INFORMATION RETRIEVAL for The Web: Techniques and Challenges

Kavya Kumar Vallurupalli

Abstract:

Information Retrieval is a growing field that has touched every sub field in computer science . Be it data science to web development , information retrieval is required . But how and why is it important to the ever-growing WORLD WIDE WEB ?This paper aims at introducing Information Retrieval , a contrast to how it is different for the web , the techniques used for information retrieval and the challenges that exist for an efficient retrieval.

Introduction

The world wide web is the standard foundation of information across the world. This plethora of information it offers can be cumbersome to sort through . Access to relevant information is important. As of now there are many tools for performing a search that is fairly effective .But the amount of relevant information returned from those searches is less. Also, the order of return of information requires a lot of time to sort through the returned information before reaching the relevant information.

Hence, the efficiency of information retrieval is directly proportional to the potential of the web .[1]. There have been various techniques that came up for this purpose. Starting with navigational strategies for searching the web to evaluation methods for presenting the information on the web and various model to retrieve it . Each technique has its own parameter and focal point . Examination of these techniques allows one to determine the most efficient way to retrieve the information they are looking for and how to ensure an efficient information retrieval .

Techniques for Information Retrieval

WEB ROBOT :

A web robot is a software program that allows a user to put in a query . The query is then used to parameterize the documents to be retrieved . These documents are returned in a prioritized order depending on the query. The main issue faced by this program is the hugeness of the documents .

Along with the amount of time required to sort through the documents , the amount of data that needs to be sorted through reduces the feasibility of this method. But a web robot can be used to a searchable index of the documents . The search can be done using this index thus reducing both the time taken and the amount of information to be sorted through . Once an efficient index has been created , the graph traversal algorithm can be used to traverse through the documents since they are structured like a directed graph.

Search Tools and Search Services :

Automated methods of information retrieval make up the major part for search tools and search services. The web robot makes the indexes. The search tool uses these indexes to retrieve relevant information from the user . The search services hide the functionality of the search tool and the database from the users. Different search tools and search services follow the same base but incorporate different techniques for indexing and information retrieval .

Some of the search tools and their functionalities are as follows :-

AltaVista:

Its uses a spider to index the full text documents and updates the index at least once a day . If the query appears in the first few lines of the documents, it is deemed as a relevant document .

HotBot:

It uses Slurp for indexing documents. The index is updated based on the HTML and Meta-Data. The specialty of HotBot is it supports parallel searching process since it has indexes distributed across servers. The relevance calculation is based on factors such as document length and frequency . If the query appears in the Meta-Data or in the title the document is tagged as high relevance .

InfoSeek Guide :

It retrieves the HTML and PDF documents .It supports searches for symbols , phrases and captions for images. The relevance calculation of this method is if the query appears in the first few lines of the document.

Lycos:

It retrieves all the documents that contain full or partial terms given in the query . The user can select the degree of match. The relevance calculation of this method is the weights of the matched terms in the document.

WebCrawler:

It has a list of URL's and web servers . It uses the round-robin logic to prevent parsing already parsed files for the same server. The relevance calculation of this method is done by the regularity in the documents.

Google:

The ranking is done by the pages linked to the index that is requests by the query. Each link to a page has a vote and the page with the most votes is given the maximum relevance.

The search services pass the given user query to various search engines and retrieve the relevant results simultaneously . The duplicated data is removed, and the results are presented to the user .

The Web Search Engine Watch is a rating service of all the web search engines. They are evaluated based on size, pages crawled per day, freshness and depth of the information that is retrieved.

Evaluation of Search Engines.

Manning and Shuzte [3] have stipulated many quantitative measurements to measure the performance of a classical information retrieval system. But these measurements cannot apply to the web search retrievals. For example, interactive response times are more important to users to not have to access the second page or the scroll down option to get information pertaining to their queries.

There are two parameters that can judge the effectiveness of the search engine Recall ratio is the number of relevant documents retrieved to the total number of relevant documents.

$$recall = \frac{number\ of\ relevant, retrieved\ documents}{total\ number\ of\ relevant\ documents}$$

$$precision = \frac{number\ of\ relevant\ documents}{number\ of\ retrieved\ documents}$$

The higher the recall ratio the more efficient is the retrieval of information. Precision is the ratio of the total number of relevant documents retrieved to the total documents in the system . Maintaining a higher precision ratio and recall ratio allows the system to be more efficient . Combining it with the web robot allows the information retrieval to be effective.[2]

But web users rely more on the precision of the results displayed on the first page rather than the traditional measure of precision. Only the highest valued pages are given importance. [4]

A successful search engine incorporates new algorithms specifically designed for fast and accurate retrieval of valuable information . Some of the other aspects are attractive interfaces and free access time .

Improving the effectiveness of the retrieval of information.

Most focus is put on query processing speed and database size. A new feature, the META TAG gives an insight as to what the document is about. Since a query returns various relevant documents , it becomes hard for the user to navigate through them . Hence another main area of focus is to return a short number of documents that contain the relevant data .

Relevance Feedback Techniques

Relevance feedback techniques is a way to return results to a query , take the user feedback and use the information to check if the results are relevant to perform a new query .

There are three main types of feedbacks :

Explicit Feedback :

The assessors of relevance indicated the relevance of a document retrieved for a query. This type of feedback is an explicit feedback . Users use the graded or binary relevance system to indicate the relevance explicitly . Binary relevance is a simple relevant or not relevant system while the graded relevance feedback indicates the relevance of the document to the query using a grading system that can be made of numbers or letters.

Implicit Feedback:

User behavior is assessed . The number of times a document is opened, the time it is open for ,page browsing and scrolling actions are monitored with respect to the query .Since the user is not informed that their actions are used for feedback this type of feedback is called the implicit feedback.

Blind Feedback:

The manual part of the relevance feedback is automated, and the user gets improved retrieval performance without extended interaction. The results returned by the initial query (usually top k relevant results) are taken into consideration. The top common 20-30 terms are taken from these documents. A query expansion is done to include these terms . The results that have been returned from the new query are then matched to the relevant results of the previous query . The pseudo relevance feedback works better than global analysis. Query expansion enhances the results by including previously missed relevant documents. While assessors are not required this automated process may not be able to deal with query drift effectively.

Term Frequency :

The term frequency is a numerical statistic that reflects how important a word is to a document . It is a weighting factor in information retrieval. The tf-idf increases proportionally to the number of times a word appears in a document. On of the most widely used ranking systems is computed by summing the tf-idf for each query term .The documents are ranked based on the weight of query terms present in that document.

Inward Link:

An inward link helps to determine which web pages are added to the collection of relevant documents and how to order them .

Conclusion

The applications of Information Retrieval are widely studied . But information retrieval is the base of web search. Although there has been significant debate that information retrieval techniques are unable to keep up with the growing World Wide Web, visions and research into this aspect point towards coming up with concrete and user-friendly solutions. The main drawbacks that are being faced are the performance issues and answer explanation i.e. how well does a document satisfy being classified in a category. While quality of indexing has made a significant impact on web information retrieval , an effective algorithm for information retrieval is also needed to meet the current challenges of information retrieval in the web. Future work can incorporate the building of a system that uses all the effective information retrieval techniques while prioritizing the evaluations of these techniques by considering different dimensions .

References

[1]V.N. Gudivada; "Information Retrieval in the World Wide Web"; IEEE Internet Computing";1997

[2]Robng, Tane, Srivastava Jaideep, and Nert, Cooley Pang. "WEBSIFT: The Website Information Filter SYstem." Supported by NSF, ARL, 1999.

[3]Chistopher Manning, et al. “Introduction to Information Retrieval Systems.” Information Storage and Retrieval Systems The Information Retrieval Series, pp. 1–25., doi:10.1007/0-306-47031-4_1.

[4]venkat, N Gudivada, Raghavan Vijay V, Grosky William L, and Kasanagothi Rajesh. "Information Retreival on the World Wide Web." IEEE Internet Computing, 1999.

[5] Lewandowski, Dirk Web Information Retrieval. *Information Wissenschaft & Praxis*, 2005, vol. 56, n. 1, pp. 5-12. [Journal article (Paginated)]

[6]Monika Henzinger . “Link Analysis in Web Information Retrival.”

[7] Allen, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, W. B., ... Zhai, B. C. (2003). Challenges in Information Retrieval and Language Modeling. *SIGIR forum*, *37*(1), 31-47. [10.1145/945546.945549]. DOI: 10.1145/945546.945549