INFORMATION RETRIEVAL FROM WEB – VARIOUS MODELS

[1]Dr. Mercy Paul Selvan

[1]Assistant Professor
[1] Department of Computer Science and Engineering,
[1]Sathyabama Institute of Science and Technology, Sholinganallur, India

Abstract : Information retrieval is a field related to the analysis, structure, storage, searching, organization, and retrieval of information. It includes variety of applications related to search and a wide range of types of information. In this paper a brief study about information retrieval, various models, issues are discussed.

***IndexTerms* -** Information Retrieval, Models of IR, Search Engine

I. Introduction

Since the 1950s, the primary focus of the field has been on text documents such as email, web pages, scholarly papers, books and news stories. All these documents have similar structure with title, author, date and abstract information associated with the content of papers in scientific journals. When referring to database records, the elements of this structure are called attributes or fields. Information in the document is relatively unstructured in the form of text. The user relates to the system directly for the Information retrieval as shown in Figure 1.3. They are easy to compare fields with well-defined semantics to queries in order to find matches. For example the records are easy to find for example bank database query. The semantics of the keywords also plays an important role which is send through the interface. System includes the databases, the interface of search engine servers and the indexing mechanism that comprises the stemming techniques. The user describes the search strategy and provides the requirement for searching .The documents present in the web put on ranking, subject indexing, and clustering. The relevant matches can be easily found by comparing it with the field values of records. It is simple for the database in terms of retrieval and maintenance of records, whereas it is complex for the unstructured documents where text is used (Baeza-Yatesand Ribeiro-Neto 1999). The phases involved in the IR process are illustrated in Figure 1.2.

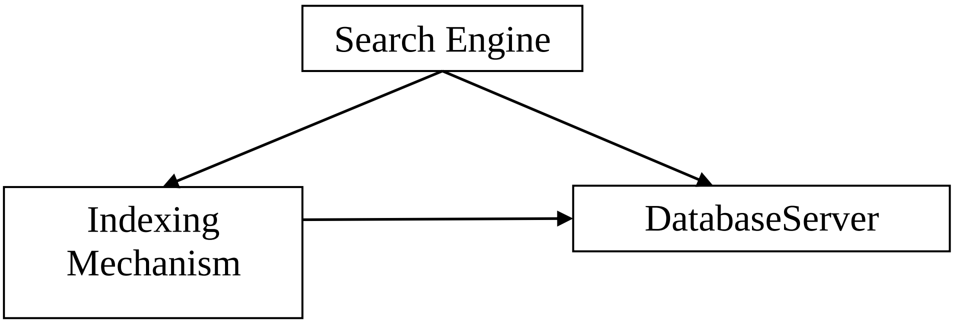

Figure 1.1 IR System Components

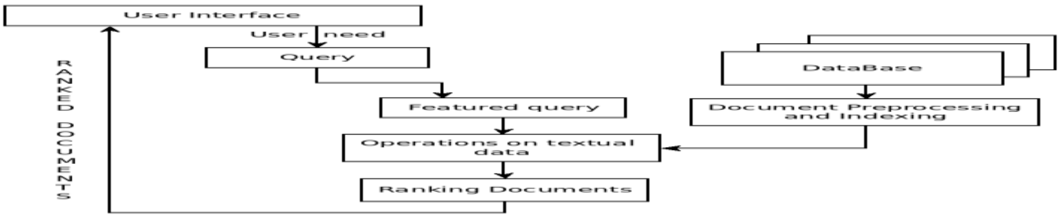

Figure 1.2 Phases of IR Process

The applications of information retrieval include significant text content, multimedia documents with structure, and other media applications. Popular information on media comprises pictures, audio, and video. It is used in Recommender systems. Its application spreads widely from digital libraries, Search engines etc.

1.1 Related Work:

[2] has designed a filtering tool which can extract the components of a web page such as text, images etc, based on a filtering condition. It eliminates the use of semantic specifications, offline parsing and compilation processes. It operates on the technique of information filtering which is nothing but removing all the unwanted information from the page. Initially, a tree data structure is formed containing all the elements of the web page. Then, this data structure is navigated to match the relevant nodes based on the filtering condition. Finally, the nodes except the matching nodes are deleted from the tree. But the tool bar needs to be extended for semantic queries and XML documents. [1] tells about the introduction to information retrieval. The Eigen Rumor

algorithm ranks each blog entry on basis of weighting the hub and authority scores of the bloggers based on eigenvector calculations. So, this algorithm enables a higher score to be assigned to a blog entry entered by a good blogger but not linked to by any other blogs based on acceptance of the blogger's prior work [3]. YoungDeokSeo et al., (2015) have proposed a new personal information retrieval system. In order to protect the personal or private data of users from exposing in Internet, this system retrieves private data from the web which are then blocked and removed. Based on the exposure degree of personal information, the web pages are ranked and the personal information is retrieved from the web. The system uses Google's PageRank for basic ranking of web pages [4]. [5] explains about modern retrieval system.

1.2 Information Retrieval (IR) Models

Information retrieval models that can be applied on any text collection are discussed in the following. Not all the IR models are easily scaled up to be able to deal with a very large collection, such as pages collected from the Web. The most important IR models are (Goker, A., and Davies, 2009)

- Boolean Model
- Vector Space Model
- Region Model
- Two poisson Model
- Bayesian network Model
- Probabilistic Model
- Language Model

1.2.1 Boolean model: It is the simplest to implement. A document is represented as a set of keywords. Queries are Boolean expressions of keywords, connected by AND, OR, and NOT, including the use of brackets to indicate the scope of these operators. The output of the system is a list of documents that are relevant, but there will be no partial matches or ranking. The Boolean model is very rigid: AND means "all"; OR means "any". All matched documents will be returned, making it difficult to control the number of documents retrieved. Figure 1.5 shows the Boolean model.

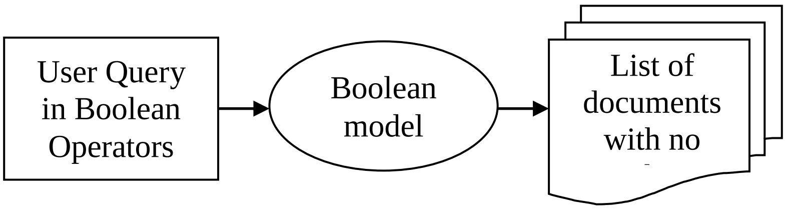

Figure 1.3 Boolean Model

1.2.2 Vector Space Model: It is a successful statistical method for generating weighted term vectors for the user query and for each document in a collection. Basedon the similarity between the document vectors and query vector, the output documents are ranked. The similarity depends on the occurrence frequencies of the keywords in the documents and query. It permits efficient implementation for large document collections.

1.2.3 Probabilistic Model: In this model, there is a set of documents with exactly relevant documents for a given user query, called the ideal answer set. The query is a process for specifying the properties of the answer set, but we do not know what are these properties. Therefore, an effort has to be made to guess a description of the answer set and retrieve an initial set of documents. Then, the user inspects the top retrieved documents, looking for the relevant ones. The IR system uses this information to refine the description of the ideal answer set. By repeating this process, it is expected that the description of the ideal answer set will improve (Inkpen2007).

1.2.4 Region Model: It is an extension of the Boolean model that deals with parts of textual data, called segments, extents or regions. The region consists of a sequence of consecutive words, which can be identified by the beginning and ending positions. Region models use two additional operators namely, CONTAINING and CONTAINED BY apart from the AND, OR, NOT Boolean operators.

1.2.5 Two-Poisson Model: The number of occurrences of terms in a document can be modeled by a combination of two Poisson distributions. It is assumed that the documents consist of a random stream of term occurrences where each term is divided into two sub sets. In documents in subset one, the subject referred by a term is given higher priority than the documents in subset two.For each term, the model needs the three terms λ, μ_1, μ_2. Here $\lambda$is the proportion of the documents in subset one, μ_1 and μ_2 are the mean number of occurrences of the term in the respective subsets.

1.2.6 Bayesian network Model: A Bayesian network is an acyclic directed graph that encodes probabilistic dependency relationships between random variables. An inference network model (Turtle and Croft 1991) was designed from the Bayesian network model. It consists of four layers of nodes: document nodes, representation nodes, query nodes and the information need node. All nodes of this network represent binary random variables with values {0,1}.

1.2.7. Language Model: These models were derived from probabilistic models of language generation developed for automatic speech recognition systems. Language models consider the same starting point as the probabilistic models.

1.3 ISSUES OF INFORMATION RETRIEVAL

The issues of information retrieval are discussed as follows:

1.3.1 Heterogeneous Data:

IR Algorithms usually work based on matching words in documents. However, the web consists of huge unstructured documents linked together which create a massive graph. This poses new challenges to IR (Bidokiand Yazdani2007).The matched documents apply the query to the same degree, thereby making it complex to rank the output.

Semantic information such as word sense and syntactic information such as phrase structure, word order, proximity information should be considered. It should have the control of a Boolean model.

1.3.2 Query Based:

It is not easy for the users to express complex queries. Hence, the query should be simple. An inverted index cannot be used when the documents are located by index keywords.

References:

[1] Christopher D. Manning, PrabhakarRaghavan and HinrichSchütze (2008), “Introduction to Information Retrieval”, Cambridge University Press.

[2] Josep Silva (2009),”Information Filtering and Information Retrieval with the Web Filtering Toolbar”, Electronic Notes in Theoretical Computer Science, Vol. 235, pp.125-136.

[3] LaxmiChoudhary and Bhawani Shankar Burdak (2012), "Role of Ranking Algorithms for Information Retrieval", International Journal of Artificial Intelligence and Applications (IJAIA), Vol.3, No.4

[4] Young DeokSeo, Jun Hyung Oh, JaeYoung Chang, Il-Min Kim (2015), “A Personal Information Retrieval System in a Web Environment”, Advanced Science and Technology Letters, Vol.87, pp.42-46.

[5] Ricardo Baeza-Yates and BerthierRibeiro-Neto (1999), “Modern Information Retrieval”, Addison-Wesley Publishing Company.