Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Trends and Techniques in Information Retrieval System (IRS): A Comparative Study

[1]Telkapalli Murali Krishna*, [2]Habtamu Fanta, [3]Sreedhar Appalabatla
[1, 3] Asst. Prof. in IT, Wolaita Sodo University, Ethiopia
[2] Lecturer in IT, Wolaita Sodo University, Ethiopia

Abstract: ***Information Retrieval (IR) system is intended to examine, evaluate and accumulate the sources of information and get back those that match user's requirements. IR is a rapidly expanding field and there are many changes in traditional techniques for IR day by day. This paper mainly focuses on learning the new theory, methods, procedures and tools that form the traditional approaches to the society and processing of information. In case of digital library environment, large amount of knowledge may be applicable in the storing and retrieval of electronic information. This paper discusses about the latest techniques used in multimedia IR; web IR; Natural Language Processing (NLP) Systems; IR in digital libraries and various trends in IR Research and comparing the various information retrieval strategies.***

Keywords: Information Retrieval, Natural Language Processing, Multimedia, Digital Library, Vector Space Model

I. INTRODUCTION

Information retrieval (IR) systems were originally developed to help and manage the large number of scientific literature. Contemporary era, many public libraries, universities and corporate offices uses IR systems to provide access to scientific research papers, journals, books and other relevant documents. IR has been useful in such different areas such as software engineering and office automation. Undoubtedly, any discipline that mainly depends on documents to do its work could potentially use and benefit from IR.

A query is a proper statement of user information need. The main goal of an IR system is to match the user query with the documents stored in a database. A document may contain tabular data, pictures, graphs and mainly textual. An IR system must support the certain fundamental operations such as to enter, change and delete the documents. Along with these, there should be a way to search for documents, retrieve and there must also be some way to search for documents, and giving them to the user.

II. FILE STRUCTURES IN IR

The file structures used in IR systems are flat files, inverted files, signature files, PAT trees, and graphs. Though it is possible to keep file structures in main memory, in practice IR databases are usually stored on disk because of their size.

Using a flat file approach, one or more documents are stored in a file, in the form of ASCII or EBCDIC text. Flat file searching is usually done via pattern matching. In UNIX, one can store a document collection one per file in a UNIX directory, and search it using pattern searching tools such as grep or awk.

An inverted file is an indexed file. The structure of an inverted file entry is usually keyword, document-ID, field-ID. A keyword is an indexing term that describes the document, document-ID is a unique identifier for a document, and field-ID is a unique name that indicates from which field in the document the keyword came. Search process is done by looking up query terms in the inverted file.

Signature files contain signatures and its patterns that represent documents. In this, documents are split into logical blocks each containing a fixed number of distinct significant, i.e., non-stoplist words. Each word in the block is hashed to give a signature--a bit pattern with some of the bits set to 1. The signatures of each word in a block are grouped together with 'OR' to create a block signature. The block signatures are then concatenated to produce the document signature. Searching is done by comparing the signatures of queries with document signatures.

Graphs, or networks, are ordered collections of nodes connected by arcs. They can be used to represent documents in various ways. A semantic net can be an example of a graph. Graph-based techniques for IR are impractical because of the amount of manual effort that would be needed to represent a large document collection in this form.

A. Query Operations

Queries are formal statements of information needs put to the IR system by users. Parsing is one common query operation, in which the query is broken down into its constituent elements. In Boolean queries, a query must be parsed into terms and operators. The set of document identifiers associated with each query term is retrieved, and the sets are then combined according to the Boolean operators.

© 2015, IJARCSSE All Rights Reserved

B. Term Operations

Operations on terms in an IR system include stemming, truncation, weighting and stoplist operations. Stemming is the automated conflation (fusing or combining) of related words, usually by reducing the words to a common root form. Truncation is manual conflation of terms by using wildcard characters in the word, so that the truncated term will match multiple words. For example, a searcher interested in finding documents about truncation might enter the term "truncat?" which would match terms such as truncate, truncated, and truncation. Another way of conflating related terms is with a thesaurus which lists synonymous terms, and sometimes the relationships among them. A stoplist is a list of words considered to have no indexing value, used to eliminate potential indexing terms. Each potential indexing term is checked against the stoplist and eliminated if found there.

In term weighting, indexing or query terms are assigned numerical values usually based on information about the statistical distribution of terms, that is, the frequencies with which terms occur in documents, document collections, or subsets of document collections such as documents considered relevant to a query.

C. Document Operations

Every document added to the database should be given unique ID (Identifier), parsed into fields, and these fields further divided into terms and field identifiers. The searcher may wish to search only the title and abstract fields of documents for a given query, or may wish to see only the title and author of retrieved documents. One may also wish to sort retrieved documents by some field, for example by author. Many sorting algorithms are available which are based on Precision and Recall. These algorithms are mainly based on probability of relevance to each document in a retrieved set, allowing retrieved documents to be ranked in order of probable relevance.

III. HARDWARE FOR IR

Most IR systems in use today are implemented on von Neumann machines--general purpose computers with a single processor. The computing speeds of these machines have improved enormously over the years, yet there are still IR applications for which they may be too slow. In response to this problem, some researchers have examined alternative hardware for implementing IR systems. There are two approaches--parallel computers and IR specific hardware. IR specific hardware has been developed both for text scanning and for common operations like Boolean set combination. Along with the need for greater speed has come the need for storage media capable of compactly holding the huge document databases that have proliferated. Optical storage technology, capable of holding gigabytes of information on a single disk, has met this need.

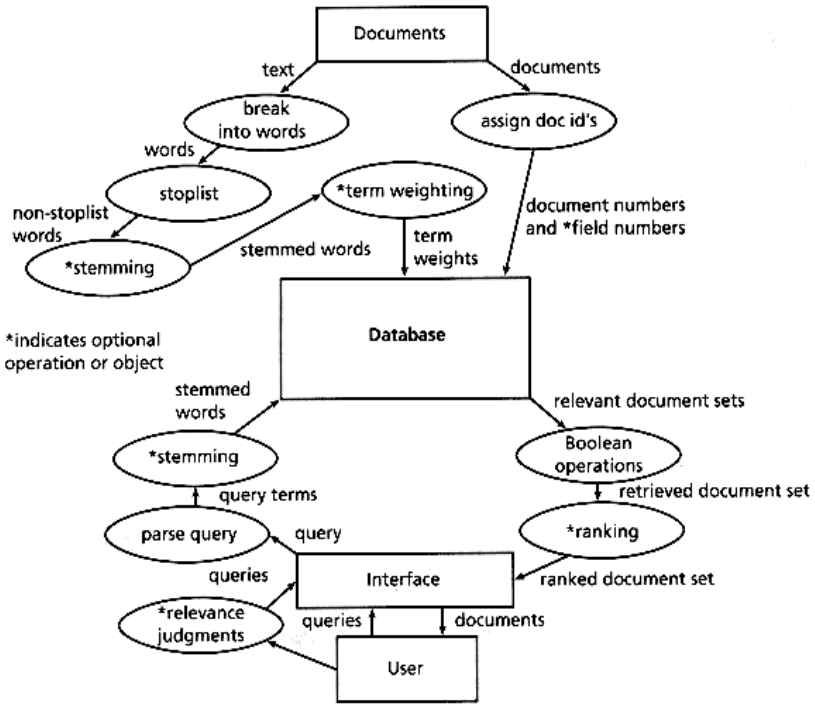

Fig 1. Functional view of an IR System

IV. MULTIMEDIA INFORMATION RETRIEVAL (MMIR OR MIR)

MIR aims at extracting semantic information from multimedia data sources such as audio, image and video directly and indirectly perceivable sources such as text, biosignals. The various methods used in traditional MIR can be organized in three groups:

- Feature extraction: summarization of media content and the result of feature extraction is a description.
- Filtering of media descriptions such as elimination of redundancy

© 2015, IJARCSSE All Rights Reserved

- Categorization of media descriptions into classes.

A. Characteristics of Multi Database Systems (MDBS):

- MDBS are distributed over different geographic locations connected via some form of communication network. Geographical distance between component databases can vary from inter-continental to distinct databases within the same DB server.
- Component DBs may have differing data management systems
- There is likely to be different schema architectures among the component databases.
- The MDB management system, MM-DBMS, maintains some form of global description of the component systems, commonly called a *global schema*
- The MM-DBMS supports global query and transaction processing.

A *multi- multimedia information retrieval system*, MMIRS will have all of the above characteristics. However, by nature media objects are seldom updated so that an MMIRS may not include transaction support.

V. ARCHITECTURES FOR MULTIPLE MIRS INTEGRATION

A. Homogeneous Systems

Homogeneous multi-DB systems are **tightly coupled** in the sense that they are designed from a single DB and then geographically distributed. Each component DB has the same schema structure so that the schema metadata are the same and synonyms are avoided. These systems are relatively easy to query and manage, but can be difficult to extend to cover new DB components.

B. Heterogeneous Systems

Heterogeneous multi-DB systems are **loosely coupled** in the sense that they are 'constructed' as an integration of existing heterogeneous systems, each of which has been independently designed and implemented and is in use for a local application set. The integration schema is constructed through the union of the schemas for the participating databases. A synonym table and a thesaurus may be constructed to support single query access to the multiple component databases. The objective of the global or federated schema is to hide the diversity of structure, location and naming conventions used in the component schemas.

C. Interoperable Systems

Disjoint or language based systems are very loosely connected. They have no or at best a very primitive, locally stored 'global' schema that defines the location and access paths to cooperating database systems. These systems have an extended query language processor that can access the local DB schema/metadata of cooperating systems and use domain ontologies to map a user query to relevant databases and documents within these.

A language-based approach to data integration in their development of tools for the semantic web was developed by W3C which contains XML (Extended Markup Language), DTD (Document type Definition), RDF (Resource Description Framework) schemas, and OWL(Web Ontology Languages). The strategy used is to 'package' each data element within a tag set defined by XML and its DTD or RDF schema that is accessible to both the sender and receiver. The primary application area has been that of e-commerce.

D. Query Processing in Multiple Multimedia Systems

The tools needed to access Web-data include those that are familiar to database management, i.e. data description (specifying metadata), indexing, search & retrieval, and presentation. Tools from both traditional *SQL3* and *Information Retrieval Systems* are being adapted for use in the multi DB environment of the Web.

VI. WEB INFORMATION RETRIEVAL (WEB IR)

For fast access to web information, the major industries like Google and Yahoo!, were the primary contributors of a technology. Searching is done in a collaborative manner about business management software, customer relationship systems to social networks and mobile phone applications etc. Even the technology for searching the Web is an important part of CS/IT/IS students, researchers and practitioners who wish to continuously educate themselves.

Relevance Ranking using Terms

TF-IDF (Term Frequency/Inverse Document Frequency) ranking

- Let n(d) = number of terms in the document d
- n(d , t) = number of occurrences of term t in the document d
- relevance of a document d to a term t
  $TF(d,t) = \log (1 + n(d,t) / n(d))$
  The log factor is to avoid excessive weight to frequent terms
- relevance of document to query Q
  $r(d,Q) = \sum TF(d, t) / n(t)$

© 2015, IJARCSSE All Rights Reserved

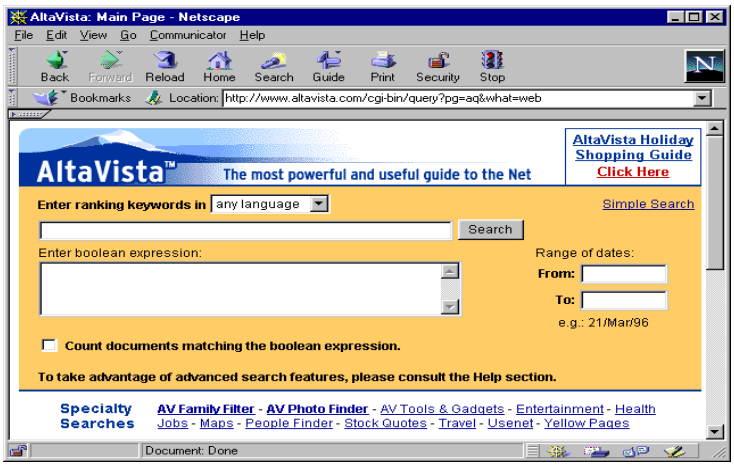

Fig 2. Form-based Query Specification (AltaVista)

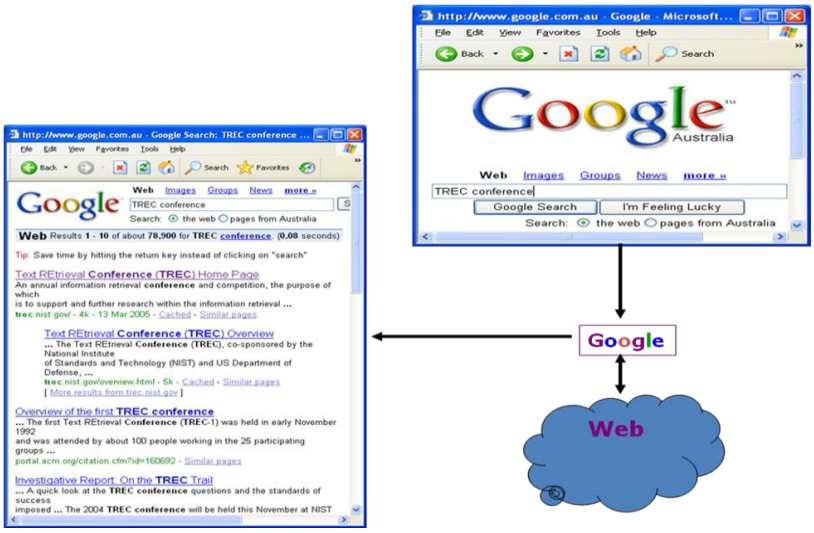

Fig 3. Form-based Query Specification (Google)

The goal of Web IR is to find relevant documents to an information need from a large document set.
The two main types of evaluation are (a) Formative (b) Summative. Formative evaluation is done at different stages of development to check the product meets users' needs, where as summative evaluation assesses the quality of a finished product.
The various search strategies used in Web IR are Web directories, query-based searching, link-based browsing and bookmarking or combination of the above strategies.
For effective web IR, we are supposed to use Term Clustering, user modeling, personalized systems (for eg., in a TV guide only show programs of interest). The three major information retrieval models (strategies) are Boolean Model, Vector Model and Probabilistic models.

VII. NATURAL LANGUAGE PROCESSING (NLP)

Huge amounts of text are available on the internet and several company intranets and most of the information is in human language. NLP is much helpful in processing such large amount of text in the form of Indexing & search, Text categorization, information extraction and knowledge acquisition.
The primary difference between natural language (NL) and Computer Languages (CLs) is CLs are designed by grammar that produce a unique parse for each sentence in the language. Some of the NLP applications are Information Extraction, Machine Translation, Automatic Summarization, Opinion Mining and Text Categorization.

VIII. CONCLUSION

There are various information retrieval models such as Boolean Model, Vector Model and Probabilistic models. Boolean Retrieval model is a simple model based on set theory and Boolean algebra. The advantage is that the Boolean expressions have precise semantics and structured queries and the disadvantage of this is frequently it is not simple to translate an information need into a Boolean expression. And also most users find it difficult and awkward to express their query requests in terms of Boolean expression. The Vector space model represents documents and queries as vectors in the term space. The Advantage of this model is "no output flattening", i.e., each query term contributes to the ranking in an equal way, depending on its weights and it showed much better effectiveness w.r.t. the Boolean Model. The Probabilistic model computes the similarity coefficient between queries and documents as the probability that a

© 2015, IJARCSSE All Rights Reserved

document will be relevant to a query. The advantages of this model are documents are ranked in decreasing order of probability of being relevant and it includes a mechanism for relevance feedback. On the other side, there is a need to guess the initial separation of documents into relevant and irrelevant and also the assumption of independence of index terms is difficult. In conclusion, for generic collections, the vector model does better performance than probabilistic model.

IX. FUTURE RESEARCH: CURRENT STATUS OF MMIRS

IR is difficult because of several reasons. Some of them are vocabulary mismatching, queries are ambiguous, content representation may be inadequate and incomplete. Though multi-database management has been a research area for more than three decades, only tightly coupled systems are well understood and even these still need to be hand crafted. There have been numerous research projects and prototypes for the loosely coupled and language-based systems, but no feasible general system has yet evolved. One approach, at least to study the problems, could be to define new SQL3 functions to define a multi-database set and to search them. i.e., we could design a new *multi-database* extension. OR-DBMS technology can be a powerful tool for development of database systems for organizations that have a combination of structured and multimedia data.

REFERENCES

[1] Chin-Ming Hong , Chih-Ming Chen , Chao- Yang Chiu, Automatic extraction of new words based on Google News corpora for supporting lexicon-based Chinese word segmentation systems, Expert Systems with Applications: An International Journal, v.36 n.2, p.3641-3651, March, 2009

[2] Charles Inskip, How different social groups within the music industry communicate meaning in order to satisfy their information needs, Proceedings of the 1st BCS IRSG conference on Future Directions in Information Access, August 28-29, 2007, Glasgow, Scotland

[3] Tao Peng , Lu Liu , Wanli Zuo, PU text classification enhanced by term frequency-inverse document frequency-improved weighting, Concurrency and Computation: Practice & Experience, v.26 n.3, p.728-741, March 2014

[4] Yanhui Gu , Zhenglu Yang , Guandong Xu , Miyuki Nakano , Masashi Toyoda , Masaru Kitsuregawa, Exploration on efficient similar sentences extraction, World Wide Web, v.17 n.4, p.595-626, July 2014

[5] Charles Inskip , Richard Butterworth , Andrew MacFarlane, A study of the information needs of the users of a folk music library and the implications for the design of a digital library system, Information Processing and Management: an International Journal, v.44 n.2, p.647-662, March, 2008

[6] Peng Tang , Tommy W. S. Chow, Mining language variation using word using and collocation characteristics, Expert Systems with Applications: An International Journal, v.41 n.17, p.7805-7819, December, 2014

[7] Hao Wen , Liping Fang , Ling Guan, Automatic Web Page Classification Using Various Features, Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing, December 09-13, 2008, Tainan, Taiwan

[8] Introduction to Information Retrieval by Jian-Yun Nie

© 2015, IJARCSSE All Rights Reserved