International Journal of Scientific Research in Computer Science and Engineering
Vol.6, Issue.2, pp. 109-112, April (2018)
Review Paper
E-ISSN: 2320-7639

A Comparative Study on the Assortment of Information Retrieval Systems

L. Senthilvadivu

Mahendra Arts & Science College, Kalippatti, India

Available online at: www.isroset.org

Received: 10/Mar/2018, Revised: 20/Mar/2018, Accepted: 04/Apr/2018, Online: 30/Apr/ 2018

***Abstract*-** For thousands of years people have realized the importance of archiving and finding information. With the advent of computers, it became possible to store large amounts of information; and finding useful information from such collections became a necessity. Over the last forty years, the field has matured considerably. Several Information Retrieval (IR) systems are used on an everyday basis by a wide variety of users. This paper presents a brief overview of the comparisons of the few assortments of Information Retrieval (IR) models and the description of the connoisseur in the field. The information retrieval by submitting the queries bring out millions of documents which consume the precious time of the user. This effort gives the information to the user to save their time in retrieving the information from the massive information sources.

***Keywords*: -** Information Retrieval (IR), assortment, connoisseur

1. INTRODUCTION

The field of Information Retrieval (IR) was born in the 1950s out of the necessity of retrieving the relevant information for the user. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. Information retrieval (IR) is finding material usually documents of an unstructured nature such as text that satisfies an information need from within large collections which is stored on computers.

IR can also cover other kinds of data and information problems beyond that specified in the core definition. 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 one 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. IR is also used to facilitate ``semi structured" search such as finding a document where the title contains Java and the body contains threading.

The First chapter represents the Introduction for the information retrieval sources available in different forms. The second chapter includes the different Information Retrieval (IR) models. The third chapter gives the evaluation of different IR models. The fourth chapter refers the related work and the fifth chapter gives away the conclusion from the comparative study of the different models.

2. IR MODELS

Most IR systems compute a numeric score on how well every object in the database matches the query, and ranks the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query [9].

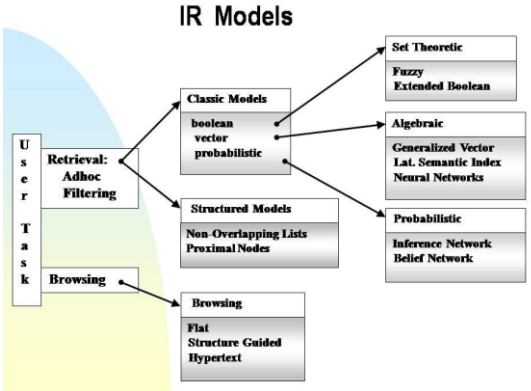

Fig 1. IR Models

Fig 1 shows the assortment of IR models in which Information retrieval systems can be distinguished by various scales at which they operate, and it is useful to distinguish three prominent scales. In web search, the system has to provide search over billions of documents stored on millions of computers. Distinctive issues need to gather documents for indexing, being able to build systems that work efficiently at this enormous scale, and handling particular aspects of the web, such as the exploitation of hypertext and not being fooled by site providers manipulating page content in an attempt to boost their search engine rankings, given the commercial importance of the web. Along which the important IR models such as Vector Space model, Boolean model, Probabilistic model and INAR system have been taken into consideration.

2.1 Vector Space Model

Information retrieval in the vector space model [10], [20] is based on literal matching of terms in the documents and the queries. The model is implemented by creating the term-document matrix, which is formed on the base of frequencies of terms in documents. Literal matching [11] of terms does not necessarily retrieve all relevant documents. Synonymy i.e, multiple words having the same meaning and polysemy i.e., words having multiple meaning are two major obstacles for efficient information retrieval. Latent semantic indexing (LSI) and concept indexing (CI) are information retrieval techniques embedded in the vector space model, which address the problem of synonymy and polysemy. The method of LSI is an information retrieval technique using low-rank singular value decomposition (SVD) of the term-document matrix. Although the LSI method has empirical success, it suffers from the lack of interpretation for the low-rank approximation and, consequently, the lack of controls for accomplishing specific tasks in information retrieval. The method of CI uses centroids of clusters or so-called concept decomposition (CD) for lowering the rank of the term-document matrix. Here we compare SVD/LSI and CD/CI in terms of matrix approximations and precision of information retrieval.

2.2 Boolean Model

Despite decades of academic research on the advantages of ranked retrieval, systems implementing the Boolean retrieval model [8], [4] were the main or only search option provided by large commercial information providers for three decades until the early 1990s approximately the date of arrival of the World Wide Web. However, these systems did not have just the basic Boolean operations (AND, OR, and NOT) which have been presented so far. A strict Boolean expression over terms with an unordered results set is too limited for many of the information needs that people have, and these systems implemented extended Boolean retrieval models by incorporating additional operators such as term proximity operators. A proximity operator is a way of specifying that two terms in a query must occur close to each other in a document, where closeness may be measured by limiting the allowed number of intervening words or by reference to a structural unit such as a sentence or paragraph.

2.3 Probabilistic Model

During the discussion of relevance feedback, it is observed that if some known relevant and non relevant documents, then straightforwardly start to estimate the probability of a term appearing in a relevant document , and that this could be the basis of a classifier that decides whether documents are relevant or not. Systematically it is introduced that this probabilistic approach to IR, which provides a different formal basis for retrieval model and results in different techniques for setting term weights. There is more than one possible retrieval model which has a probabilistic basis. The combination of the probability theory and the Probability Ranking Principle and the Binary Independence Model [21] is the original and still most influential probabilistic retrieval model. Finally, it will be introduced the related but extended methods which use term counts, including the empirically successful Okapi BM25 weighting scheme, and Bayesian Network models for IR. The alternative probabilistic language modeling approach to IR is presented, which has been developed with considerable success in recent years.

2.4 INAR system

The INAR system is the imperative part of the prevailing system [5] named as the Information Assimilation and Retrieval (**INAR)** system. The Cleaving of Phrases with NLQ (Natural Language Query) Parser which will provide a number of services for accessing relevant information, ranging from a straight forward keyword based search, to a topic search. In the INAR system, NLQ parser separates the phrases given by the user. In Domain based search, the input given by the user is separated as Domain, Child Domain and Concept with the NLQ parser. The result set can be known as hierarchical result set which means, if the user gives Domain only the user gets the result found within Domain. If the user puts the query to search the Child domain the results are fetched by the system from the Domain and Child domain and so on. Hence the searching techniques are trying to refine the searching area in respond to the queries by the user. NLQ Parser first counts the given term. Different counts define the application way. The terms can be understood by the application and the terms will be sent to persistence and matched with appropriate fields to build the query.

3. EVALUATION

The Boolean retrieval model [8],[4] contrasts with ranked retrieval models such as the vector space model, in which users largely use free text queries , that is, just typing one or more words rather than using a precise language with

operators for building up query expressions, and the system decides which documents best satisfy the query. Users start with information needs, which they translate into query representations. Similarly, there are documents, which are converted into document representations the latter differing at least by how text is tokenized, but perhaps containing fundamentally less information, as when a non-positional index is used. Based on these two representations, a system tries to determine how well documents satisfy information needs. In the Boolean or vector space models of IR, matching is done in a formally defined but semantically imprecise calculus of index terms. Given only a query, an IR system has an uncertain understanding of the information need. Given the query and document representations, a system has an uncertain guess of whether a document has content relevant to the information need. Probability theory provides a principled foundation for such reasoning under uncertainty. This chapter provides one answer as to how to exploit this foundation to estimate how likely it is that a document is relevant to an information need.

Phrase collection are separated to make easy analysis and assist the application the NLQ parser comes to divide the data and make them as partitioned blocks in INAR system. The blocks can be identified as Inter Relationship task which identifies the domain and related items. After the segregation, the communicator communicates with persistence memory that holds the data and prepares itself to provide the data based on domain counts. In persistence memory, lots of static and dynamic information are stored for the web data retrieval to the user. Domain relationship Algorithm fetches the data and Finder process correctly defines the phrases using counter process. Information retrieval from the link set is another focal process and it helps the navigation process very effectively.

4. RELATED WORK

In the case indexed information retrieval [6], [16], [17] within a document collection, it is assumed that each document has a unique serial number, known as the document identifier docID. During index construction, it can be simply assigned successive integers to each new document when it is first encountered. The input to indexing is a list of normalized tokens for each document, which we can equally think of as a list of pairs of SORTING term and docID. The core indexing step is sorting this list so that the terms are alphabetical gives the representation. Multiple occurrences of the same term from the same document are then merged. Instances of the same term are then grouped. Not only can inverted files be used to evaluate typical queries in less time than can signature files, but inverted files [1] require less space and provide greater functionality. The results also show that a synthetic text database can provide a realistic indication of the behavior of an actual text database. The tools used to generate the synthetic database have been made publicly available.

The basic idea of the heuristic methods is to process the query terms in an order so that as many top documents as possible can be identified without processing all of the query terms. The first heuristic was proposed by Smeaton and van Rijsbergen and it serves as the basis for comparison with the other two heuristic methods proposed in this paper. These three heuristics are evaluated and compared by experimental runs based on the number of disk accesses required for partial document ranking [2], in which the returned documents contain some, but not necessarily all, of the requested number of top documents.

The paper [7] presents an intelligent Internet information system, Automatic Classifier for the Internet Resource Discovery (ACIRD), which uses machine learning techniques to organize and retrieve Internet documents. ACIRD consists of a knowledge acquisition process, document classifier and two-phase search engine. The knowledge acquisition process of ACIRD automatically learns classification knowledge from classified Internet documents. The document classifier applies learned classification knowledge to classify newly collected Internet documents into one or more classes. Experimental results indicate that ACIRD performs as well or better than human experts in both knowledge acquisition and document classification. By using the learned classification knowledge and the given class lattice, the ACIRD two-phase search engine responds to user queries with hierarchically structured navigable results instead of a conventional flat ranked document list, which greatly aids users in locating information from numerous, diversified Internet documents.

Bayesian networks [19] use directed graphs to show probabilistic dependencies between variables and have led to the development of sophisticated algorithms for propagating influence so as to allow learning and inference with arbitrary knowledge within arbitrary directed acyclic graphs. Turtle and Croft used a sophisticated network to better model the complex dependencies between a document and a user's information need. The model decomposes into two parts: a document collection network and a query network. The document collection network is large, but can be precomputed: it maps from documents to terms to concepts. The concepts are a thesaurus-based expansion of the terms appearing in the document. The query network is relatively small but a new network needs to be built each time a query comes in, and then attached to the document network. The query network maps from query terms, to query sub expressions built using probabilistic or ``noisy'' versions of AND & OR operators, to the user's information need.

5. CONCLUSION

There are many retrieval models contrast with models such as the Boolean model, vector space model and probabilistic model and INAR system in which users largely use free text queries, that is, just typing one or more words rather than using a precise language with operators for building up query expressions, and the system decides which documents best satisfy the query. Straightforwardly start to estimate the probability of a term appearing in a relevant document, and that this could be the basis of a classifier that decides whether documents are relevant or not which brings out millions of links to confuse the user. The INAR system explains the search based on the queries posted by the user to the multi related, heterogeneous information sources which gives out the content merely matching the queries. This kind of searching is trying to give simply the precise content retrieval for the user based on the queries posted by the user which can eliminate the searching of millions of links for the same queries and saves the time of the user.

REFERENCES

[1] Zobel, Justin, and Alistair Moffat. "Inverted files for text search engines", ACM Computing Surveys 38 (2),2006.

[2] Zobel, Justin, Alistair Moffat, Ross Wilkinson, and Ron Sacks-Davis."Efficient retrieval of partial documents". IP&M 31 (3): 361-377. DOI: dx.doi.org/10.1016/0306-457300052 -5, 1995.

[3] Zaragoza, Hugo, Djoerd Hiemstra, Michael Tipping, and Stephen Robertson. "Bayesian extension to the language model for ad hoc information retrieval". In Proc. SIGIR, pp. 4-9, 2005.

[4] Zukowski, Marcin, Sandor Heman, Niels Nes, and Peter Boncz. "Super -scalar RAM-CPU cache compression" .In Proc. International Conference on Data Engineering, p. 59. IEEE Computer Society. dx.doi.org/10.1109/ ICDE. 2006.150.

[5] L.Senthilvadivu, K.Duraiswamy "Conniving the Information Assimilation and Retrieval (INAR) system for the heterogeneous, multi related Information Sources", World of Computer Science and Information Technology Journal pp.357-363,2011.

[6] Zavrel, Jakub, Peter Berck, and Willem Lavrijssen. "Information extraction by text classification: Corpus mining for features",2000.

[7] Shian-hua Lin, Meng Chang Chen, Jan-ming Ho, Yueh-ming Huang - ACIRD: "Intelligent Internet Documents Organization and Retrieval" IEEE Transactions on Knowledge and Data Engineering 2002

[8] Lashkari, A.H.; Mahdavi, F.; Ghomi, V. " A Boolean Model in Information Retrieval for Search Engines", doi:10.1109/ICIME101, 2009

[9] Frakes, William B. (1992). Information Retrieval Data Structures & Algorithms, Prentice-Hall, Inc. ISBN 0-13-463837-9.

[10] Jasminka Dobsa, Faculty of Organization and Informatics, Comparison of information retrieval techniques: Latent semantic indexing (LSI) and Concept indexing (CI) published: Feb. 25, 2007, views: 1308

[11] Zobel, Justin, and Philip Dart. "Phonetic string matching". In Proc. SIGIR, pp. 166-173,1996. ACM Press.

[12] C.D. Manning, P. Raghavan, H. Schütze. Cambridge UP, "Classical and web information retrieval systems: algorithms", mathematical foundations and practical issues, 2008.

[14]Managing Gigabytes. I.H. Witten, A. Moffat, T.C. Bell. Morgan Kaufmann, "The authority on index construction and compression",1999.

[15]Readings in Information Retrieval. K. Sparck Jones, P. Willett. Morgan Kaufmann, " A collection of classical IR papers", 1997.

[16]Information Storage and Retrieval Systems. G. Kowalski, M.T. Maybury. Springer, "Takes a system approach, discussing all aspects of an Information Retrieval System.", 2005

[17] G.G.Chowdhury. Neal-Schuman, "Introduction to Modern Information Retrieval", 2003.

[18] D.H. Kraft, C.L. Barry, C.T. Meadow, B.R. Boyce "Text Information Retrieval Systems", 2007, Academic Press.

[19] Jensen fv. Bayesian Networks and Decision Graphs. Technometrics, Volume 45, Number 2, pp. 178-179(2) 2003.

[20] Anh, Vo Ngoc, Owen de Kretser, and Alistair Moffat "Vector-space ranking with effective early termination". In Proc. SIGIR, pp. 35-42,2001. ACM Press.

[21] Crestani, Fabio, Mounia Lalmas, Cornelis J. Van Rijsbergen, and Iain Campbell. " A survey of probabilistic models in information retrieval".ACM Computing Surveys30 (4):528-552,1998.doi.acm.org/10.1145/ 299917.

