IMPROVING INFORMATION RETRIEVAL FROM PDF DOCUMENTS USING INDEXING

[1]SHWETA J. PATIL, [2]DILIP K. BUDHWANT

[1,2]Department of Computer Science and Engineering, Jawaharlal Nehru Engineering College, Aurangabad, M.S, India

Abstract - For thousands of years people have realized the importance of archiving and finding information. With the advent of computers, it became possible to store large amounts of information; and finding useful information from such collections became a necessity. Several Information retrieval (IR) systems are used on an everyday basis by a wide variety of users. Information retrieval is become an essential research area in the field of computer science. Information retrieval is generally concerned with the searching and retrieving of knowledge-based information from database. This paper describes working comparison of different document indexing software to improve efficiency of information retrieval from large volume of documents.

Keywords - Information Retrieval (IR), Indexing, Searching

I. INTRODUCTION

Information retrieval (IR) is the task of representing, storing, organizing, and offering access to information items[1]. Information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)[6]. Information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents[8].
IR is different from data retrieval, which is about finding precise data in databases with a given structure in efficient way. There is a strong need to gather information according request when it possible or to point researcher to systems where information can be found. It is very important to know if the gathered information is actual and complete. So, it a necessity to find a solution for a problem data integration, which will be 1) easy to implement for any participator as well as flexible enough to embrace diversity and data meaning and structure in organizations 2) powerful to go provide sophisticated information retrieval services for users. There is a strong need to integrate information from different sources and to provide access to all information to users, enabling them to utilize a wide range of sources. Data retrieval from different sources has become a hot topic during the last years. For instance, there are such data sources as employee data source, student data source, library data source etc within the same enterprise (talking of academic institution). When someone wants piece of information we need to execute n queries and possibly provide user with n such results, retrieved from n data sources[1].

A. Information Retrieval System

Information retrieval is the art of demonstration, storage, organization of and access to information items. The representation and organization of information should be in such a way that the end user can access information to meet his/her information needed. In Dataspace, information retrieval finds the structured; semi-structured or unstructured data that satisfies information need from within in the Dataspace, and it inform the end user on the existence and where about of data relating to his or her query[1].

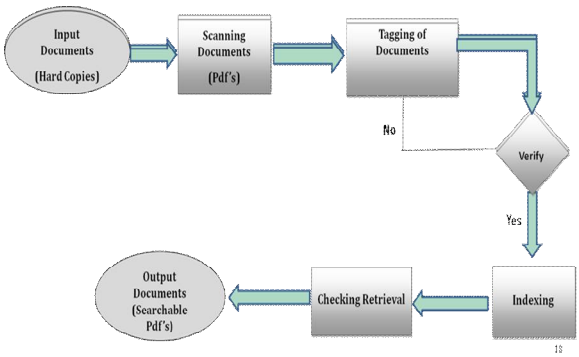

Fig 1- System Flow of Information Retrieval System

Different Phases of Information Retrieval System

1. Scanning of Documents

- Normally scanning at 300 dpi is recommended.
- Maximum dpi limit can be up to 600**.**

2. Tagging of PDF Documents

- Read page images
- Analyze page images
- Recognize the contents of image.

3. Indexing of PDF Documents

- Collect all PDF documents to be indexed into one or more folders.

4. Searching of PDF Documents

- Searches can be done using a single string, multiple search strings, and patterns.

In this paper emphasis is given on last two phases i.e. indexing of PDF documents and searching of PDF

documents. For indexing of PDF documents it is required to perform scanning and tagging of PDF documents using OCR technique. The aim is to improve content retrieval from PDF documents by reducing time.

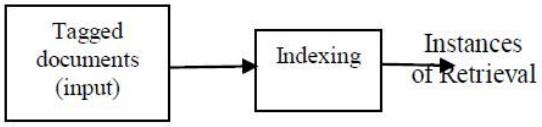

Fig 2- Indexing Workflow

II. RELATED WORK

A. Different Indexing Software

1. Nuance PDF Converter Professional

Nuance PDF Converter Professional gives you extensive control over your PDF. You can edit pages and documents, annotate and review them, adjust document security, sign your documents and more. PDF documents can be compiled from different sources and pages can be rearranged. The program delivers a document management system: use it to create archives from related documents to index and search them later[10].

Testing Observation-

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

2. Adobe Reader

A document that consists of scanned images of text is inherently inaccessible because the content of the document is a graphic representing the letters on the page, not searchable text. Scanned images of text must be converted into to searchable text using optical character recognition (OCR) before addressing accessibility in the document. Searches for artifacts, OCR suspects, and unmarked (untagged) content, comments, links, and annotations. Options allow searching the page or document and adding tags to found items. Indexing in Adobe Reader gives effective retrieval[11].

Testing Observation-

- Simple User Interface.
- It gives all correct instances while searching.
- Retrieval of data is not satisfactory.
- Accuracy for content retrieval is more.

B. Feature Comparison of Indexing Softwares

Based on above comparison of indexing software we decided to use Adobe Reader software for indexing PDF documents.

Table 1- Feature Comparison of 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

III. PROPOSED INDEXING SOFTWARE- ADOBE READER

The basic concepts apply to searching many different collections of PDF files. These PDF files are tagged documents with identified specific keywords from scanned PDF documents. We can reduce the time required to search a long PDF by embedding an index of the words in the document. Adobe Reader can search the index much faster than it can search the document. Users search PDFs with embedded indexes exactly as they search those without embedded indexes; no extra steps are required. Begin by creating a folder to contain the PDFs you want to index. All PDFs should be complete in both content and electronic features, such as links, bookmarks, and form fields. If the files to be indexed include scanned documents, make sure that the text is searchable. Break long documents into smaller, chapter-sized files, to improve search performance. We can also add information to a file's document properties to improve the file's searchability. The main aim of the system was to integrate the speed of digital search with the visual overview of keyword locations that comes with document indexing[11].

A. Indexing of PDF Documents Using Adobe Reader

1. Information Retrieval Process

Information retrieval is generally considered as a subfield of computer science that deals with the representation, storage, and access of information[7]. Information Retrieval (IR) is the process by which a collection of data is represented, stored, and searched for the purpose of knowledge discovery as a response to a user request (query)[9]. The goal of any information retrieval system is to satisfy user's information need. Unfortunately, characterization of user information need is not simple. User's often do not know clearly about the information need. Query is only a vague and incomplete description of the information need[1].This process involves various stages initiate with representing data and ending with returning relevant information to the user. Intermediate stage includes filtering, searching, matching and ranking operations.

The main goal of information retrieval system (IRS) is to "finding relevant information or a document that satisfies user information needs". To achieve this goal, IRSs usually implement following processes:

1) In indexing process the documents are represented in summarized content form.

2) Searching is the core process of IRS.
There are various techniques for retrieving documents that match with users need[2].

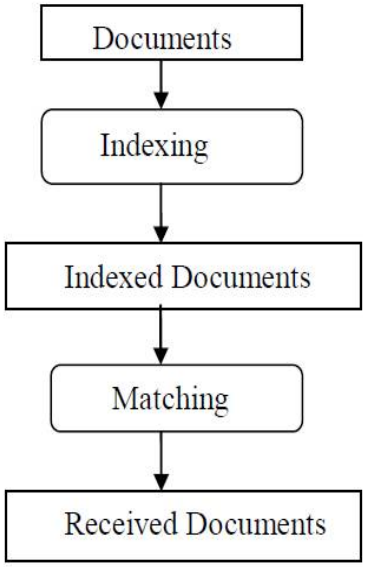

Fig 3 – Information Retrieval Process

Information retrieval process includes-

1. Create index with the model according to the text of database. Indexing can greatly improve the speed of in-formation retrieval. Which way do you use depends on the scale of information retrieval system. Large-scale information retrieval systems such as Google, Baidu take advantage of the approach of inverted index.

2. Search-After indexing the documents, you can start to search information you need. Search requests are submitted by the users and information retrieval systems to preprocess and search the information eventually return user the information[4].

2. Indexing Techniques
There are several popular information retrieval (IR) indexing techniques; the technique which we have used is called Inverted Indices.

Inverted Indices
Inverted indices are widely used in industry. They are easy to implement. The inverted lists could be rather long, making the storage requirement quite large[8].
Each document can be represented by a list of keywords which describe the contents of the document for retrieval purposes which we called tagged PDF document. Fast retrieval can be achieved if we invert on those keywords. The keywords are stored, eg alphabetically; in the index file for each keyword we maintain a list of pointers to the qualifying documents in the inverted file. This method is followed by almost all the commercial systems[2].

Importance is given on how to reduce time costs, with particular emphasis on environments in which indexing has been used[3].

In IR, inverted indices consist of a search structure for all searchable words called a dictionary, and lists of references to documents containing each searchable word, called inverted lists. An inverted index for an attribute in a decision support system consists of a dictionary of the distinct values in the attribute, with pointers to inverted lists that reference tuples with the given value through tuple identifiers (TIDs)[5]. Index the documents that each term occurs in by creating an inverted index, consisting of a dictionary and inverted lists[6].

Steps for Indexing-

- Collect all tagged PDF documents to be indexed into one or more folders.
- You can build an index file from all the tagged PDF files in a set of folders you define. Before starting you choose a folder where the index will be stored. Indexing proceeds in the background. A small index definition file is created, with the extension pdx.
- Carry out text searches in the currently open PDF, all PDFs in a given folder or on a prebuilt index file.
- Searches can be done using a single string, multiple search strings, patterns or masks. Use the Organizer to specify the search criteria. A PDF Index file (.pdx) is a searchable archive of PDF documents.

IV. EXPERIMENTAL RESULT

There is training data of organization containing Approximately 15 Lakhs tagged pages which need to be retrieved efficiently and correctly with less amount of time.

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. Result is presented for four cases having 20 keywords per page and obtain following result.

Table 2- 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%

CONCLUSION

This paper proposes an effective information retrieval indexing method using the evaluation copy of Adobe Reader. The work extensively has tested the same using Nuance PDF Converter Professional also.
Retrieval accuracy is also more in Adobe Reader. Adobe Reader is good solution for indexing in all respect as it gives correct instances.
Accuracy of correctly retrieved documents by Adobe Reader is 90%.
After indexing the documents, retrieval is much faster than simple tagged documents. Though PDF documents tagged with indentifying keywords to enhance searchabilty, indexing gives better solution to retrieve information efficiently. It reduces time for information retrieval. Thus indexing needs to be carried out after tagging of documents to increase retrieval accuracy.

REFERENCES

[1] Information Retrieval System and challenge with Dataspace, International Journal of Computer Applications (0975 – 8887) Volume 147 – No. 8, August 2016.

[2] Akram Roshdi and Akram Roohparvar, Review: Information Retrieval Techniques and Applications, International Journal of Computer Networks and Communications Security, VOL. 3, NO. 9, SEPTEMBER 2015, 373–377.

[3] ALISTAIR MOFFAT and JUSTIN ZOBEL, Self-Indexing Inverted Files for Fast Text Retrieval, ACM Transactions on Information Systems, Vol. 14, No. 4, October 1996, Pages 349–379.

[4] Rujia Gao, Danying Li, Wanlong Li, Yaze Dong, Application of Full Text Search Engine Based on Lucene, Advances in Internet of Things, October 2012, 2, 106-109.

[5] 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.

[6] Christopher DM, Prabhakar R, Hinrich S. Introduction to Information Retrieval. Cambridge University Press; 2008.

[7] M.François Sy, S.Ranwez, J.Montmain,"User centered and ontology based information Retrieval system for life sciences", BMC Bioinformatics, 2012 Jan.

[8] R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", IJCER, sep 2012, Vol. 2 Issue. 5, PP: 1443-1444.

[9] Anwar A. Alhenshiri, "Web Information Retrieval and Search Engines Techniques",2010,Al- Satil journal,PP: 55-92.

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

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

★★★