IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114108

WEB SEARCH AND INFORMATION RETRIEVAL

Shrutika Doiphode, Sheetal A. Wadhai

Department of Computer Engineering, Universal College of Engineering and Research, Pune

Abstract: The Internet is one of the main sources of information for millions of people. One can find information related to practically all matters on the internet. Moreover, if we want to retrieve information about some particular topic, we may find thousands of Web Pages related to that topic. But our main concern is to find relevant Web Pages from among that collection. So, in this paper, I have discussed how information is retrieved from the web and the efforts required for retrieving this information in terms of system and users' efforts.

Keywords: Information retrieval, Page ranking, Evaluation of information retrieval system.

1. INTRODUCTION:

The Web has undergone exponential growth in the past few years. It has been estimated that there are approximately 15-20 billion pages present on the Web and recently this count has hit the mark of 1 trillion. According to the studies, only 80-85 % of the total Web pages that are available on the Web give useful information and the remaining 20-15% are mostly duplicates of the original pages or near-duplicates and some of them are completely irrelevant pages. Thus, the Web explosion offers lots of new problems for the information retrieval systems. These information retrieval systems help users complete the search tasks, by finding a handful of relevant documents among thousands and thousands of pages of text with little structural organization. At the same time, developers of retrieval systems must be able to evaluate the overall effectiveness of these systems i.e., the relevance of results it retrieves in response to a user query.

2. INFORMATION RETRIEVAL ON THE WEB

Information Retrieval on the Web has always been a different and difficult task as compared with a classical information retrieval system (Library System). To explain the difference between classical information retrieval and information retrieval on the Web we compare the two. The differences can be partitioned into two parts, namely differences in the documents and differences in the users.

We first discuss the differences in the documents.

- **Hypertext:** Documents present on the web are different from general text-only documents because of the presence of hyperlinks. It is estimated that there are roughly 10 hyperlinks present per document.

- **Heterogeneity of document:** The contents present on a web page are heterogeneous i.e., in addition to the text they might contain other multimedia contents like audio, video, and images.

- **Duplication:** On the Web, over 20% of the documents present are either near or exact duplicates of other documents and this estimation have not included the semantic duplicates yet.

- **Several documents:** The size of the Web has grown exponentially over the past few years. The collection of documents is over trillions and this collection is much larger than any collection of documents processed by an information retrieval system. According to estimation, the Web currently grows by 10% per month.

- **Lack of stability:** Web pages lack stability in the sense that the contents of Web pages are modified frequently. Moreover, any person using the internet can create a Web page even if it contains authentic information or not.

The users on the Web behave differently than the users of the classical information retrieval systems. The users of the latter are mostly trained librarians whereas the range of Web users varies from a layman to a technically sound people. Typical user behavior shows:

- **Poor queries:** Most of the queries submitted by users are usually short and lack useful keywords that may help in the retrieval of relevant information.

- **Reaction to results:** Usually users don't evaluate all the result screens, they restrict themselves to only results displayed on the first result screen.

- **Heterogeneity of users:** There is a wide variance in education and Web experience between Web users.

Thus, the main challenge of information retrieval on the Web is how to meet the user needs to give the heterogeneity of the Web pages and the poorly made queries.

3. IR (INFORMATION RETRIEVAL) TOOLS ON THE WEB

Information from the Web can be retrieved by several tools available ranging from General Purpose Search Engines to Specialized Search Engines. Following are the most used Web IR tools:

- **General-Purpose Search Engine:** They are the most used tool for information retrieval. Google, AltaVista, and Excite are some of the examples. Each of them has its own set of Web pages which they search to answer a query.

- **Hierarchical directories:** In this approach, the user is required to choose one of a given set of categories at each level to get to the next level. For example, Yahoo! or the Dmoz open directory project

- **Specialized Search Engines:** These Search Engines are specialized in an area and provide a huge collection of documents related to that specific area. E.g., PubMed, a Search Engine that specialized in medical publications. It offers a collection of millions of research papers, articles; journals related to biomedical sciences, life sciences, etc.

4. GENERAL PURPOSE SEARCH ENGINE:

General Purpose Search Engines are used to index a sizeable portion of the Web across all topics and domains to retrieve the information. Each such Engine consists of three major components:

- A spider or crawler browses the Web by starting with a list of URLs called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks on the page and adds them to the list of URLs that are visited recursively to form a huge collection of documents called a corpus. The corpus is typically augmented with pages obtained from direct submissions to search engines and various other sources. Each crawler has different policies concerning which links are followed, how deep various sites are explored, etc. As a result, there is surprisingly little correlation among corpora of various engines.
- The indexer processes the data and represents it usually in the form of fully inverted files. However, each major Search Engine uses different representation schemes and has different policies concerning which words are indexed.
- The query processor processes the input query and returns matching answers, in an order determined by a ranking algorithm. It consists of a front end that transforms the input and brings it to a standard format and a back end that finds the matching documents and ranks them.

4.1 A Brief History of Search Engines

They have evolved a lot since their inception. This evolution witnessed three major generations; each generation considered its approach to retrieving relevant documents. Following are the three main generations:

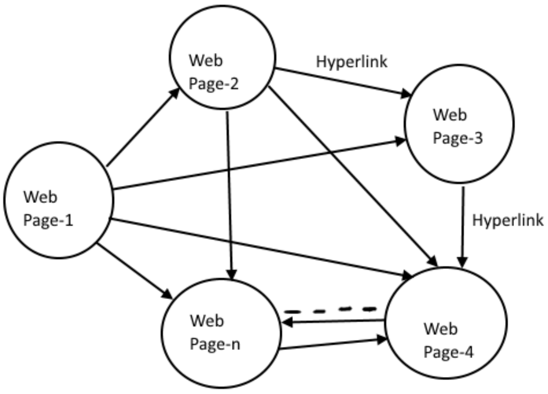
- **1st Generation:** This generation came around 1996. It searches ranked sites based on-page content. Documents are treated as a collection of words and no importance is given to the semantics of the documents. The main disadvantage of this generation was that any document can be made relevant by keyword stuffing to increase the content similarity examples are Excite, Alta Vista, and InfoSec.
- **2nd Generation:** This generation relies on contents and as well as on-link analysis for ranking- so they take the structure of the Web as a graph into account. It considers site popularity as the criteria for ranking the document as relevant. But this approach has its flaws spammers can create link farms i.e., heavily interconnected sites which may make any document or page of lesser importance more important. For example, Lycos.
- **3rd Generation:** Apart from page contents and web structure this generation considers page reputation as one of the major criteria. According to this approach if a page is referred by a highly reputed page, then it has more relevance, and more links to a page mean that the page has a high reputation. Examples of 3rd generation search engines are Google and Yahoo! from the above discussion we inferred that the main task of a search engine is to retrieve information for a user query. To make this retrieval more relevant number of approaches are used as discussed above. But the best and universally accepted approach is to rank a page according to its relevance, this approach called Page Rank is discussed below.

4.2 Ranking is used to order the answer to a query in decreasing order of value. For this, a numerical value called score is assigned to each document and the documents are arranged in the decreasing order of the score. This score is typically a combination of two criteria query-independent and query dependent criteria. A query-independent criterion assigns an

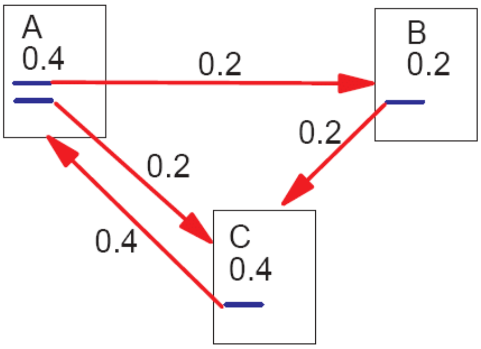
DOI: 10.17148/IJARCCE.2022.114108

intrinsic value to a document, regardless of the actual query by considering the publication data (like the site to which it belongs, the date of the last change, etc.), the number of citations (in degree), etc. A query-dependent criterion is a score that is determined only concerning a particular query.

4.3 Graph Structure of Web Before we study the details of each criterion, we must represent the Web as a directed graph [10], where each node represents a page and any link from one page to another page represents an edge i.e., if a page u contains a hyperlink for page v, then that link is represented by a directed edge (u, v). Every page on the web has several forward links called out edges and some number of backlinks called in edges. The number of out edges can be easily found by considering all the hyperlinks present on that page, but it is difficult to find all the in edges to a page i.e., to find all the pages pointing to that page. For example, in figure 1 page B has two backlinks.



Query-independent ranking criterion: According to this criterion if a web page has a larger number of hyperlinks pointing to it (also called links) then it is considered a better page. The main drawback of this criterion is that each link is equally weighted. Thus, it cannot distinguish the quality of a page that gets pointed to by low-quality pages from the quality of a page that gets pointed to by high-quality pages. It is easier to make a page appear to be high-quality- just creates many other pages that point to it. To remedy this problem, Brin invented the Page Rank measure. Page Rank is defined as follows:



Query-dependent ranking criterion: It was developed by Kleinberg. It is described as follows: Forgiven a user query, the algorithm first constructs a graph specific to that query which is a subgraph of the main graph representing the Web. In this query-specific graph, nodes represent the pages and edges represent the hyperlink. For each page, two types of scores are calculated: Authority Score and Hub Score. If a Web page has more relevant content, then its authority score is more and if a Web page contains hyperlinks to relevant pages, then it has more hub score.

4.4 Duplicate Filtering Experiments indicate that over 20% of the publicly available documents on the Web are duplicates or near-duplicates. There is a need to adopt some approaches to find these duplicate documents, as discussed, we can calculate the resemblance among Web pages in terms of a set intersection problem. The reduction to a set intersection problem is done via a process called shingling. In this, each document is viewed as a sequence of tokens. We can take tokens to be letters, words, or lines. We assume that we have a parser program that takes an arbitrary document and reduces it to a canonical sequence of tokens. ―Canonical here means that any two documents that differ only in formatting or other information that we chose to ignore, for instance

DOI: 10.17148/IJARCCE.2022.114108

5 TEST COLLECTION:

Before starting the evaluation of an information retrieval system we need to understand that a user uses these systems for retrieval tasks like he may want to find all relevant documents for a query, to filter the relevant documents from the retrieved result set, etc. All these retrieval tasks are done from a vast collection of documents called test collection. A test collection encapsulates the experimental environment. It is meant to model users with information needs that are instances or examples of the task. These information needs are generally treated as if they do not change over time; if they are representative of the needs of users of the system in general, then showing that a system can perform well on them suggests that a system will perform well.

Test collections have three components:

- A corpus of documents to search.
- A set of user information needs.
- Judgement of the relevance of information needs to document in the corpus.

5.1 Relevance Judgement: The relevance judgments tell us which documents are relevant to each of the information needs. As described above, since it is people that will be using the documents, relevance is something that must be determined by people. The system itself can only try to predict relevance; an evaluation determines how good the system is at predicting what will be relevant, and an experiment tells us whether one system is better at it than another. Once the topics have been finalized, human assessors can start judging documents for relevance. Assessors read documents, compare them to the topic definition, and say whether they are relevant or not (or possibly how relevant they are).

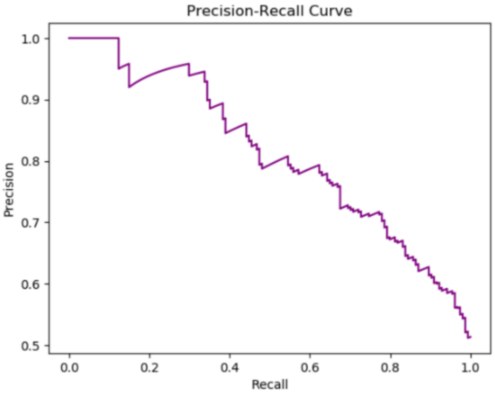
Exhaustively judging relevance—that is, judging every single document in the corpus to every single topic—is the only way to guarantee that all relevant documents are known. This is often impossible due to time and budget constraints, however. One assessor judging a million documents at a relatively quick rate of 10 per minute would take over ten months of 40-hour weeks to complete just one topic.

Focusing judgment effort on a small portion of the complete corpus can usually provide enough of the relevant documents for most evaluation and experimentation purposes. One simple approach is the pooling method: each topic in the collection is submitted to a variety of different retrieval systems, and the top N-ranked documents from all those systems are pooled for judging.

5.2 Evaluation Measures: Once a test collection has been finalized, at any time someone may submit a query derived from one of its topics to a retrieval system, obtain the ranked list of retrieved documents, and measure the system’s effectiveness using the relevance judgments for that topic. The IR literature is awash with different evaluation measures meant to measure different aspects of retrieval performance; we will focus on a few of the most widely used.

5.4.1.1 Precision – Recall: Curve Plotting recall and precision over a series of rank cut-offs produce the precision-recall curve. To understand the behavior of the precision-recall curve, we calculate the value of precision and recall at different ranks. For example, consider the above-mentioned case in which the system retrieves 10 documents. Suppose instead of 10 documents our system retrieves 50 documents out of which 20 are relevant, then precision =20/50 =0.4 and recall =20/162 ≈ 0.05. Here we see as the rank increases the value of precision decreases and the value of recall increases this is because of the increase in at he the number to f he retrieved documents. Using raw values of precision and recall at every possible rank cut-off produces a jagged curve like the one shown in Figure 3. This jagged curve represents that recall can never decrease with rank cutoff, while precision increases with every increase in recall and decreases while recall stays constant. To produce a smoother curve, we use a technique called interpolation. Interpolated precision is defined by a value of recall rather than by a rank cut-off; specifically, for a given recall level r, interpolated precision at r is defined to be the maximum measured precision at any rank cut-off k at which recall is no less than. We formulate this as:

i-precision@r = max precision@k



5.4.2 Modelling User Effort: One factor of system performance that precision and recall-based measures do not directly address is the amount of effort a user can be expected to put in while interacting with the system. There are various families of measures that attempt to address this; the most used is the discounted cumulative gain (DCG) family. 5.4.2.1 Discounted Cumulative Gain Family Discounted cumulative gain (DCG) is defined by a gain function and a discount function. The gain function tells us the value of a particular relevant document to a user, allowing DCG to take advantage of grades of relevance. For instance, relevance judgments may be made on a three-point scale (not relevant, relevant, highly relevant) or a five-point scale (poor, fair, good, excellent, perfect); DCG's gain function can take advantage of these grades by mapping them to numeric values to reflect their utility to a user. Traditional precision and recall can only use binary judgments. Two typical gain functions are the linear and exponential functions. Linear gain simply assigns incrementally increasing values to each relevance grade, e.g., nonrelevant→0, relevant→1, highly relevant →2. Exponential gain multiplicatively increases values, e.g., poor→0, fair→1, good→3, excellent→7, perfect→15. By tuning the gain function, a developer can model users that have varying degrees of preference for different grades of relevance. The discount function reflects the patience a user has for proceeding down the ranked list. It is assumed that as the rank increases the gain function is likely to increase and discounts never increase or increase by a small margin.

6. CONCLUSION:

In this paper, we discussed Web information Retrieval methods and tools that take advantage of the Web particularities to mitigate some of the difficulties that Web information retrieval encounters. To quantify the results of Information Retrieval we used evaluation measures like Precision and Recall and studied how to calculate them effectively. Since the degree of effectiveness greatly depends on the user's effort so we discussed how to model the user's effort using the gain function and discount function of DCG (Discount Cumulative Gain Family). Effectiveness evaluation is an important aspect of research and design of information retrieval systems. Much research has been done on the topic, and more continues to appear every year. The issue of cost-effective relevance judging and evaluation remains important. Interest in devising user models for evaluations that go beyond individual, independent document relevance has recently increased; ongoing work in novelty and diversity is investigating the tradeoffs between the relevance of documents and the redundancy of relevant information within the documents.

7. FUTURE SCOPE:

The present Information Retrieval Systems are effective enough to retrieve the relevant pages but still there are some open problems that we discussed like whether these pages are the result of an exhaustive search from the Web, how to uniformly sample Web Pages on a Web Site if one does not have a complete list of Web Pages. Also, we know lots of resources are wasted (memory and time) for dealing with duplicate pages so while finding the duplicate pages we also need to work on finding the pages which are semantic duplicates of each other.

REFERENCES

[1] James M. Abello, Panos M. Pardalos, Mauricio G. C. Resende —Algorithmic Aspects of Information Retrieval on the Web‖ in —Handbook of Massive Data Sets‖, Kluwer Academic Publisher, 2002, pp 3-10

[2] Sergey Brin and Lawrence Page —The Anatomy of a Large Scale Hypertextual Web Search Engine‖, in Proceedings of World-Wide Web'98.

[3] J.Klenberg, —Authoritative sources in a hyperlinked environment‖ Proc. ACM-SIAM Symposium on Discrete Algorithms (1998).

[4] A. Z. Broder (1997). On the resemblance and containment of documents. In Proceedings of Compression and Complexity of Sequences 1997, pp 21-29. IEEE Computer Society.

[5] J.Cho, H.Garcia-Molina, L.Page (1998). Efficient Crawling Through URL Ordering. In (WWW7, 1998) pp 161-172

[6] Monika R. Henzinger (2004). Algorithmic Challenges in Web Search Engines. Internet Mathematics Vol. 1, No. 1: 115-126.

[7] K. Bharat and A.Z. Broder (1998). A technique for measuring the relative size and overlap of public web search engines. In (WWW7, 1998) pp 379-388