The Study of Information Retrieval

Ms. J. Sharmeli  
B.E Student  
Department of Computer Science & Engineering  
Valliammai Engineering College, Kattankulathur, Kanchipuram

Ms. V. Sanjeeya  
B.E Student  
Department of Computer Science & Engineering  
Valliammai Engineering College, Kattankulathur, Kanchipuram

Ms. G. Sangeetha  
Assistant Professor  
Department of Computer Science & Engineering  
Valliammai Engineering College, Kattankulathur, Kanchipuram

Abstract

The meaning of the term information retrieval can be very broad. Just getting a credit card out of your wallet so that you can type in the card number is a form of information retrieval. An information retrieval is a system where the end users extract information from www. However, as an academic field of study information retrieval might be defined as Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Keywords: Information Retrieval, IR Process, IR Models

I. INTRODUCTION

As defined in this way, information retrieval used to be an activity that only a few people are engaged in: reference librarians, paralegals, and similar professional searchers. Now the world has changed, and hundreds of millions of people engage in information retrieval every day when they use a web search engine or search their email. Information retrieval is fast becoming the dominant form of information access, overtaking traditional database style searching.

Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.” (Salton, 1968).

General definition that can be applied to many types of information and search applications. Primary focus of IR since the 50s has been on text and documents. The purpose of establishing an information retrieval system

IR can also cover other kinds of data and information problems beyond that specified in the core definition above. The term “unstructured data” refers to data which does not have clear, semantically overt, easy-for-a-computer structure. It is the opposite of structured data, the canonical example of which is a relational database, of the sort companies usually use to maintain product inventories and personnel records.

In reality, almost no data are truly “unstructured”. This is definitely true of all text data if you count the latent linguistic structure of human languages. But even accepting that the intended notion of structure is overt structure, most text has structure, such as headings and paragraphs and footnotes, which is commonly represented in documents by explicit markup (such as the coding underlying web pages).

II. TYPES OF INFORMATION SYSTEM

A. Database Management Systems (DBMS)

Relatively small schema, large body of homogeneous data Minor or no deductive capability, Extensive formal update capability, Shared use for both read and write

B. Information-Retrieval Systems (IR)

Search large bodies of information which are not specifically formatted as formal data bases. Web search engine, Keyword search of a text base, typically read-only.

C. Knowledge-Base Systems (KBS)

Relatively small body of heterogeneous information. Significant deductive capability, typical use: support of an intelligent application.
III. INFORMATION RETRIEVAL PROCESSING SYSTEM

The general process of IR system has three primary processes to support:
1) representation of the information of the documents,
2) interpretation of the user's information need, and
3) Comparison of these two representations.

In the figure 1, square shape box represent data and round shape box represent processes.

Indexer collects, parses, and stores data to facilitate fast and accurate information retrieval. Index design incorporates interdisciplinary concepts from linguistics, cognitive psychology, mathematics, informatics, and computer science. Popular search engines focus on the full-text indexing of online, natural language documents. Media types such as video and audio and graphics are also indexed and searchable.

The user need information is interpreted as query. The process of representing the information need of the user is called query formulation process.

The three broad categories of web search queries can be grouped into (i) informational, (ii) navigational and (iii) transactional.

Informational queries seek general information on a broad topic. There is typically not a single web page that contains all the information sought; indeed, users with informational queries typically try to assimilate information from multiple web pages. Navigational queries seek the website or home page of a single entity that the user has in mind, say Spicejet airlines. In such cases, the user's expectation is that the very first search result should be the home page of Spicejet. The user is not interested in documents containing the term Spicejet; for such a user, the best measure of user satisfaction is precision at 1.

A transactional query is one that the user performing a transaction on the Web - such as purchasing a product, downloading a file or making a reservation. In such cases, the search engine should return results listing services that provide form interfaces for such transactions.

User queries are matched against the database information. However, as opposed to classical SQL queries of a database, in information retrieval the results returned may or may not match the query, so results are typically ranked.

This ranking of results is a key difference of information retrieval searching compared to database searching.

IV. IR MODELS

An IR model specifies the details of the document representation, the query representation and the retrieval functionality. The fundamental IR models can be classified into Boolean, Vector, Probabilistic and Inference network model.

A. Boolean Model

The Boolean model is the first model of information retrieval and probably also the most criticized model. The Boolean model of information retrieval is a classical information retrieval (IR) model and, at the same time, the first and most adopted one. It is used by many IR systems to this day.

The Boolean model is based on Boolean logic and classical set theory in that both the documents to be searched and the user's query are conceived as sets of terms. Retrieval is based on whether or not the documents contain the query terms. The model can be explained by thinking of a query term as an unambiguous definition of a set of documents. For instance, the query term economic simply defines the set of all documents that are indexed with the term economic. Using the operators of George Boole's mathematical logic, query terms and their corresponding sets of documents can be combined to form new sets of documents. The Boolean model allows for the use of operators of Boolean algebra. It is easy to implement but more like data retrieval than information retrieval.
B. Vector Space Model

Gerard Salton and his colleagues suggested a model based on Luhn's similarity criterion that has a stronger theoretical motivation (Salton and McGill 1983). They considered the index representations and the query as vectors embedded in a high dimensional Euclidean space, where each term is assigned a separate dimension. The vector space model can best be characterized by its attempt to rank documents by the similarity between the query and each document [10]. In the Vector Space Model (VSM), documents and query are represented as a Vector and the angle between the two vectors are computed using the similarity cosine function.

Vector Space Model have been introduce term weight scheme known as if-idf weighting. These weights have a term frequency (tf) factor measuring the frequency of occurrence of the terms in the document or query texts and an inverse document frequency (idf) factor measuring the inverse of the number of documents that contain a query or document term. It is a simple model based on linear algebra and computes continuous degree of similarity between queries and documents. But still long documents are poorly represented because they have poor similarity values. Semantic sensitivity; documents with similar context but different term vocabulary won't be associated, resulting in a "false negative match".

C. Probabilistic Model

Whereas Maron and Kuhns introduced ranking by the probability of relevance, it was Stephen Robertson who turned the idea into a principle. He formulated the probability ranking principle, which he attributed to William Cooper, as follows (Robertson 1977). The most important characteristic of the probabilistic model is its attempt to rank documents by their probability of relevance given a query. Documents and queries are represented by binary vectors ~d and ~q, each vector element indicating whether a document attribute or term occurs in the document or query, or not. Instead of probabilities, the probabilistic model uses odds O(R), where O(R) = P(R)/1 − P(R), R means "document is relevant" and ~R means "document is not relevant".

D. Inference Network Model

Do in this model, document retrieval is modeled as an inference process in an inference network. Most techniques used by IR systems can be implemented under this model. In the simplest implementation of this model, a document instantiates a term with certain strength, and the credit from multiple terms is accumulated given a query to compute the equivalent of a numeric score for the document. From an operational perspective, the strength of instantiation of a term for a document can be considered as the weight of the term in the document, and document ranking in the simplest form of this model becomes similar to ranking in the vector space model and the probabilistic models described above. The strength of instantiation of a term for a document is not defined by the model, and any formulation can be used effectively.

V. THE VARIOUS INFORMATION RETRIEVAL SCHEMES THAT HAS BEEN DONE BY OTHER AUTHORS ARE STUDIED BELOW

<table>
<thead>
<tr>
<th>Paper Title</th>
<th>Author</th>
<th>Year</th>
<th>Concept</th>
<th>Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology-based Digital Photo Annotation Using Multi-source Information</td>
<td>Yanmei Chai, Xiaoyan Zhu, Sen Zhou, Yiting Bian, Fan Bu, Wei Li and Jing Zhe Tsinghua University</td>
<td>May 11-13, 2009</td>
<td>To overcome the difficulties of Users who encounter severe difficulties with the management and retrieval of information, especially when they want to find a desired one among tens of thousands of photos using just a simple query.</td>
<td>It provides manual annotation by automatically extracting semantic concepts from text, EXIF metadata and face detection result. It improves the performance of annotation by allowing users to freely edit the annotation information.</td>
</tr>
<tr>
<td>Research and Implementation of Automatic Question Answering System based on Ontology</td>
<td>Xingbo Xie, Wei Song, Lizhen Liu, Chao Du, Hanshi Wang.</td>
<td>2015</td>
<td>To implement the automatic question answering system. After fusing Ontology into the automatic question-answering, the system not only can analyze the user’s questions in the semantic level but also combine with the user’s questions for semantic reasoning. Therefore, the users can get a better understanding an more accurate result.</td>
<td>Ontology has a good concept hierarchy and supports logical reasoning. It is better in reflecting the logical and hierarchical relationships between knowledge than ordinary database. This system is not only able to answer user’s questions but also can provide relevant recommendations, which can help students study systematically.</td>
</tr>
<tr>
<td>Document Retrieval Using Knowledge-Based Fuzzy Information</td>
<td>Shyi-Ming Chen and Jeng-Yih Wang</td>
<td>5, MAY 1995</td>
<td>The knowledge is represented by a concept matrix, where the elements in a concept represent relevant values between concepts. The implicit relevant values between</td>
<td>Efficient retrieving capability and flexible user’s queries are consequently provided.</td>
</tr>
</tbody>
</table>
Through this paper, we have tried to highlight how to gain information, by using the WWW, which is considered to be the largest pool of information resources. Before starting the information retrieval process the knowledge to be extracted must be clearly identified. This technology helps in gaining meaningful insights related to the day-to-day activities by using the useful information made accessible through the web. It not only enables discovery of relevant information from millions of data on the WWW, but it also monitors and predicts user visit habits.

Using various information retrieval techniques discussed in the paper, helps in digging out layers of information about the keyword phrases, and the data collected from. Web sites present enormous potential content of information. Searchers can fine-tune their strategies by building profiles and using these to identify the segments upon which retrieval activities are focused. In the conclusion, information retrieval can be summed up as a viable technology, application meant for discovering knowledge pertaining to the routine search activities and can prove to be a very important approach for gaining competitive advantage.
REFERENCES