Zahra et al LGURJCSIT 2018

LGU International Journal for
Electronic Crime Investigation

LGU (IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

Vol. 2 Issue 4, October - December 2018

Desktop Based: Off-line Information Retrieval System

Syeda Binish Zahra [1], Syed Muhammad Shabih-ul-Hassan [2]
binishzahra@lgu.edu.pk, binishzahra@ncbae.edu.pk, shabih@pac.edu.pk
[1] Department of Computer Science, Lahore Garrison University
[2] Professional Academy of Commerce

Abstract:

Information retrieval is rapidlydeveloping field and there are many changes are introduced day by day in traditional techniques for IR. IR system is intended to evaluate examine and accumulate the sources of information and get back those that match user's requirement. The need of today's fast moving life is to get maximum information but within minimum time. For getting maximum information in minimum time requires more efforts. The main functionality of IR is to provide access of documents (that document may be in collection of thousand, or millions). With the help of an appropriate description, user can recover any one document from a collection of documents. In this paper I describe my IR system which retrieves information from any directory and this information may be in terms of image, audio or in text form. The selection of good features also allows the space, time and costs of the retrieval process to be reduced. Two documents may be considered similar in this system if they have same name and places in different folders or directories. To explore the retrieval process from that system, I used Apache Lucene with JAVA implemented in IntelliJ.

Keywords: Information Retrieval (IR), human-interaction

I. Introduction:

The task of Information Retrieval is to find appropriate document from a repository. IR system typically accepts a query as a string of words and returns a ranked list of document from that repository [1]. With the help of good human-interaction methods with systems, an information retrieval system can give the ease-of-use of given system even if it having two distinct features. The mean of two distinct features are natural language interpretation and visual interface. Visual interaction with system for local search is done by visual interface and linguistic interaction with system for global search is done by natural language interpretation [2]. It is beneficial to use visual interface for IR systems because by using this user can create expressions of query easily, and can consult with system and even react with the system easily. Most tools are devising for IR to aid information. Many tools are already available to support electric mail, data analysis, manipulation of spreadsheet, preparation of text and even some tools can edit and playback audio\video data [3]. However, for retrieving and sorting above mention information; tools are remaining primitive. In 1951 Bagley's suggested that searching 50 million item records each containing 30 index terms would take approximately 41,700 hours. Since 1960s and 1970s major IBM innovations are introduced such as IBM desktop filing system and free-text search systems. But these systems deals with only information explosion. Our goal is to improve the user ability of retrieving information through their system more easily and also within minimum time slot. We design a system for retrieve the information through directories of user system and that system handle easily the process of finding information of their relevant query. For design such type of IR system we use JAVA in IntelliJ and include libraries of Lucene; the well-known open source engine, for diverse retrieval experiments and in different ways.

II. Proposed System:

Database systems are design to store the entire information in the form of data (all data in documents is structural), but IR system use simpler model than any database system used [5]. IR organized information as a document collection even these documents are not in structure form (no schema).IR locates relevant documents, on the basic of input (enter by user) such as finding documents containing the words "database system". When user wants to retrieve information, then user temporarily enters into a new world. This world is confine from the rest of their computer environmentwhen user returns; they are with result of their requested query. In given system a short query in form of number will do a remarkable job of searching material which is relevant to the query. In this paper we discuss the algorithms that we use in implementation of our IR system. In which we test our system by given it different queries. In final session in this paper our system will generate results of the given system.

a) How our System Works:

Our system offers a full data retrieval mechanism base on their directories numbering. User just enter a desire number of their directory (that numbering is assign to directories by our system) to extract their desire information. After user enters their desire directory number, the system will return all sub-directories of that selected directory. Our system will return a list of all folders which that directory contains.

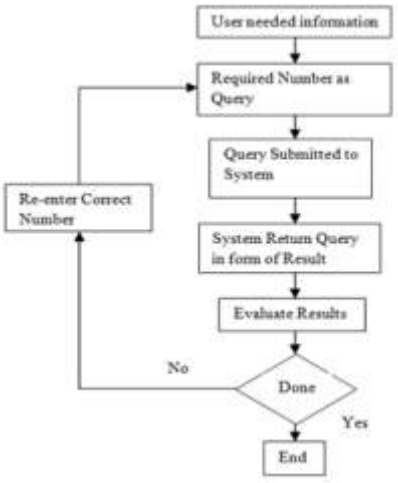

Figure 1: Standard Information Retrieval interaction Model

b) Feature of Our Proposed System:

Our system can easily be supported any new directory that added at run time (a new folder or even a new USB data). For example if we attached a new USB device with our desktop system and that device have multiple folders in it. If we want to retrieve information from that device, our system will also handle those queries too with proper and accurate retrieval of information.

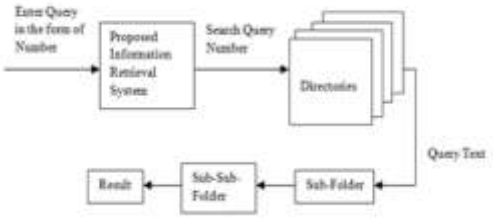

Figure 2: Information Retrieval Proposed Model
It is our system ability to index all types of documents such as plain text documents, audio\video files, Adobe PDF files, Microsoft Excel, Word, PowerPoint files, HTML documents, Web Documents (those web documents that are store somewhere in our system directories) and any image file (jpg, tiff, png and gif format images). Each term extracted from a directory has three fundamental properties

1. Actual directory
2. Theposition at which the number assign to subdirectories
3. The fields in which that query occur.

c) Index Structures [4]:

Our system index consists of 4 main data structures:

Lexicon:

The terms and term id (a unique number for each term) are store in lexicon, along with the global statistics of the term (term document frequency and global term frequency) and the offsets of the postings list in the Inverted Index.

Inverted Index:

Thepostings lists of a term are store in inverted index. In particular, for each term, the inverted index stores:

ü The document id of the matching document; and the term frequency of the term in that document.

Document Index:

- The Document Index stores
- The number of document (an external of the document's unique identifier),
- The id of document (internal document's unique identifier),
- The document length in terms of tokens;
- The document offset in the Direct Index.

Direct Index:

The Direct Index stores

- The terms and term frequencies of the terms present in each document.
- The main purpose of the Direct Index is to efficiently handle query expansion and also provide easiness.

• III. Implementation And Results:

To retrieve the whole data from specific folder, our system use given algorithm.

```
private ArrayList<File>getAllFiles(String path)
{ArrayList<File> files = new ArrayList<>();
Collections.addAll(files, (new File(path).listFiles()));
intsize = files.size();
for (inti = 0; i < size; i++) {
if (files.get(i).isDirectory()) {
files.addAll(getAllFiles(files.get(i).getAbsolute Path())); } } return files;
}
```

The following figure shows the implementations of proposed system and their corresponding results.

Figure3: Results after running the code

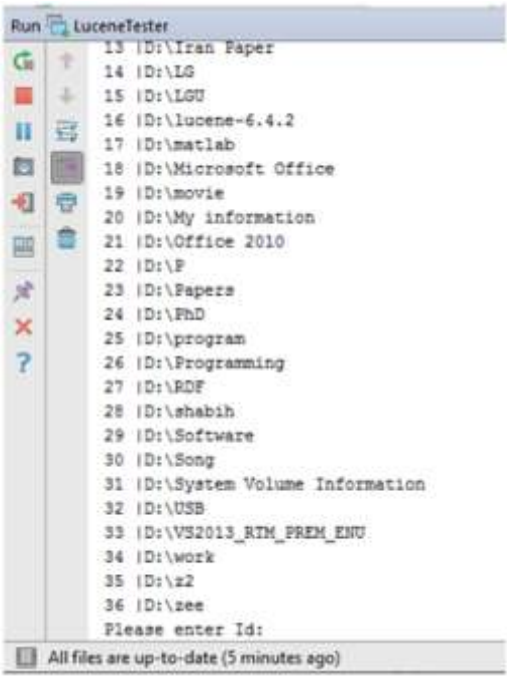

Figure 4: Results after enter the id 1, which represent D directory

Figure5: Results after enter the id 24, which represent D directory folder PhD

Figure6: Results after enter the id 8, which represent D directory folder PhD, sub-folder IR

Figure7: Results after enter the id 5 which represent D directory folder PhD, sub-folder IR, sub-sub-folder Research Paper

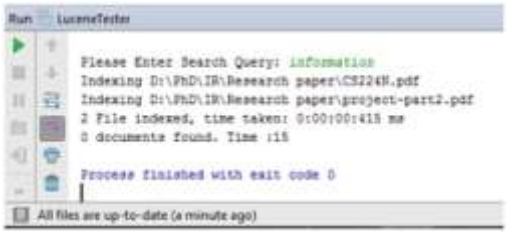

Figure 8: Results after enter the id 0, and then Query "information".

IV. Conclusion:

Our goal with this work is to stimulate new thinking about how to retrieve data through directories in a fast manner. In the future, we hope to evaluate our system with just entering required query and given query will return as output of system with proper data path at only 1 time. With small changes our implementation can be improved and overall retrieval tools will replace by our designed system.

V. References:

1. Hsieh-Chang Tu, Jieh Hsiang, "An Effectiveness Measure for Evaluating Open Retrieval System" in workshop of OSIR (Open Source Information Retrieval)' 06 Seattle, USA

2. Thomas Erickson, Gitta Salomon, "Designing a Desktop Information System: Observations and Issues", in Human Factors in Computing Systems: CHI' 91 Proceeding. ACM: 1991.

3. Michael G. Lamming, William M. Newman, "Activity-based Information Retrieval: Technology in Support of Personal Memory"

4. IadhOunis, Gianni Amati, VassilisPlachouras, Ben He, Craig Macdonald, Christina Lioma, "Terrier: A high Performance and Scalable Information Retrieval Plateform" in workshop of OSIR (Open Source Information Retrieval)' 06 Seattle, USA

5. T. Grust, M. van Keulen, and J. Teubner. Staircase join: Teach a relational DBMS to watch its (axis) steps. In Proceedings of the 29th Conference on Very Large Databases (VLDB),2003.