International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Document Retrieval Techniques using Vector Space Model

Tanmayee Tushar Parbat[1], Rohan Benhal[2], Honey Jain[1]
[1]B.E IT, Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India
[2]BBA IT, Dr. Vishwanath Karad MIT World Peace University, Pune, Maharashtra, India

Article Info
Volume 8, Issue 5
Page Number: 93-99

Publication Issue :
September-October-2021

Article History
Accepted : 02 Oct 2021
Published : 26 Oct 2021

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. The field of Document Retrieval (DR) was born in the 1950s out of this necessity. Over the last forty years, the field has matured considerably. Several DR systems are used on an everyday basis by a wide variety of users. Information retrieval is become a important research area in the field of computer science. Information retrieval (IR) is generally concerned with the searching and retrieving of knowledge-based information from database. In this paper, we represent the various models and techniques for information retrieval. In this Review paper we are describing different indexing methods for reducing search space and different searching techniques for retrieving a information. We are also providing the overview of traditional IR models.

Keywords:- Document Retrieval (IR), Indexing, IR mode, Searching, Vector Space Model (VSM)

I. INTRODUCTION

Information retrieval is generally considered as a subfield of computer science that deals with the representation, storage, and access of information [1]. Information retrieval is concerned with the organization and retrieval of information from large database collections [2]. Information Retrieval (IR) is the process by which a collection of data is represented, stored, and searched for the purpose of knowledge discovery as a response to a user request (query) [3].this process involves various stages initiate with representing data and ending with returning relevant information to the user. Intermediate stage includes filtering, searching, matching and ranking operations. The main goal of information retrieval system (IRS) is to "finding relevant information or a document that satisfies user information needs". To achieve this goal, IRSs usually implement following processes: 1) In indexing process the documents are

represented in summarized content form. 2) In filtering process all the stop words and common words are remove. 3) Searching is the core process of IRS. There are various techniques for retrieving documents that match with users need. There are two basic measures for assessing the quality of information retrieval [2]. Precision: This is the percentage of retrieved documents that are in fact relevant to the query. Recall: This is the percentage of documents that are relevant to the query and were in fact retrieved. There are three basic processes an information retrieval system has to support: the representation of the content of the documents, the representation of the user's information need, and the comparison of the two representations. The processes are visualized in Figure 1. In the figure, squared boxes represent data and rounded boxes represent processes. Representing the documents is usually called the indexing process. The process takes place off-line, that is, the end user of the information retrieval system is not directly involved. The indexing process results in a representation of the document [5]. Users do not search just for fun, they have a need for information. The process of representing their information need is often referred to as the query formulation process. The resulting representation is the query [5]. Comparing the two representations is known as the matching process.

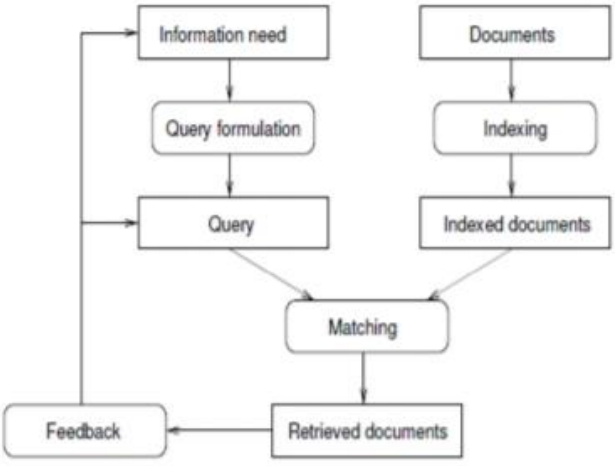


Fig 1: Information retrieval processes

II. ITEM NORMALIZATION

The first step in any integrated system is to normalize the incoming items to a standard format. Item normalization provides logical restructuring of the item. Additional operations during item normalization are needed to create a searchable data structure: identification of processing tokens (e.g., words), characterization of the tokens, and stemming (e.g., removing word endings) of the tokens. The processing tokens and their characterization are used to define the searchable text from the total received text. Figure 1.5 shows the normalization process. Standardizing the input takes the different external formats of input data and performs the translation to the formats acceptable to the system. A system may have a single format for all items or allow multiple formats. One example of standardization could be translation of foreign languages into Unicode. Every language has a different internal binary encoding for the characters in the language. One standard encoding that covers English, French, Spanish, etc. is ISO-Latin.

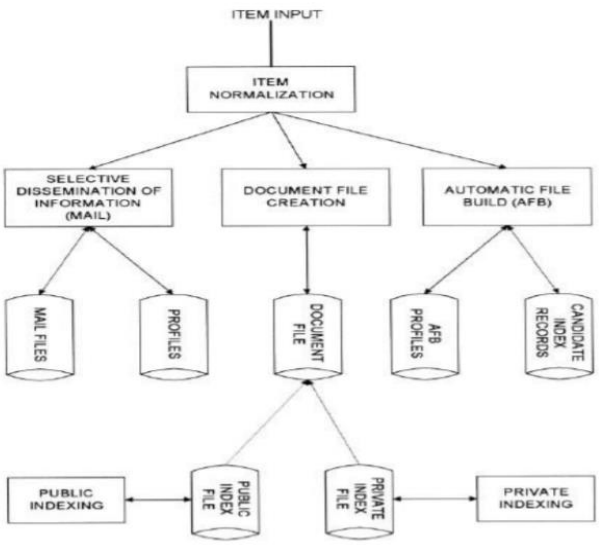


Fig. 2: Total Information retrieval System

Multi-media adds an extra dimension to the normalization process. In addition to normalizing the textual input, the multi-media input also needs to be standardized. There are a lot of options to the

standards being applied to the normalization. If the input is video the likely digital standards will be either MPEG-2, MPEG-1, AVI or Real Media. MPEG (Motion Picture Expert Group) standards are the most universal standards for higher quality video where Real Media is the most common standard for lower quality video being used on the Internet. Audio standards are typically WAV or Real Media (Real Audio). Images vary from JPEG to BMP.

III. DOCUMENT DATABASE SEARCH

The Document Database Search Process provides the capability for a query to search against all items received by the system. The Document Database Search process is composed of the search process, user entered queries (typically ad hoc queries) and the document database which contains all items that have been received, processed and stored by the system. Typically items in the Document Database do not change (i.e., are not edited) once received. Index Database Search When an item is determined to be of interest, a user may want to save it for future reference. This is in effect filing it. In an information system this is accomplished via the index process. In this process the user can logically store an item in a file along with additional index terms and descriptive text the user wants to associate with the item. The Index Database Search Process (see Figure 2) provides the capability to create indexes and search them. There are 2 classes of index files:

1) Public Index files

2) Private Index files

Every user can have one or more Private Index files leading to a very large number of files. Each Private Index file references only a small subset of the total number of items in the Document Database. Public Index files are maintained by professional library services personnel and typically index every item in the Document Database. There is a small number of Public Index files. These files have access lists (i.e., lists of users and their privileges) that allow anyone to search or retrieve data. Private Index files typically have very limited access lists. To assist the users in generating indexes, especially the professional indexers, the system provides a process called Automatic File Build shown in Figure 2 (also called Information Extraction).

Multimedia Database Search

From a system perspective, the multi-media data is not logically its own data structure, but an augmentation to the existing structures in the Information Retrieval System.

Relationship to Database Management Systems

From a practical standpoint, the integration of DBMS's and Information Retrieval Systems is very important. Commercial database companies have already integrated the two types of systems. One of the first commercial databases to integrate the two systems into a single view is the INQUIRE DBMS. This has been available for over fifteen years. A more current example is the ORACLE DBMS that now offers an imbedded capability called CONVECTIS, which is an informational retrieval system that uses a comprehensive thesaurus which provides the basis to generate "themes" for a particular item. The INFORMIX DBMS has the ability to link to RetrievalWare to provide integration of structured data and information along with functions associated with Information Retrieval Systems.

Digital Libraries and Data Warehouses (DataMarts)

As the Internet continued its exponential growth and project funding became available, the topic of Digital

Libraries has grown. By 1995 enough research and pilot efforts had started to support the 1ST ACM International Conference on Digital Libraries (Fox-96). Indexing is one of the critical disciplines in library science and significant effort has gone into the establishment of indexing and cataloging standards. Migration of many of the library products to a digital format introduces both opportunities and challenges. Information Storage and Retrieval technology has addressed a small subset of the issues associated with Digital Libraries. Data warehouses are similar to information storage and retrieval systems in that they both have a need for search and retrieval of information. But a data warehouse is more focused on structured data and decision support technologies. In addition to the normal search process, a complete system provides a flexible set of analytical tools to "mine" the data. Data mining (originally called Knowledge Discovery in Databases - KDD) is a search process that automatically analyzes data and extract relationships and dependencies that were not part of the database design.

IV. METHODOLOGY

IR Model

An IR model specifies the details of the document representation, the query representation and the retrieval functionality [3]. The fundamental IR models can be classified into Boolean, vector, probabilistic and inference network model [8] [3]. The rest of this section briefly describes these models.

Boolean Model

The Boolean model is the _rst model of information retrieval and probably also the most criticized model. The Boolean model is the _rst model of information retrieval and probably also the most criticized model. The model can be explained by thinking of a query term as a 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, AND, OR and NOT, for query formulation, but has one major disadvantage: a Boolean system is not able to rank the returned list of documents [4]. In the Boolean model, a document is associated with a set of keywords. Queries are also expressions of keywords separated by AND, OR, or NOT/BUT. The retrieval function in this model treats a document as either relevant or irrelevant [3]. In Figure 3, the retrieved sets are visualised by the shaded areas.

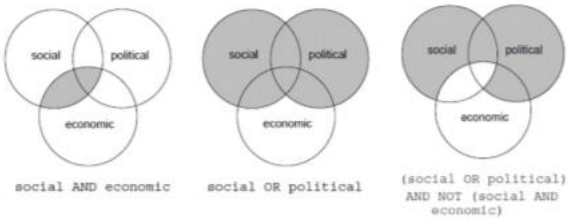


Fig 3: Boolean combinations of sets visualized as Venn diagrams

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 represent 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 [4].

There are various searching algorithms, including linear search, binary search, brute force search etc. some general searching algorithms are described below:

1) In linear search algorithm is a method of finding a particular element or keyword from list or array that checks every element in list, one at a time and in sequence. Linear search is a simplest search algorithm. One of the most important drawbacks of linear search is slow searching speed in ordered list. This search is also known as sequential search.

2) Brute force search is a very general problemsolving technique that consists of systematically enumerating all possible candidates for the solution and checking whether each candidate satisfies the problem's statement. Brute force algorithm is simple to implement and it will always find a solution if it exist.

3) Binary search algorithm, finds specified position of the element by using the key value with in a sorted array. In each step, the algorithm compares the search key value with the key value of the middle element of the array. If the keys match, then a matching element has been found and its index, or position, is returned. Otherwise, if the search key is less than the middle element's key, then the algorithm repeats its action on the sub-array to the left of the middle element or, if the search key is greater, on the subarray to the right.

V. SIMULATION RESULTS

Three classical text retrieval techniques have been defined for organizing items in a textual database, for rapidly identifying the relevant items and for eliminating items that do not satisfy the search. The techniques are

1) Full text scanning (streaming)
2) Word inversion
3) Multiattribute retrieval

In addition to using the indexes as a mechanism for searching text ininformation systems, streaming of text was frequently found in the systems as anadditional search mechanism. The basic concept of a text scanning system is the ability for one or moreusers to enter queries, and the text to be searched is accessed and compared to thequery terms. When all of the text has been accessed, the query is complete.

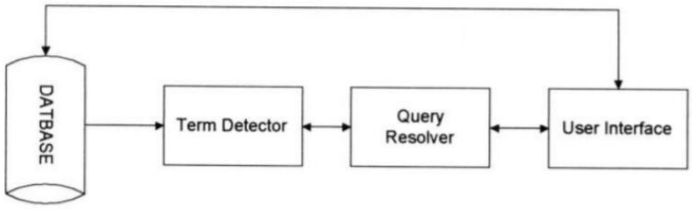
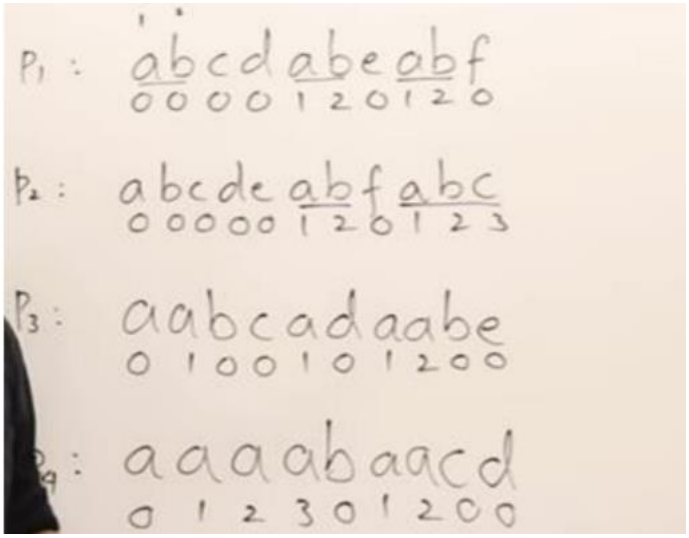


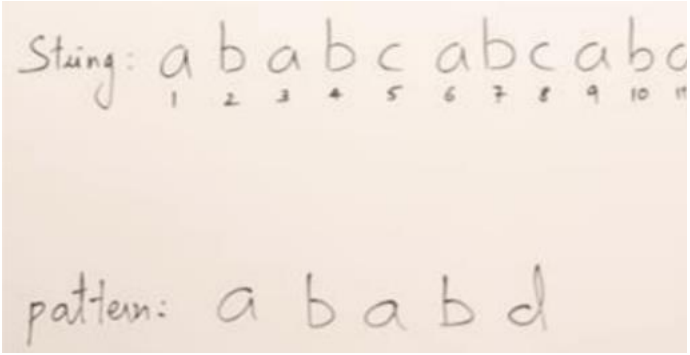
Fig. 4: Testing Stream Architecture

The database contains the full text of the items. The term detector is the special hardware/software that contains all of the terms being searched for and in some systems the logic between the items. It will input the text and detect the existence of the search terms. It will output to the query resolver the detected terms to allow for final logical processing of a query against an item. The query resolver performs two functions.

Now find out substrings as prefix, suffix by taking any number of characters from left to right and right to left. Prefix: a, ab, abc, abcdetc Suffix: c, bc, abc, dabcetc From above prefix, suffix substrings we can observe a substring “abc” is there in both and also that is repeated twice in given pattern.



Example: Given string and pattern is



VI. CONCLUSION

At last we conclude that, information retrieval is a process of searching and retrieving the knowledge based information from collection of documents. This studied has dealt with the basics of the information retrieval. In first section we are defining the information retrieval system with their basic measurements. After this we concerns with traditional IR models and also discuss about the different indexing techniques and searching techniques. This paper also includes the area of IR applications.

VII. REFERENCES

[1] M.François Sy, S.Ranwez, J.Montmain,"User centered and ontology based information Retrieval system for life sciences", BMC Bioinformatics,2105.

[2] R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", IJCER, sep 2012, Vol. 2 Issue. 5, , PP: 1443-1444,.

[3] Anwar A. Alhenshiri, "Web Information Retrieval and Search Engines Techniques",2010,Al- Satil journal,PP: 55-92.

[4] D.Hiemstra,P. de Vries, "Relating the new language models of information retrieval to the traditional retrieval models", published as CTIT technical report TR-CTIT-00-09, May 2000.

[5] Djoerd Hiemstra, "Information Retrieval Models", published in Goker, A., and Davies, J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, November 2009,Ltd., ISBN-13: 978-0470027622.

[6] Christos Faloutsos, Douglas W. Oard, "A Survey of Information Retrieval and Filtering Methods", CS-TR-3514, Aug 1995. "Algorithms for Information Retrieval – Introduction", Lab module 1.

[7] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval",2009, ACM Press, ISBN: 0-201-39829-X.

[8] S.E. Robertson and K. Sparck Jones. "Relevance weighting of search terms. Journal of the American Society for Information Science", 1976, 27:129–146.

[9] G. Salton and M.J. McGill, "editors. Introduction to Modern Information Retrieval". McGraw-Hill ,1983.

[10] H. Turtle, "Inference Networks for Document Retrieval". Ph.D. thesis, Department of Computer Science,University of Massachusetts, Amherst, MA 01003. Available as COINS Technical Report 90-92, 1990.

[11] C. J. van Rijsbergen. "Information Retrieval. Butterworths", London,1979.

[12] T. Strzalkowski, L. Guthrie, J. Karlgren, J. and et. "Natural language information retrieval: TREC-5 report". In Proceedings of the Fifth Text REtrieval Conference (TREC-5), 1997.

[13] Gerard Salton and M. J. McGill. “Introduction to Modern Information Retrieval”. McGraw Hill Book Co.,New York, 1983.
[14] Gerard Salton and Chris Buckley. “Termweighting approaches in automatic text retrieval”. Information Processing and Management, , 1988, 24(5):513–523.
[15] Gerard Salton, editor. “The SMART Retrieval System—Experiments in Automatic Document Retrieval”.Prentice Hall Inc., Englewood Cliffs, NJ, 1971.
[16] N. J. Belkin and W. B. Croft.” Information filtering and information retrieval: Two sides of the same coin? “,Communications of the ACM, 1992,35(12):29–38.

Cite this article as :

Tanmayee Tushar Parbat, Rohan Benhal, Honey Jain, "Document Retrieval Techniques using Vector Space Model ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 5, pp. 93-99, September-October 2021. Journal URL : https://ijsrcseit.com/CSEIT217680