensures data access authorization.
5) Integrity can be improved.

3.3.2 Disadvantage of the Proposed System

1. Time consuming
2. File loss if not properly saved.
3. Failure if power outbreak occurs.

3.4 JUSTIFICATION FOR THE NEW SYSTEM

The new system will help sanities in document retrieval system in Nigeria. The software will be of immense benefit to government. The software will among other things:

1. Facilitate information retrieval system

2. Information management

4. Fast retrieval of documents

Flow Chart Design on Document Retrieval System

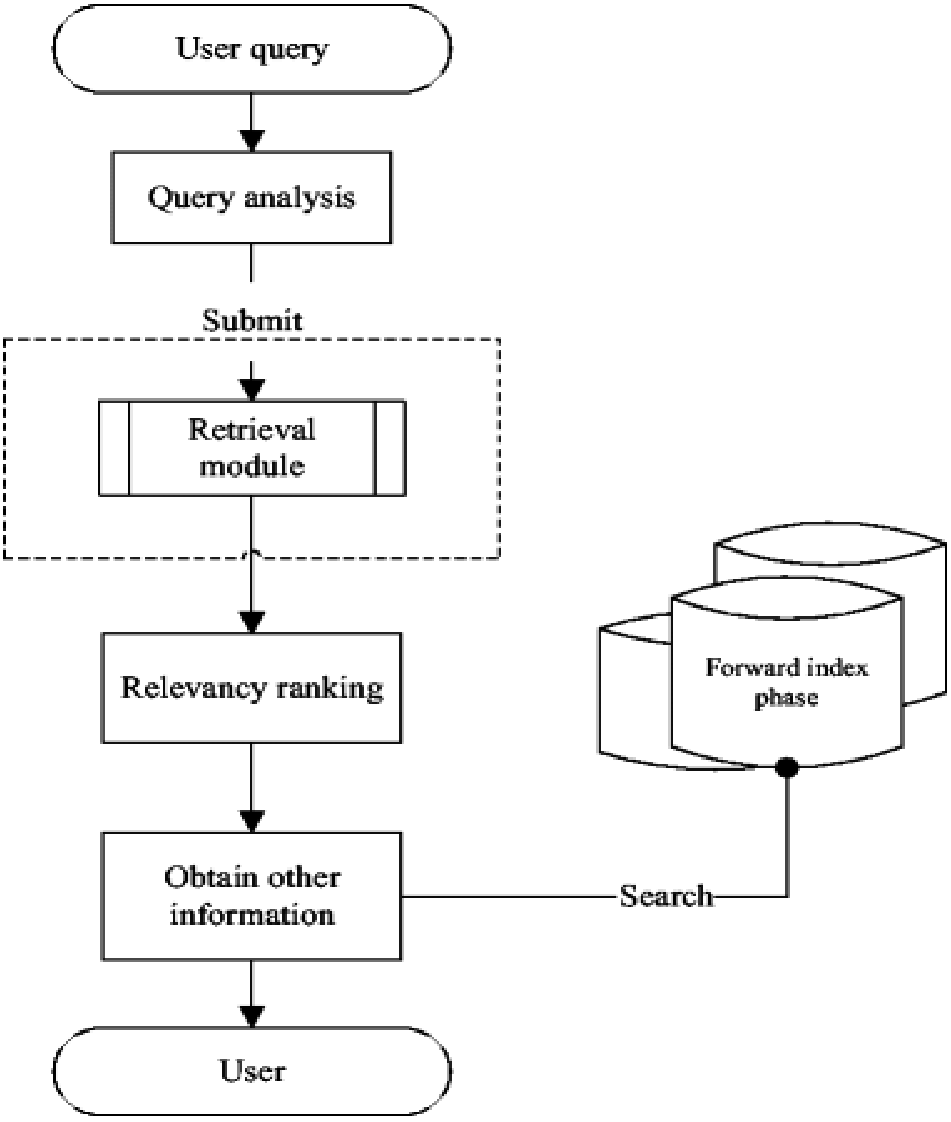


Figure 3.1 Flow Chart Design

3.5 PROGRAMS SPECIFICATION AND DESIGN

This takes care of the billing system in terms of the required information provision to aid the programmer in writing the program and in the maintenance of programs for future purposes.

3.6 INPUT DESIGN SPECIFICATION

The input design entails how the data can be supplied, considering the format data item, type and size. With the specification, the maximum characters for each item and the type, which it belongs to and are used in the entering of data for processing.

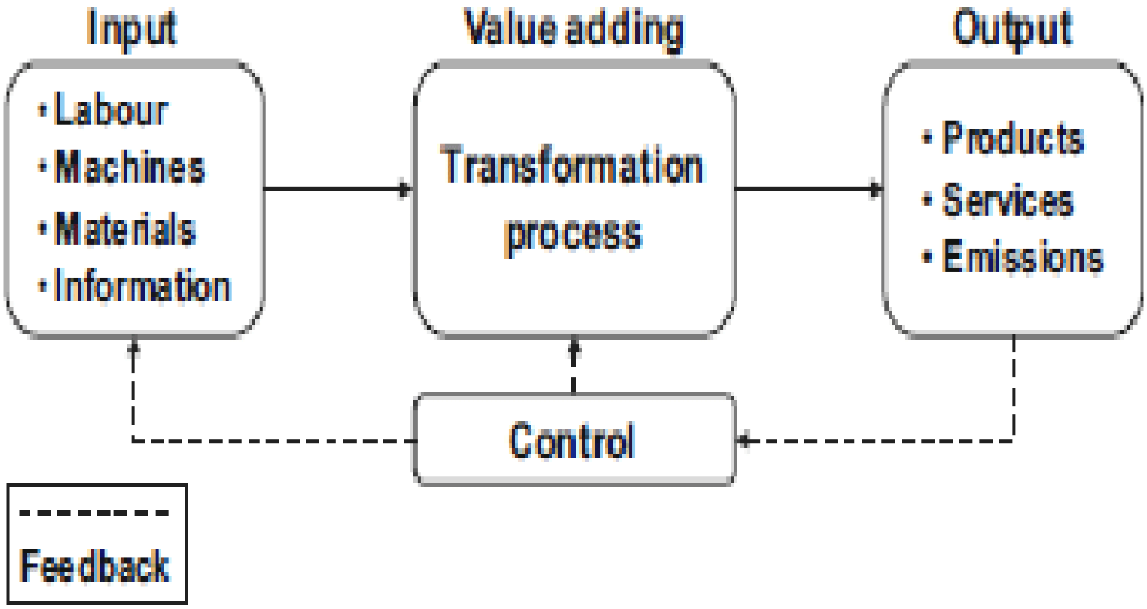


Figure 3.2: Input-Output Specification

3.7 PROCESSING DESIGN SPECIFICATION

The processing design specification is the stage whereby the data processing system would be broken down into a number of routines and then future into a number of computer runs. A routine may be part or all an application and its outputs are usable outside the system. The amount of processing per computer run should be large as possible; it also reveals the processing technique of the system with regards to the type file management would be the most appropriate one to be used for the correctness of input before processing continues.

3.8 OUTPUT DESIGN SPECIFICATION

The output requirement specification may be considered before input needs because knowledge of that is needed for output largely determines the type or nature of input, which can generate the required output. Specification would be based on numbers and types of output required in patient billing system, there could be another output showing summary of income for the day/week month etc. arranged by naira volume, hard copy or soft copy. The hard copy allows the display-generated output to be printed on paper while the soft copy will only allow display on screen. The hard copy will also make it possible for the printing of registration card, billing from services and purchase of medication and master file containing type total amount that has come into the hospital for the day.

3.9 STORAGE DESIGN SPECIFICATION

As to the storage design there is always a general question asked on whether there are more records for processing or not. Immediately the necessary calculation is performed with output stored in the computer memory except on request would it be displayed on screen or printed.

3.10 CONTROL SPECIFICATION

This specification handles the computation of incorrect password and other necessary data item that is typed into the generated question demanding the accuracy or validity of the entered data.

CHAPTER FOUR

SYSTEM IMPLEMENTATION AND TESTING

4.0 INTRODUCTION

Our application package for students' project allocation is realized using ASP.NET (Active Server Pages) as main scripting language, JQuery to simplify menus, CSS (Cascading Style Sheet) to style the interface, MSSQL server as database server, and Vertigo as web server. The application can be accessed suing any web browser.

4.1 JUSTIFICATION OF THE PROGRAMMING LANGUAGE

Adobe Dreamweaver is the most powerful web design software program on the market today. Adobe Dreamweaver gain its popularity through its WYSIWYG (wee-see-wee- what you see is what you get) feature. Dreamweaver is use in this research work for the design of the application. Dreamweaver is a powerful but easy-to-use web site development program that bridges the gap between designer and developer. Although it includes advanced features for developing complex web-based data-driven applications, Dreamweaver's intuitive interface and extensive libraries let even the novice web designer develop a professional web site quickly and easily.

Dreamweaver's interface makes it easy to design and manage both simple and complex web sites by providing a point-and-click interface that simplifies most tasks. Designers can drag and drop page elements in Design view, while developers can work directly with the page's code, making use of the various tools Dreamweaver provides for ensuring correct syntax. The programming Language used is Microsoft ASP.NET. This Language was chosen because of its object oriented features and class libraries for developing online applications.

4.1.1 SWISH MAX VERSION 2

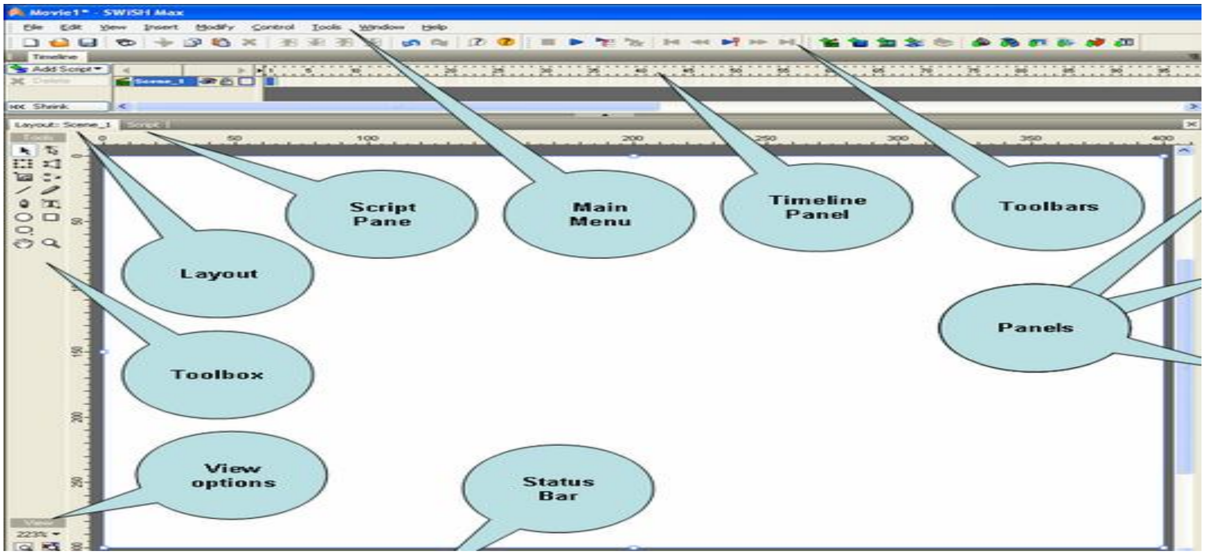
SWISH max is also use in the design for creating animation. SWISH Max is a complete Flash™ animation authoring application. Create stunning and powerful Flash™ animations without using Adobe Flash™.

SWiSH Max is easy to use and produces complex animations with text, images, graphics, video and sound. SWiSH Max has tools for creating lines, rectangles, ellipses, vector and freehand curves, motion paths, movie clips, rollover buttons, and input forms all in an intuitive easy-to-use interface.(Katy Perry, 2017).

Earlier versions were called SWiSH Lite, SWiSH2 then SWiSH Max. SWiSH Max version 2 is the latest addition to the SWiSHzone.com family of Flash™ authoring tools and is an upgrade from the first version of SWiSH Max.

SWiSH Max exports the SWF file format used by Adobe Flash™, so the animation will play on any machine that has the Flash™ Player installed.

SWiSH Max animations can be incorporated into any web page or imported into Flash™. They can also be sent in an email, embedded in a Microsoft PowerPoint presentation or included in a Microsoft Word document. Below is a typical diagram of SWiSH Max interface use for the design of this research work.



4.1.2 HOW TO INSTALL AND RUN THE APPLICATION

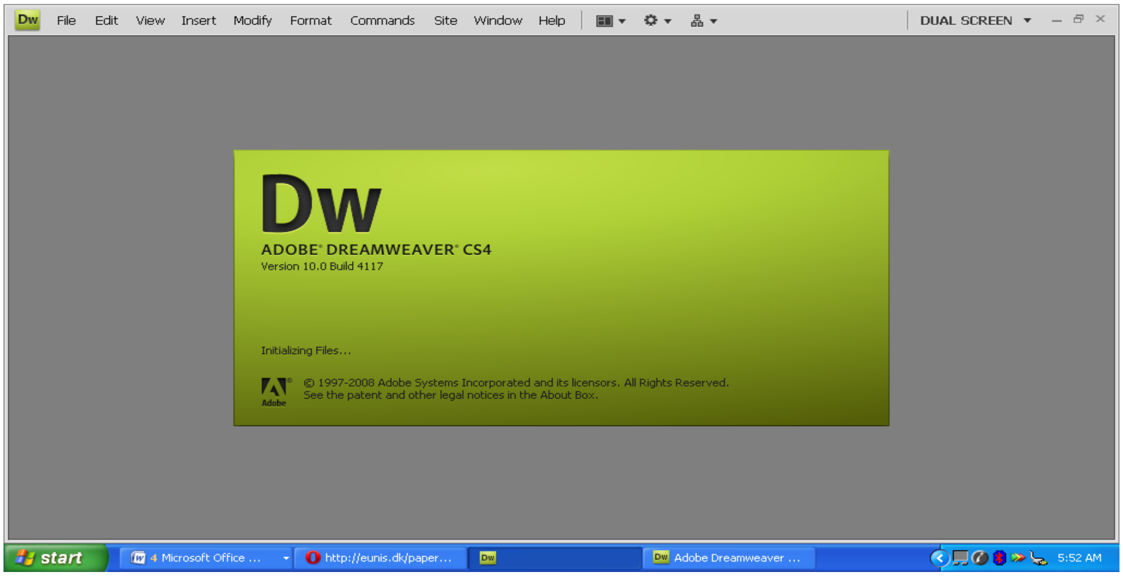
In this section, we will briefly discuss how to setup this web application on a standalone computer system and then how to operate it. The installation that is discussed in this section applies to the ordinary user that wants to implement or use the web application on a standalone computer or over a network. At this level, it is assumed that the web application has been developed already.

To run web application, you need a computer that is capable of running any versions of window XP SP2, Windows Vista and Windows 7.

After that, you need to install Adobe Dreamweaver CS3 or CS4. Because a database is required to power the web application you have to install one, such as internet information server (IIS), but I'll recommend VertrigoServ 2.19

4.1.3 INSTALLING DEVELOPMENT TOOL (DREAMWEAVER)

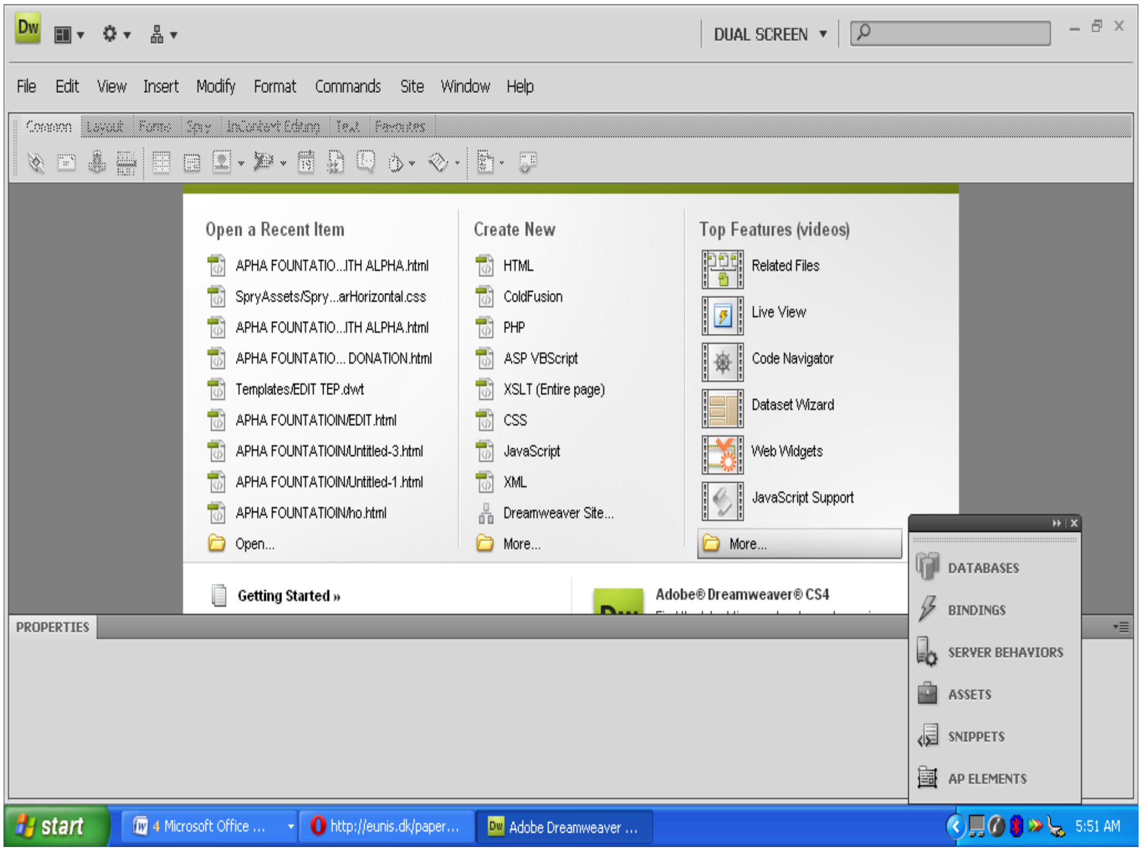
The installation process of DREAMWEAVER CS4 is very easy. Just insert the installation CD/DVD ROM and then follow the instructions on the screen that appears.



Version 10.0 Build 4117

Initializing Files...

© 1997-2008 Adobe Systems Incorporated and its licensors. All Rights Reserved.
See the patent and other legal notices in the About Box.



4.1.4 INSTALLING THE WEBSERVER

In this project a web server is required to provide components that will enable the web Application to run. VERTRIGOSERV is the recommended web server. To enable the installation of the Apache HTTP server and MSSQL Database, download Vertrigoserv from VertrigoServ Project page on http://vertrigo.sf.net install the application by following the steps.

4.1.5 USING VERTRIGO IN THIS PROJECT

The apache web server’s duty is to host and serve the web application’s output to the web browser that requested it. This include receiving the request for a resource that the web server has by the web browser, triggering the necessary server side scripting languages to interpret server side script code (if need be), collect the result HTML document and sending it to the web browser that made the request for the web page. It is important to note that the web server and web browser can both exist on the same machine especially because of design and testing purposes as with this project. But the web server, its utilities and the web browser must be present either together in the same machine or remotely for a web application system to be complete.

Web browser is needed to retrieve the student personal data identification system from the host server (Apache HTTP Server) over the internet or a local area network. It receives the HTML codes for the contents of the current page it is accessing and interprets the HTML codes to produce the interface where data can be collected from the user and sent to the server.

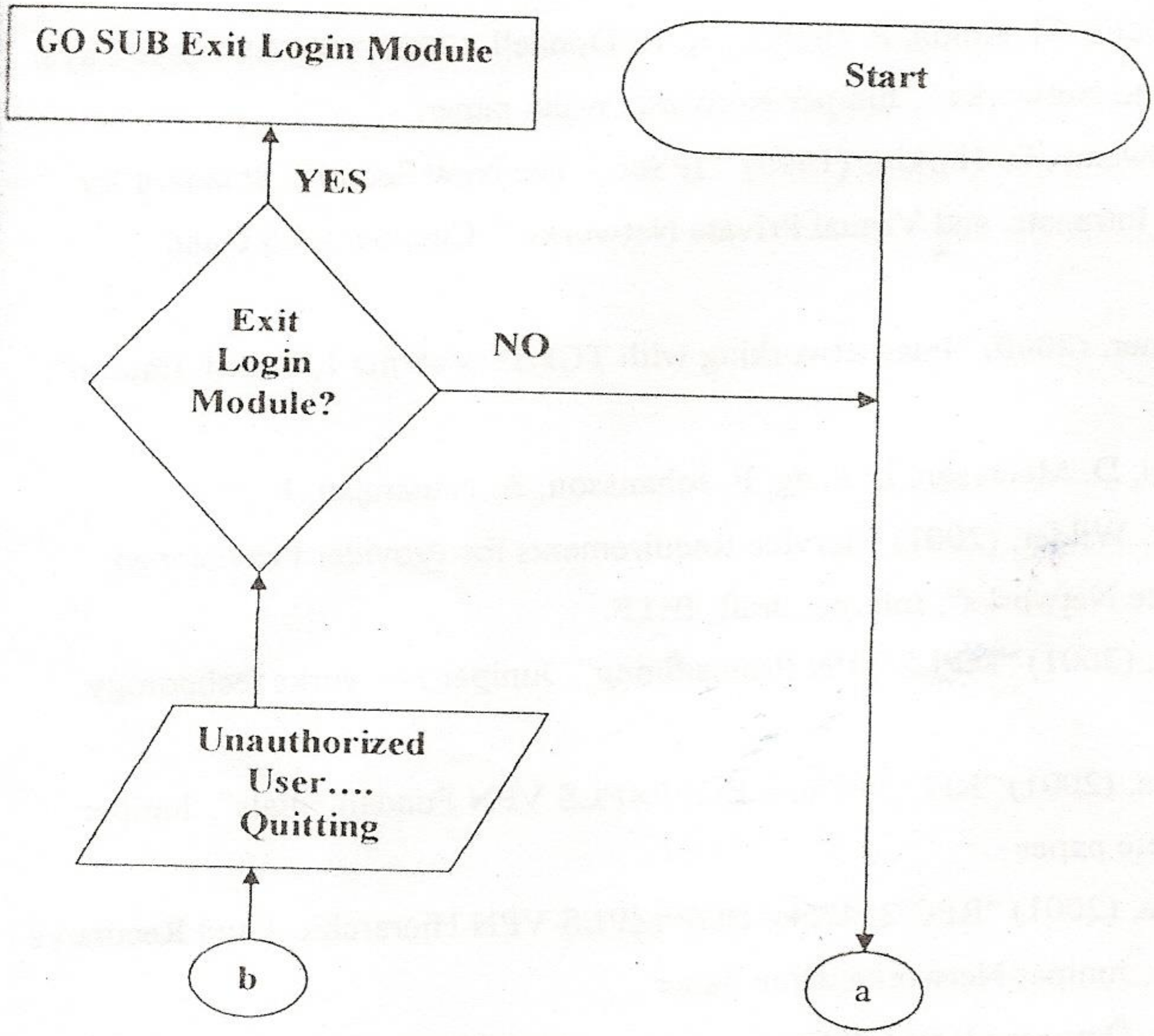
4.2 SETTING UP THE MSSQL DATABASE

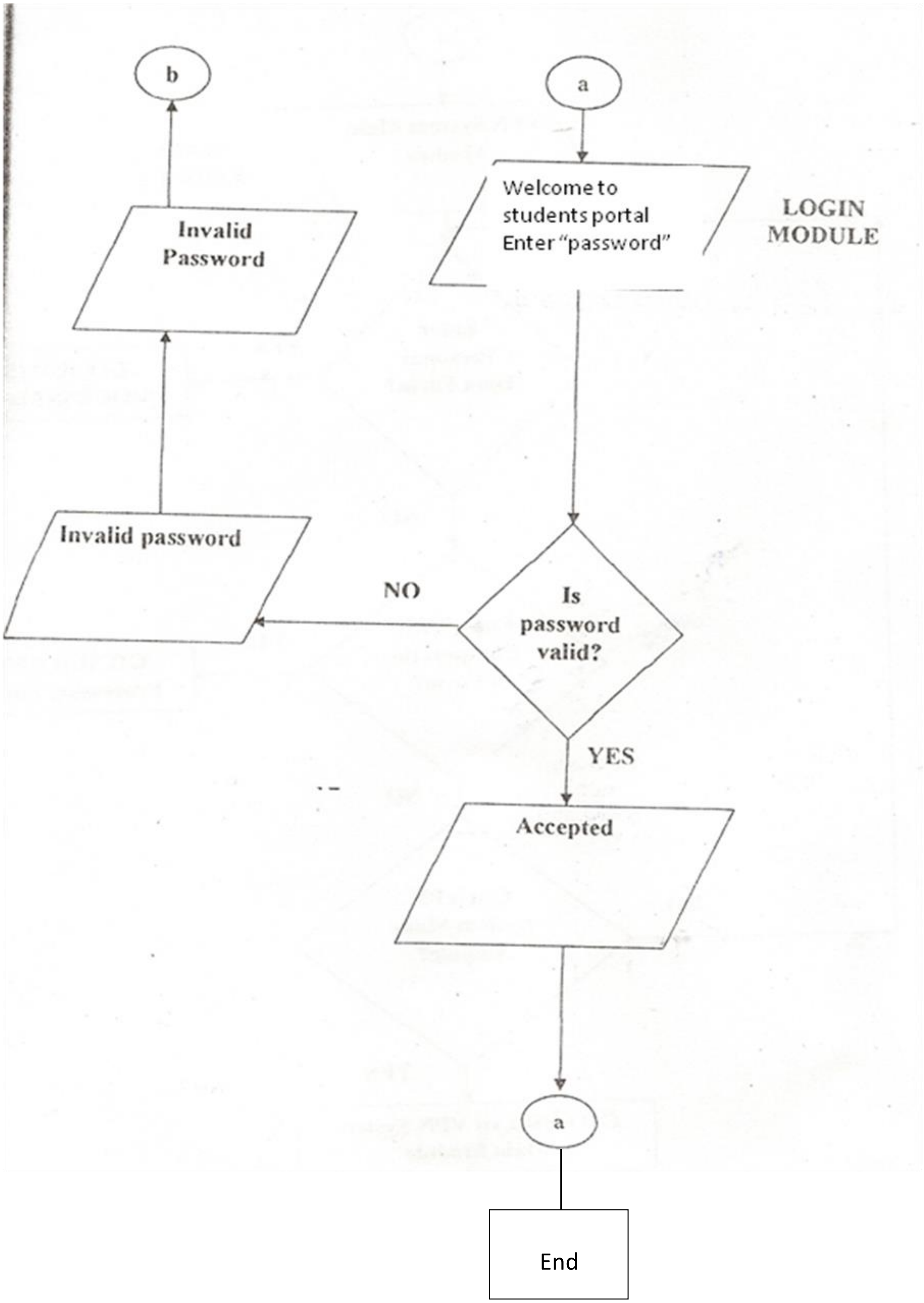
Setting up the MSSQL database is relatively simple. It can be done in two ways; by writing SQL code and with the use of Microsoft SQL Management Server, it is controlled entirely through SQL scripts.

4.3 FLOWCHART OF THE STRUCTUURED DOCUMENT RETRIEVAL

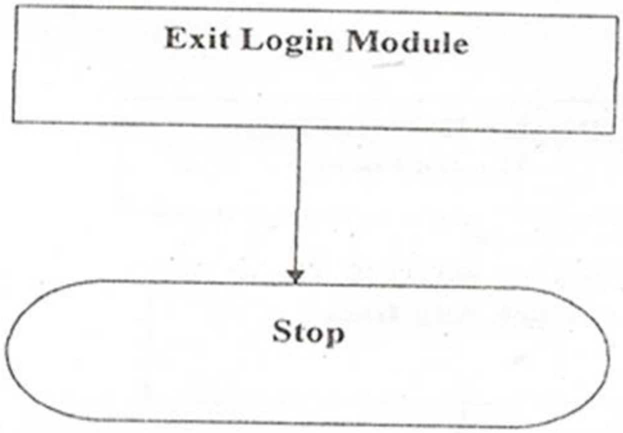
A flowchart is a common type of diagram that represents an algorithm or process showing the steps as boxes of various kinds, and their order by connecting these with arrows. This diagrammatic representation can give a step by step solution to a given problem. Data is represented in these boxes, and arrows connecting them represent flow/direction of flow of data. Flowcharts are used in analyzing, designing, documenting or managing a process or program in various fields. (Wikipedia flowchart 2000)

PROGRAM FLOWCHART

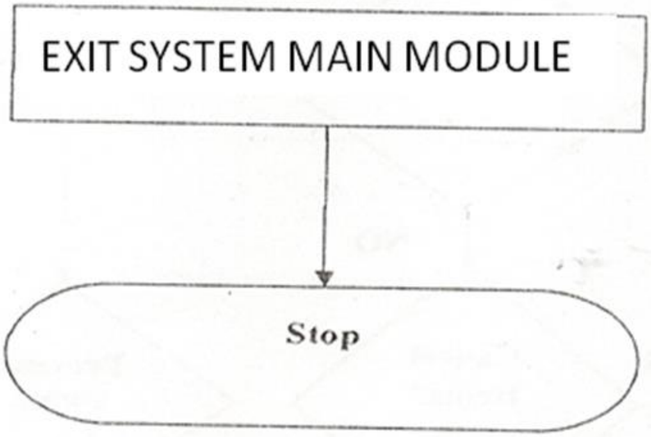



b
a
Welcome to students portal Enter "password"
LOGIN MODULE
Invalid Password
Invalid password
NO
Is password valid?
YES
Accepted
a
End

EXIT PROGRAM LOGIN MODULE



EXIT STUDENT MAIN MODULE



(Isaac Nassi & Ben shnerderman, 1972)

4.4 SYSTEM REQUIREMENT

4.4.1 Software Requirement

This application has been tested and is compatible with all web browser eg. Operal, internet explorer, Mozila Firefox etc.

Due to the flash content in the application, it therefore requires adobe flash to be installed in the system for it to display all the flash contents.

4.4.2 Hardware requirement

System must have internet connection to run the application. Internet speed as low as 256Kbps can run the application very well with great speed. And 32-64bits operating system, such as Windows 7, 8, temporary storage device 2-4 GB RAM, 50-100GB Free Hard Disk space, Compatible CPU (intel i5/i7/xeon) JEPG images about 14-20 MP clarity.

4.4.3 People

With its easy-to-use interface, it requires little or no pre-internet knowledge before use lecturer or student can use the application.

4.5 IMPLEMENTATION

When preparing system implementation plans, certain things must be considered. For this project, the new system differs a little from that of the existing one because of its nature.

It is computerized and requires the services of competent and well trained staff for it to be effectively operational.

Implementation of the new system involves:

(I) Training of staff

(2) System testing

(3) System change over

(4) System review and maintenance

4.5.1 TRAINING OF-STAFF

It is important to prepare training schedule for the staff before the new system is to be installed. The user of the new system should be given specific time for training courses. This will enable them fit into the new system. Also, user manual will be produced in regards to the operation of the new system.

4.5.2 SYSTEM TESTING

For the implementation of the new system, data must be prepared for live testing. The result from the new system is compared to that of the existing system to check if the expected result was achieved. It is also necessary to formulate the operation of the new system to check the overall time and ability of the staff to handle the operation of the new system.

4.5.3 SYSTEM CHANGE OVER

The parallel method is adopted in the changeover process. This method was adopted because it creates an avenue where by the old and new systems are being run concurrently. With this method, the users of the system will gain a practical knowledge of how the new system is being operated. When this is achieved, the old system is discontinued and the new system takes its place. This method also helps to introduce the new system to users having little or no notice of the change over process.

4.5.4 SYSTEM REVIEW MAINTENANCE

The system should be reviewed and maintained periodically in order to deal with unforeseen operational problems that may arise and to make sure that the new system meets its planned objectives or standard.

4.6 DOCUMENTATION

The administrator controls the logging in process in such a way that unauthorized user do not log in, add new lecturer/supervisor to the list, update lecturer/supervisor’s profile, determine if a student should be given project supervisor after students assessment, add and delete student or supervisor below requirement.

CHAPTER FIVE

5.0 SUMMARY AND CONCLUSION AND RECOMMENDATION

5.1 SUMMARY

In drawing this project study to a close, we can say to a large extent that the major aims and objectives of this study were achieved in data and information correlation in the Auchi Polytechnic, Auchi, using Computer Science Department as a Case Study that entails the output of daily report billing sheet, document retrieval and at the same time structured document retrieval system. We could also say that the services of a system analyst needs to be employed of a reliable and effective computer based information system is to be achieved.

5.2 CONCLUSION

Retrieval depends primarily on classifying/indexing. But the main thing is the way we look at it. If our view is dynamic, that is, if we classify/index to retrieve, then everything may change. And if, in theory or in practice, we classify/ index because it is a job and we are told to do so, nothing will change. Even in the case of information engine providers, although they try to satisfy users by gathering databases, it seems that their main idea is to attract audiences' attention by their abundance of information, not by methodology and their help systems. This is the same as a static view to the library and information system, in which the accumulation of information is more important than successful retrieval. It means that every library and information system, as well as information databases, in order to become a super power as an information provider, tries to increase its assets by collecting that which relates or does not relate to it. This may also be because there has never been a clear definition for their activities. In such situations, serving the clients may be considered to be the secondary task. Another factor may lie in the fact that the libraries, databases and information providers are not established primarily for the sake of needs, but rather for the sake of wants. And although wants lead to more creation, bring research and development, and initiate new activities, they may be beyond actual service to the real needs of clients.

5.3 RECOMMENDATIONS

Efforts have been made to design and develop software that support network. But there are still areas that may be considered as a further and important area to improve on, and my suggestion go thus.

- There is the need for the magistrate segments and structured document retrieval system
- The development of DNA database on structured document retrieval system.

REFERENCES

Tazzite N., Yousfi A., Bouyakhef E. H, (2008)"Conception et réalisation d'un système de recherché d’information intégrant des connaissances sémantiques dans la phase d’indexation". Colloque International sur les technologies de l’information. IRCAM, Rabat, 24-25 Novembre.

Kim,W., Aronson, A.R. and Wilbur, W.J. (2011). "Automatic MeSH term assignment and quality assessment". Proc AMIA Symp: 319–23. PMC 2243528. PMID 11825203.

Lin, J.I. and Wilbur, W.J. (2007). " A probabilistic topic-based model for content similarity". BMC Bioinformatics. PubMed related articles: **8**: 423. doi:10.1186/1471-2105-8-423. PMC 2212667. PMID 17971238.

Hubert, P. Y. (2005). Information Theory, Evolution, and the Origin of Life. Cambridge University Press. p. 7. ISBN 9780511546433.

Luciano, F. (2010). Information - A Very Short Introduction. Oxford University Press. ISBN 978-0-19-160954-1.

Webler, F. (2022). "Measurement in the Age of Information". Information. **13** (3): 111. doi:10.3390/info13030111.

. Jiri Hynek. (2002). Document Classification in a Digital Library. Technical Report no.
DCSE/TR 2002-04, p. 20. http://www.kiv.zcu.cz/publications/2002/tr-2002-04.pdf

Arthur Maltby. (1975). Sayer’s Manual of Classification for Librarians. 5th ed. Gt. Brit.: Andre Deutsch/Agrafton Book, p. 309; Jennifer Rowley and John Farrow.
(2000). Organizing Knowledge, 3 rd ed. London: Gower, p. 341; Masse Bloomfield

Masse Bloomfield. (2001). “Indexing–Neglected and Poorly Understood,” Cataloging
& Classification Quarterly: 33(1):70.

Joseph. E. Ochiedu (2003) Introduction to System Analysis and Design, Marlon Technologies Auchi Edo State.