ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
www.computerscijournal.org

ISSN: 0974-6471
December 2011,
Vol. 4, No. (2):
Pgs. 459-462

Importance of Information Retrieval

C.S. NAGA MANJULA RANI

Assistant Professor in CSE, K.S.I.T., Raghuvanahalli, Kanakapura Road, Bangalore (India).

(Received: September 19, 2011; Accepted: September 29, 2011)

ABSTRACT

Information retrieval (IR) technique stands today at a crossroads. With the enormous increase in recent years in the number of text databases available on-line, and the consequent need for better techniques to access this information, there has been a strong resurgence of interest in the research done in the area of information retrieval (IR). Originally an outgrowth of librarianship, it has expanded into fields such as office automation, genome databases, fingerprint identification, medical image management, knowledge finding in databases, and multimedia management. This paper deals with importance of IR and classical models of IR.

Key words: Information Retrieval- Evaluation-models.

INTRODUCTION

The problem of information storage and retrieval has attracted increasing attentions since 1940. The problem states that we have vast amounts of information to be stored and the relevant information should be accessed accurately and efficiently. A great deal of work has been done to provide rapid and intelligent retrieval systems. However, the problem of effective retrieval remains largely unsolved. Improving the effectiveness is an important goal for the research of IR system.

IR is a science that aims to store and allow fast access to a large amount of information. This information could be of any kind: textual, visual or auditory. An Information Retrieval System(IRS) is a computing tool which stores this information to be retrieved for future use. Most actual IRS store and enable the retrieval of only textual information or documents.

IR deals with the processing of documents containing free text, so that they can rapidly retrieved based on keywords specified in a user's query. A user accesses the IRS by submitting a query, the IRS then tries to retrieve all documents that "satisfy" the query. The primary goal of an IRS is to retrieve all the documents which are relevant to a user query.

IR technology is the basis of web-based search engines and plays a vital role in the various fields. For several decades IR was an "orphan" technology researched by a relative handful of scientists. However, due to the spread of the World Wide Web, IR is now mainstream because most of the information on the web is textual. Web search engines such as Google, Yahoo, Excite, AltaVista etc. are used by millions of users to locate information on web pages across the world on any topic.

IR and other types of Information Systems

IRS can be related to different types of information systems such as DBMS and AI systems. The table summarizes the similarities and differences.

	Data Object	Primary Operation	Database size
IRS	Document	Retrieval (Probabilistic)	Small to very large
DBMS	Table	Retrieval (Deterministic)	Small to very large
AI	Logical statements	inference	Usually small

One difference between IR, DBMS and AI systems is the amount of usable structure in their data objects. Documents, being primarily text, in general have less usable structure than the tables of data used by RDBMS, and structures such as frames and semantic nets used by AI systems. It is possible to analyze a document manually and store information about its syntax and semantics in a DBMS or an AI system. Researchers hope to eventually develop practical systems that combine IR, DBMS and AI.

Another distinguishing feature of IRS is that retrieval is probabilistic. That is, one cannot be certain that a retrieved document will meet the information need of the user. In a typical search in an IRS, some relevant documents will be missed and some nonrelevant documents will be retrieved. This may be contrasted with retrieval form, for example, a DBMS where retrieval is deterministic.

One feature of IRS shared with many DBMS is that their databases are often very large-sometimes in the gigabyte range. Book library systems may contain several million records. Commercial on-line retrieval services such as Dialog and BRS provide databases of many gigabytes.

Another feature that IRS share with DBMS is database volatility. A typical large IR application, such as commercial document retrieval service, will change constantly as documents are added, changed and deleted.

A typical IRS must meet the following functional and non functional requirements. It must allow a user to add, delete, and change documents in the database. It must provide a way for users to search for documents by entering queries, and examine the retrieved documents. It must accommodate databases in the megabyte to gigabyte range, and retrieve relevant documents in response to queries interactively- often within 1 to 10 seconds.

Evaluation Measures

Many different measures for evaluating the performance of information retrieval systems have been proposed. The measures require a collection of documents and a query. All common measures described here assume a ground truth notion of relevancy: every document is known to be either relevant or non-relevant to a particular query. Two important metrics are precision and recall.

Precision

Precision is the fraction of the documents retrieved that are relevant to the user's information need. It is the ratio of the number of relevant documents retrieved over the total number of documents retrieved.

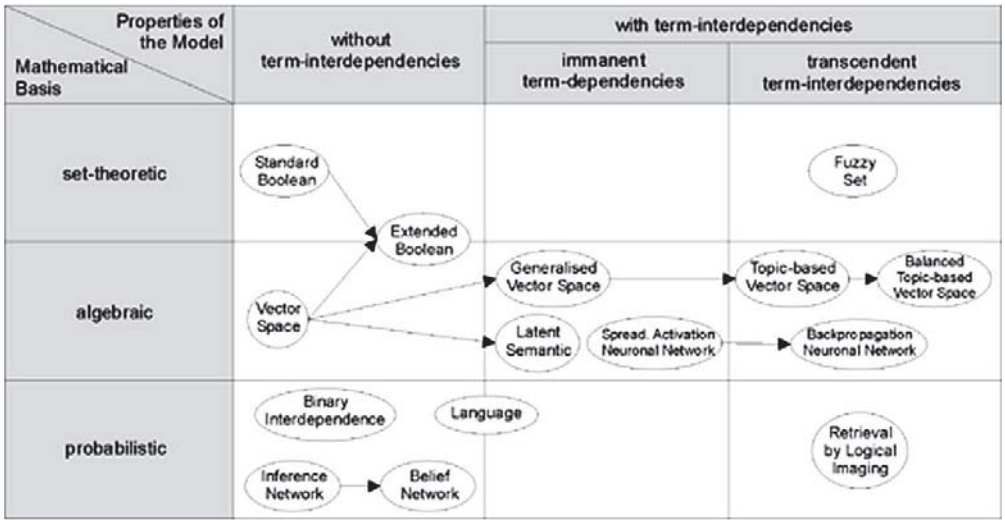
$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Recall

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved. It is the ratio of relevant documents retrieved for a given query over the number of relevant documents for that query in the database.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Except for small test collections, this



denominator is generally unknown and must be estimated by sampling or some other method. Both recall and precision take on values between 0 and 1.

IR Models

For the information retrieval to be efficient, the documents are typically transformed into a suitable representation. There are several representations. The picture below illustrates the relationship of some common models. In the picture, the models are categorized according to two dimensions: the mathematical basis and the properties of the model.

First dimension: mathematical basis

Set-theoretic models represent documents as sets of words or phrases. Similarities are usually derived from set-theoretic operations on those sets. Common models are:

- Standard Boolean model
- Extended Boolean model
- Fuzzy Retrieval

Algebraic models represent documents and queries usually as vectors, matrices, or tuples. The similarity of the query vector and document vector is represented as a scalar value.

- Vector space model
- Generalized vector space model
- Enhanced topic based vector space model
- Latent Semantic Indexing
- Neural Networks

Probabilistic models treat the process of document retrieval as a probabilistic inference. Similarities are computed as probabilities that a document is relevant for a given query. Probabilistic theorems like the Baye's theorem are often used in these models.

- Binary Independence model
- Probabilistic relevance model
- Inference Network
- Belief Network

Second dimension: properties of the model

- Models without term-interdependencies treat different terms/words as independent. This fact is usually represented in vector space models by the orthoganality assumption of term vectors or in probabilistic models by an independency assumption for term variables.
- Models with immanent term interdependencies allow a representation of interdependencies between terms. However the degree of the interdependency between two terms is defined by the model itself. It is usually directly or indirectly derived from the co-occurrence of those terms in the whole set of documents.
- Models with transcendent term interdependencies allow a representation of interdependencies between terms, but they do not allege how the interdependency

between two terms is defined. They relay an external source for the degree of interdependency between two terms. (For example a human or sophisticated algorithms).

CONCLUSION

Given the speed with which industry has adopted the results of IR research from the 1970s and 1980s, the IR community is faced with identifying major new directions. IR can also be used to reinforce the knowledge acquired in other subjects including programming, algorithms and data structures, and user interface design. The challenge for IR researchers is to define and pursue research programs that maintain their relevance in a rapidly changing environment.

REFERENCES

1. W.Bruce Croft. What Do People Want from Information Retrieval? D-Lib Magazine, (1995).
2. William.B.Frakes, Ricardo Baeza Yates. Information Retrieval Data Structures & Algorithms. LPE, Pearson Education.
3. Ricardo Baeza Yates Berthier Ribeiro-Neto. Modern Information Retrieval. Pearson Education.
4. Paul B. Kantor. Information Retrieval Techniques
5. Juan M. Ferna´ndez-Luna , Juan F. Huete, Andrew MacFarlane,Efthimis N. Efthimiadis. Teaching and learning in information retrieval. Information Retrieval (2009).
6. en.wikipedia.org