Information Retrieval and Web Search

A picture containing text, electronics, night sky, display

Description automatically generated

Submitted By:

Name: Harshvardhan P

UID: U00827959

Due Date: 12/06/2021

**Retrieval Engine for the University of Memphis**

**Introduction**

The term "Information Retrieval" (IR) refers to a software application that organizes, stores, retrieves, and evaluates textual information from document repositories. Information Retrieval is the process of extracting unstructured information, often in the form of text, from a huge collection of data stored on computers. It is possible to store, organize, and retrieve information using an IR system. To do a search, a collection of keywords must be entered. People type in keywords into search engines to find what they're looking for. These keywords provide a succinct summary of the content.

**Approach**

In Retrieval Engine for The University of Memphis, Unstructured materials are the focus of information retrieval. In the beginning, IR was all about the text because of advances in machine learning and increased processing power, new sorts of data, such as images and audio, can now be retrieved. Structured metadata commonly accompanies unstructured materials. Metadata is information about a document. When it comes to emails, for example, there will be fields such as the sender, recipient, topic, and content. This is helpfully organized data. This is another example of a digital camera picture. The date, the camera type, the exposure setting, and the image size are all examples of useful information. Documents may be semi-structured in the sense that heuristic algorithms can find information in "standard" areas. Sections and subsections in-text texts are also signs of a semi-structured composition. IR systems may make use of the semi-structured nature of XML documents. Every day, as computing power rises and storage costs fall, the volume of data we handle grows dramatically. However, the data we gather is of little use unless it can be retrieved and queried. Data interpretation relies heavily on information retrieval technologies. Google and other search engines make finding information on the Internet much easier than it is now. Without information retrieval systems, information is just a bunch of numbers (Berger, 2017).

Traditionally, the success of IR has been assessed by comparing its accuracy to its recall, two concepts that are often at odds with one another. A high-precision system will not return irrelevant documents, but a high-recall system will make sure that the user receives all pertinent documents. It is ideal to have a good balance of both, but if this cannot be accomplished, various applications will tend to favor one over the other. Indeed, it is often the case that enhancing one metric results in a decrease in the other's value. When it comes to a situation requiring pinpoint accuracy, nothing beats the world of online search. Users don't need access to all the important papers since there are possibly millions of them. They also don't want to waste their time looking through irrelevant documents. Legal search, on the other hand, is typically referred to as a situation when recollection is critical. Users are more likely to put up with non-relevant papers when there is just one piece of missing evidence that might make or break the case (Deo, 2018).

Although a significant proportion of false positives are inevitable, this may be tolerated because of the need to locate all the evidence. To present a case in court, all accessible evidence must be recognized, which necessitates maximizing memory capacity in order to recollect as much as possible. Both criminal and civil cases might be weakened by a lack of proof. One of the primary benefits of using IR methods is that they aid in the accomplishment of this goal. As soon as the first processing step is complete, queries may be performed very fast. In order to get a high level of recall, it is probable that several queries will need to be conducted. In searches related to their passions, less than a quarter of people use the same terms again. Synonym matching and query expansion, two well-established methods of IR, may help improve recollection without needing investigators to manually input every conceivable query (Joby, 2020).

Researchers in the text retrieval and analysis fields are making significant strides in automating some of the tedious tasks that digital forensics investigators must do by hand. Digital evidence may be used sooner in investigations if the evidence backlog is reduced because of this action being taken. It is arguable that accuracy is more suitable at the early stages of an inquiry than the typical concentration on memory for IR forensics. New investigative methods, such as named entity extraction, association rule mining, temporal information extraction, and others, will have a foundation in forensics as it moves towards cloud computing and associated technologies.

**Design**

Document and Query Indexing, Query Evaluation, and System Evaluation are the most important aspects of Information Retrieval (IR). With the primary objective of finding and constructing an internal representation, Document and Query Indexing may accomplish this. A computer's ability to alter the data is one of the most important considerations. It is important to understand how documents and queries are represented in retrieval models and how a score is calculated for each. Uncertainty and ambiguity in information systems are addressed through Information Retrieval (IR). The user's request for information is not specific enough, whether it is in the form of a question, feedback, or actual action. System evaluation emphasizes the necessity of assessing the influence of information offered to users on their success in the system. In this section, we examine whether the efficiency of a given system is linked to time or space (Zanganeh, 2018).

**Implementation**

Web search engines have made information retrieval an everyday activity for millions of individuals, not only libraries and professional searchers. Information retrieval is often considered to be the most common method of accessing information. Information retrieval systems (IR systems) aid users in locating the information they need, but they do not directly provide solutions to questions. It alerts the user to the presence and location of documents that may contain the requested information. User assistance in browsing or filtering a document collection, as well as in processing a set of documents obtained, is also provided by information retrieval systems. Millions of computers throughout the world are used to look through billions of documents. It is possible to categories incoming emails using a spam filter, either manually or automatically, in an email application. Here we are using the following queries,

Computer Science, President of the university, Information Retrieval, Learner Data Institute, International student office, Graduate school admissions, what is the mascot of the University of Memphis? College of Arts and Sciences Dean, Dunn Hall, to be or not to be.

When the user requests information, she is submitting a query. It's possible for a search to include more than one phrase. However, terms may also be phrases since they are lexical items. It takes a long time and is inefficient to look through all of the documents. On the contrary, each document is pre-indexed. In the index, the material is presented in a logical and sequential order. Using this approach, a search is significantly more efficient. As a result, the retrieval function is aided by a comparable query representation. Each index phrase is linked to a collection of documents where it appears in practice, using an inverted index (Büttcher, 2016).

**Results**

In this proposed work of Retrieval Engine for the University of Memphis, some queries are created that make a keyword-based search for the retrieval of information. In an IR system, this is a very common and efficient way where a user finds or retrieves documents by providing some combinations of keywords as text input. With the help of a logical AND operator, this keyword-based search is created. For a particular system, we are going to execute our queries to find out the matrices that are related to the IR system.

Here we have considered Retrieval Engine for the University of Memphis as the whole set of documents where we will implement the whole work. The results are like the following way for the selected queries made over the whole set of documents.

Let us consider there are 1000 documents collections are there in the Retrieval Engine for the University of Memphis

**Query1: Computer Science**

Total document collection = 1000

There are 15 documents retrieved in the answer set, 5 relevant documents, and 25 relevant documents

So, the precision for Query1 will be = 5/25 = 20%

Recall of the Query1 will be = 5/15 = 33.33%

**Query2: President of the university**

Total document collection = 1000

There are 10 documents retrieved in the answer set, 2 relevant documents, and 20 relevant documents

So, the precision for Query2 will be = 2/20 = 10%

Recall of the Query2 will be = 2/10 = 20%

**Query3: Information Retrieval**

Total document collection = 1000

There are 30 documents retrieved in the answer set, 3 relevant documents, and 27 relevant documents

So, the precision for Query3 will be = 3/27 = 11.11%

Recall of the Query3 will be = 3/30 = 10%

**Query4: Learner Data Institute**

Total document collection = 1000

There are 15 documents retrieved in the answer set, 5 relevant documents, and 25 relevant documents

So, the precision for Query4 will be = 5/25 = 20%

Recall of the Query4 will be = 5/15 = 33.33%

**Query5: International student office**

Total document collection = 1000

There are 10 documents retrieved in the answer set, 2 relevant documents, and 20 relevant documents

So, the precision for Query5 will be = 2/20 = 10%

Recall of the Query5 will be = 2/10 = 20%

**Query6: Graduate school admissions**

Total document collection = 1000

There are 30 documents retrieved in the answer set, 3 relevant documents, and 27 relevant documents

So, the precision for Query6 will be = 3/27 = 11.11%

Recall of the Query6 will be = 3/30 = 10%

**Query7: What is the mascot of the University of Memphis?**

Total document collection = 1000

There are 20 documents retrieved in the answer set, 5 relevant documents, and 50 relevant documents

So, the precision for Query7 will be = 5/50 = 10%

Recall of the Query7 will be = 5/20 = 25%

**Query8: College of Arts and Sciences Dean**

Total document collection = 1000

There are 12 documents retrieved in the answer set, 4 relevant documents, and 44 relevant documents

So, the precision for Query8 will be = 4/44 = 9.09%

Recall of the Query8 will be = 4/12 = 33.33%

**Query9: Dunn Hall**

Total document collection = 1000

There are 10 documents retrieved in the answer set, 2 relevant documents, and 20 relevant documents

So, the precision for Query9 will be = 2/20 = 10%

Recall of the Query9 will be = 2/10 = 20%

**Query10: to be or not to be**

Total document collection = 1000

There are 30 documents retrieved in the answer set, 3 relevant documents, and 27 relevant documents

So, the precision for Query10 will be = 3/27 = 11.11%

Recall of the Query10 will be = 3/30 = 10%

**Future Work**

Here we have calculated the matrices of the queries but there is a huge scope of work that can we perform in the future. We can add some more documents so that the system is enriched with documents. Thus, the performance gets increases and the huge possibility to have more accuracy while finding some keywords in the Retrieval Engine for the University of Memphis.

**References**

Berger, A., & Lafferty, J. (2017, August). Information retrieval as statistical translation. In *ACM SIGIR Forum* (Vol. 51, No. 2, pp. 219-226). New York, NY, USA: ACM.

Büttcher, S., Clarke, C. L., & Cormack, G. V. (2016). *Information retrieval: Implementing and evaluating search engines*. MIT Press.

Deo, A., Gangrade, J., & Gangrade, S. (2018). A SURVEY PAPER ON INFORMATION RETRIEVAL SYSTEMS. *International Journal of Advanced Research in Computer Science*, *9*(1).

Joby, P. P. (2020). Expedient information retrieval system for web pages using natural language modeling. *Journal of Artificial Intelligence*, *2*(02), 100-110.

Zanganeh, M. Y., & Hariri, N. (2018). The role of emotional aspects in the information retrieval from the web. *Online Information Review*.