Journal of Theoretical and Applied Information Technology





www.jatit.org

# A PRIMER ON THE WEB INFORMATION RETRIEVAL PARADIGM

**[1]MPS Bhatia, [2]Akshi Kumar Khalid**

Netaji Subhas Institute of Technology, University of Delhi

E-mail: [1]mpsbhatia@nsit.ac.in, [2]akshi.kumar@gmail.com

**ABSTRACT**

The unabated growth of the Web and the increasing expectation placed by the user on the search engine to anticipate and infer his/her information needs and provide relevant results has fostered the development of the field of Web Information Retrieval (Web IR). The recent surveys claim that 85% of internet users use search engines and search services to find specific information [1]. The same surveys, however, show that users are not satisfied with the performance of the current generation search engines. The slow retrieval speed, poor quality of retrieved results, handling a huge quantity of information, addressing subjective & time-varying search needs, finding fresh information and dealing with poor quality queries are commonly cited glitches. This paper expounds the Web Information Retrieval paradigm, a variant of classical Information Retrieval, by illustrating its basics, the components, model categories, tools, tasks and the performance measures that quantify the quality of retrieval results.

**Keywords:** *Web Information Retrieval, Web IR tasks, Web IR models, Web Tools, Performance Measures*

## 1. INTRODUCTION: FROM IR TO WEB IR

By all measures, the Web is enormous and growing at a staggering rate, which has made it increasingly intricate and crucial for both people and programs to have quick and accurate access to Web information and services. Thus, it is imperative to provide users with tools for efficient and effective resource and knowledge discovery. Search engines have assumed a central role in the World Wide Web's infrastructure as its scale and impact have escalated. Although the web search engine assists resource discovery, it is far from satisfying for its poor precision and recall [1, 2]. The unabated growth of the Web and the increasing expectation placed by the user on the search engine to anticipate and infer his/her information needs and provide relevant results has fostered the development of the field of Web Information Retrieval (Web IR).

### 1.1. Web Information Retrieval (Web IR)

Web IR can be defined as the application of theories and methodologies from IR to the World Wide Web. It is concerned with addressing the technological challenges facing Information Retrieval (IR) in the setting of WWW [3]. The characteristics of Web make the task of retrieving information from it quite different from the Pre-Web (traditional) information retrieval. The Web is seemingly unlimited source of information with users from cross-section of society seeking to find information to satisfy their information need. They require the Web to be accessible through effective and efficient information retrieval systems that deliver information need fulfillments through the retrieval of Web content.

Web IR is different from classical IR for two kinds of reasons: concepts and technologies [4]. The following characteristics of the Web shape up the nature of Web Information Retrieval and are what make it considerably different to traditional retrieval challenges:

- *The "Abundance" of Web:* With the phenomenal growth of the Web, there is an ever increasing volume of data and information published in numerous Web pages. According to worldwidewebsize.com, the indexed Web contains at least 27.87 billion pages (Sunday, 22 June, 2008).

- *Heterogeneity*
  - Information /data of almost all types exist on the Web, e.g., structured tables, texts, multimedia data, etc.
  - Much of the Web information is semi-structured due to the nested structure of HTML code.
  - Much of the Web information is linked
  - Much of the Web information is redundant
  - The Web is noisy: A Web page typically contains a mixture of many kinds of information, e.g., main contents, advertisement, navigational panels, copyright notices.

- *Dynamics:* The freedom for anyone to publish information on the web at anytime and anywhere implies that information on the Web is constantly changing. It is a dynamic information environment whereas traditional systems are typically based on static document collection.

- *Duplication:* Several studies indicate that nearly 30% of the web's content is duplicated, mainly due to mirroring.

- *Users Search Behavior*: The users have different expectations and goals such as Informative, Transactional and Navigational. Often they compose short, ill-defined queries and impatiently look for the results mainly in the top 10 results.

Despite the success of Web as a preferred or de-facto source of information, the retrieval of information from the Web is still an unsolved problem with many different applications probably undiscovered. Specifically, the operative challenges motivating researchers in Web IR include problems relating either to data quality or user satisfaction [1]. The problems facing successful Web Information Retrieval are a combination of challenges that stem from traditional information retrieval and challenges characterized by the nature of the World Wide Web.

The ultimate challenge of Web IR research is to provide improved systems that retrieve the most relevant information available on the web to better satisfy a user's information need.

## 2. THE WEB IR COMPONENTS

To address the challenges found in Web IR, Web search systems need very specialized architectures. Figure 1 shows an overview of the most important components of a typical Web Information Retrieval system. A Crawler is usually responsible for gathering the documents and storing them in document repository. The IR system includes the indexing and ranking functions. The Indexer distills information contained within corpus documents into a format which is amenable to quick access by the query processor. Typically this involves extracting document features by breaking-down documents into their constituent terms, extracting statistics relating to term presence within the documents and corpus, and calculating any query-independent evidence. Once the indices are built, the system is ready to process queries. The principal component of the query processor is the document ranking function. It determines what information is a good match to a user query & what information is inherently good.

The Web Search can be categorized into two phases, namely the *Offline phase* which includes the 'Crawling' & 'Indexing' Components; and the *Online phase* which includes the 'Querying' & 'Ranking' components of the Web IR system.

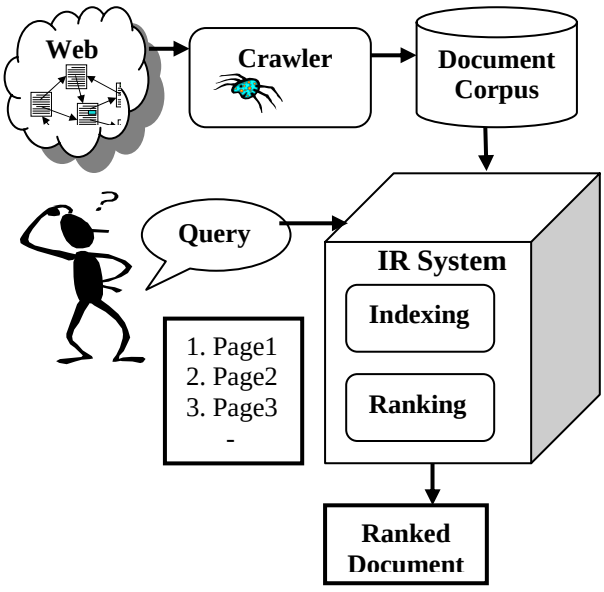


**Figure 1:** Overview of components of a typical Web Information Retrieval System

## 3. WEB IR MODELS

Retrieval models form the theoretical basis for computing the answer to a query [12]. A Retrieval Model is a formal representation of the process of matching a query and a document. The model of Web IR can be defined as a set of premises and an algorithm for ranking documents with regard to a user query. More formally, a Web IR model is a quadruple [D, Q, F, R ($q_i$,$d_j$)] where D is a set of logical views of documents, Q is a set of user queries, F is a framework for modeling documents and queries, and R($q_i$,$d_j$) is a ranking function which associates a numeric ranking to the query $q_i$ and the document $d_j$ [5]. The model is characterized by four parameters:

- Representations for documents and queries, which define the model
- Matching strategies for assessing the relevance of documents to a user query, which involves learning parameters from query.
- Methods for ranking query output, and
- Mechanisms for acquiring user-relevance feedback.

Retrieval models can describe the C*omputational process*, for example, how the documents are ranked and note that how documents or indexes are stored is implementation. The Retrieval models can also attempt to describe the *User process*, for example, the information need and interaction level. The Retrieval variables are usually depicted by queries, documents, terms, relevance judgments, users & information needs. They can have an explicit or implicit definition of relevance.

### 3.1. First Dimension: Computational Process: The Mathematical Basis

According to the first dimension, the models can be classed into three types: set theoretic, algebraic and probabilistic models. In the following sections, we describe instances of each type.

#### 3.1.1. Set theoretic models

Documents are represented by sets that contain terms. Similarities are derived using set-theoretic operations. Implementations of these models include the Standard Boolean Model, the Extended Boolean Model and the Fuzzy Model. The strict Boolean and fuzzy-set models are preferable to other models in terms of computational requirements, which are low in terms of both the disk space required for storing document representations and the algorithmic complexity of indexing and computing query-document similarities.

#### 3.1.2. Algebraic models

Documents are represented as vectors, matrices or tuples. These are transformed using algebraic operations to a one-dimensional similarity measure. Implementations include the Vector Space Model and the Generalized Vector Space Model. The strength of this model lies in its simplicity. Relevance feedback can be easily incorporated into it. However, the rich expressiveness of query specification inherent in the Boolean model is sacrificed.

#### 3.1.3. Probabilistic models

Document's relevance is interpreted as a probability. Documents and queries similarities are computed as probabilities for a given query. The *probabilistic model* takes these term dependencies and relationships into account and, in fact, specifies major parameters such as the weights of the query terms and the form of the query document similarity. Due to its simplicity and efficient computation, the Vector Model is the most widely used model in IR. The model requires term-occurrence probabilities in the relevant and irrelevant parts of the document collection, which are difficult to estimate. However, this model serves an important function for characterizing retrieval processes and provides a theoretical justification for practices previously used on an empirical basis (for example, the introduction of certain term-weighting systems).

### 3.2. Second Dimension: User Process: The Relevance Basis

Another dimension of defining different categories of Web IR models can be based on their applications as follows:

#### 3.2.1. Classical models

- Query languages, Indexing ( Boolean)
- Introducing ranking and weighting (Vector Space)

#### 3.2.2. Topical relevance models

- IR as Bayesian classification, relevance information, *tf.idf* weights (BM25)

- Probabilistic models of documents, queries, topics (Language Modeling)

### 3.2.3. User relevance models

- Combination of evidence, features, query language ( inference network, Inquery)

### 3.2.4. Linear feature-based models

- Learning weights, arbitrary features, optimizing effectiveness measures (Ranking SVM, Linear Discriminant, MRF)
- "Learning to Rank", learning ranking rather than classification, preferences

## 4. WEB IR TASKS

Web Information Retrieval research is typically organized in tasks with specific goals to be achieved. Existing tasks have changed frequently over the years due to the emergence of new fields. Below is a summary of the main tasks and also of the new or emerging ones:

- **Ad-Hoc Retrieval** Rank documents using non-constrained queries in a fixed collection. This is the standard retrieval task in Web IR.
- **Filtering** Select documents using a fixed query in a dynamic collection. For example, "Retrieve all documents related to 'Research in India' from a continuous feed".
- **Topic Distillation** Find short lists of good entry points to a broad topic. For example, "Find relevant pages on the topic of Indian History".
- **Homepage Finding** Find the URL of a named entity. For example, "Find the URL of the Indian High Commission homepage"
- **Adversarial Web IR** Develop methods to identify and address the problem of web spam, namely link spamming that affect the ranking of results.
- **Summarization** Produce a relevant summary of a single or multiple documents.
- **Visualization** Develop methods to present and interact with results.
- **Question Answering** Retrieve small snippets of text that contained an answer for open-domain or closed-domain questions.
- **Categorization/ Clustering** Grouping documents into pre-defined classes/ adaptive clusters.

## 5. WEB IR TOOLS

Automated methods for retrieving information on the Web can be broadly classed as search tools or search services.

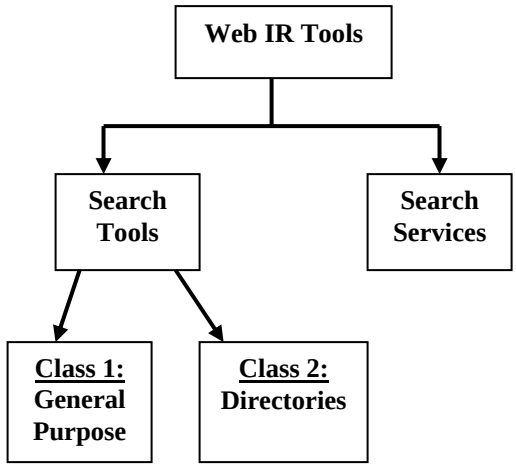


**Figure 2:** Classification of Web IR tools

### 5.1. Search Tools

The Search tools employ robots for indexing Web documents. They feature a user interface for specifying queries and browsing the results. At the heart of a search tool is the search engine, which is responsible for searching the index to retrieve documents relevant to a user query. Search tools can be distinguished into two categories on the transparency of the index to the user. The two class categories are depicted along the following dimensions:

- Methods for Web navigation,
- Indexing techniques,
- Query language or specification scheme for expressing user queries,
- Strategies for query-document matching, and
- Methods for presenting the query output.

#### 5.1.1. Class1 search tools: General Purpose Search Engine

These tools completely hide the organization and content of the index from the user. Example: AltaVista, Excite, Google, Infoseek, Lycos

#### 5.1.2. Class 2 search tools: Subject Directories

These feature a hierarchically organized subject catalog or directory of the Web, which is visible to users as they browse and search. Example: Yahoo!, WWW Virtual Library and Galaxy.

### 5.2. Search Services

The Search services provide users a layer of abstraction over several search tools and databases and aim at simplifying the Web search. Search services broadcast user queries to several search engines and various other information sources simultaneously. Then they merge the results submitted by these sources, check for duplicates, and present them to the user as an HTML page with clickable URLs. Example: MetaCrawler.

## 6. QUANTIFYING THE QUALITY OF WEB IR RESULTS

There are various ways to measure how well the retrieved information matches the intended information. The Web IR system might evaluate several aspects, namely, the assistance in formulating queries, the speed of retrieval, the resources required, the presentation of documents, the ability to find relevant documents, the appealing to users (market evaluation). The Evaluation is generally comparative.

In an Information Retrieval scenario, the most common evaluation is retrieval effectiveness and the effect of indexing exhaustivity and term specificity on retrieval effectiveness can be explained by two widely accepted measures Precision & Recall.

**Precision** is defined as the *number of relevant documents* retrieved by a search *divided by the total number of documents retrieved* by that search, and **Recall** is defined as the *number of relevant documents* retrieved by a search *divided by the total number of existing relevant documents* (which should have been retrieved). [7]

### 6.1. Precision

The proportion of retrieved and relevant documents to all the documents retrieved:

$$\frac{\text{Total Number of Relevant Retrieved Documents}}{\text{Total Number of Retrieved Documents}}$$

A perfect **Precision** score of 1.0 means that every result retrieved by a search was relevant (but says nothing about whether all relevant documents were retrieved)

### 6.2. Recall

The proportion of relevant documents that are retrieved, out of all relevant documents available:

$$\frac{\text{Total Number of Relevant Retrieved Documents}}{\text{Total Number of Relevant Documents}}$$

A perfect **Recall** score of 1.0 means that all relevant documents were retrieved by the search (but says nothing about how many irrelevant documents were also retrieved).

Both the measures are used to measure the accuracy of a system's ability to retrieve documents with respect to a given query. Ideally, you would like to achieve both high recall and high precision. In reality, you must strike a compromise. Indexing terms that are specific yields higher precision at the expense of recall. Indexing terms that are broad yields higher recall at the cost of precision. For this reason, an IR system's effectiveness is measured by the precision parameter at various recall levels.

### 6.3. Some Other Measures….

There are a number of more advanced and specific types of precision and recall measures that are used as modern evaluation measures. [9]

- *Fallout* is a measure of how quickly precision drops as recall is increased. Fallout is defined as the probability to find an irrelevant among the retrieved documents:

$$\frac{\text{Total Number of Irrelevant Retrieved Documents}}{\text{Total Number of Retrieved Documents}}$$

- R-*precision* is the precision at R where R is the number of relevant documents in the collection for the query. It is the precision after R retrieved documents, where R is the number of relevant documents that exists for that query. An R-precision of 1.0 is equivalent to perfect relevance ranking and perfect recall. However, a typical value of R-precision which is far below 1.0 does not indicate the actual value of recall (since some of the relevant documents may be present in the hit-list beyond point R).

- *Initial precision* is the precision at recall 0% in the interpolated precision-recall graph. It is an indication of relevance ranking of the top few hits. Similarly, one can define a *final precision* that is the precision at 100% recall. Final precision indicates how far down one need to go in the hit-list to find all relevant documents.

- *Precision at 0.5 Recall* is the precision after half the relevant documents have been retrieved.

- *Average Precision* is the average of precision scores at every relevant document in the retrieved set.

- *Recall (1000)* is the recall after 1000 retrieved documents. This is more practical than true recall over all documents since modern systems can return a huge number of results.

## 7. CONCLUSION

The field of Information Retrieval has become extremely important in recent years due to the intriguing challenges presented in tapping the Internet and the Web as an inexhaustible source of information. The success of Web search engines is a testimony to this fact. This paper confers the Web Information Retrieval paradigm, a variant of classical Information Retrieval, by expounding its basics, system components, model categories, and probing the Web IR tools, tasks and performance measures.

## 8. REFERENCES

[1] M. Kobayashi & K. Takeda, "Information Retrieval on the Web", ACM Computing Surveys, Vol. 32, No.2, June 2000.

[2] Monika Henzinger, "The Past, Present, and Future of Web Search Engines", Proceedings of 31st International Colloquium, ICALP 2004, Finland, July 12-16, 2004.

[3] Nunes Sérgio**, "**State of the Art in Web Information Retrieval**"**, Technical Report, FEUP, 2006.

[4] Monika Henzinger, "Information Retrieval on the Web", 39th Annual Symposium on Foundations of Computer Science (FOCS'98), Palo Alto, CA

[5] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, New York, 1999.

[6] V.N. Gudivada, V.V. Raghavan, W.I. Grosky, R. Kasanagottu, "Information retrieval on the World Wide Web"**,** Internet Computing, IEEE Volume 1, Issue 5, Sep/Oct 1997 Page(s):58 - 68

[7] Amit Singhal, "Modern Information Retrieval: A Brief Overview", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43, 2001.

[8] M.R. Henzinger, R. Motwani, C. Silverstein, "Challenges in Web Search Engines", Proceedings of the 18th International Joint Conference on Artificial Intelligence, 1573-1579, 2003.

[9] C.Buckley & E.Voorhees, "Evaluating Evaluation Measure Stability", Proceedings of the 23rd annual ACM SIGIR Conference on Research & Development in Information Retrieval, Athens, Greece, pp 33-40, 2000.

[10] Mehran Sahami, Vibhu Mittal, Shumeet Baluja, Henry Rowley, "The Happy Searcher: Challenges in Web Information Retrieval", PRICAI 2004: Trends in Artificial Intelligence, LNCS, Volume 3157, pp 3-12, 2004.

[11] James Allan et al., "Challenges in Information Retrieval and Language Modeling", Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002, SIGIR Forum 37(1): 31-47 (2003).

[12] Norbert Fuhr, "Models in Information Retrieval", ESSIR 2000: 21-50