JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

INFORMATION RETRIEVAL: A REVIEW ON INFORMATION PROCESSING AND EXTRACTION IN MACHINE LEARNING

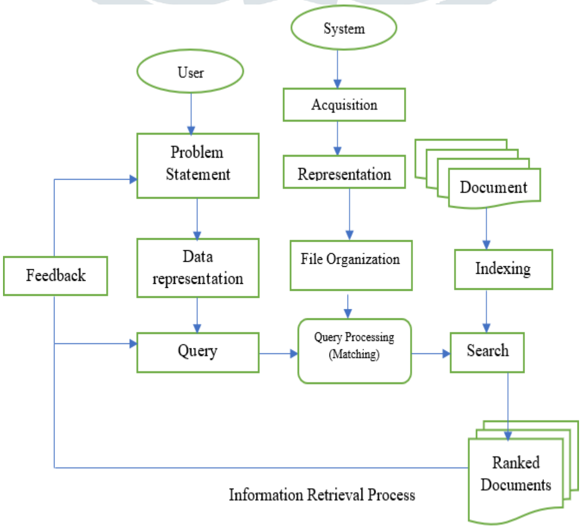
G.Balasaranya[1]., V.Lavanya[2].,
Assistant Professor[1,2], G.T.N. Arts College[1,2], Dindigul, India.

ABSTRACT
From past few decades, we know the importance of searching and storing of information. As the frequency usage of computers increases for past few years, information retrieval is the most important and become necessary option in our daily life. Information retrieval is a method of finding a material most probably documents from large collection of unstructured formats of data preferably text. In this modern era, information retrieval is based on computer science that deals with the representation, storageand retrieval of information. Information retrieval is a process that starts when a user starts searching some data. The results may depend on the search techniques used. In this review paper one can gather complete knowledge regarding the information retrieval process, IR models and IR extraction mechanisms and IR application areas in machine learning.

Keywords: *Information Retrieval (IR), IR Model, IR process, Classical Model*

1. Introduction

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 called a query [3]. When a query is given, it searches the database and results are given according to the query [4]. The IR process involves various stages starting with representing the data followed by indexing, filtering, searching, matching, ranking and ending with returning relevant information to the user [1]. The diagrammatic representation of IR process is as follows:



Information Retrieval Process

2. Information Retrieval Models:

In order to achieve the IR process, we have various models [5]. The models are Classical model,Non-Classical Model and Alternative IR Model.

Classic IR model:

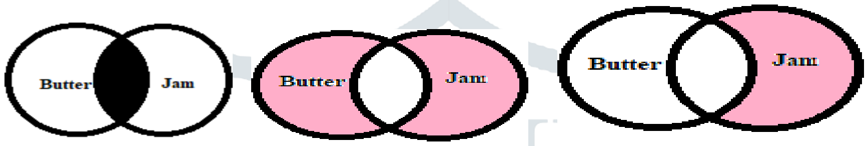
It is the most basic and straightforward IR model. This paradigm is founded on mathematical information that was easily recognized. Three traditional IR models are Boolean, Vector, and Probabilistic [5].

Boolean Model:

The Boolean model is the first form of information retrieval based on logical algebra and the principle of Exact Match. There is no room for partial matching in this form. Here documents are represented by a set of terms (also known as index terms). Then it is classified into a class in which the terms of the query mentioned, and a class in which the terms not mentioned. This classification means that there is no sort of arrangement in evaluating the relevance of the documents to the query. The user'sneeds are identified by a combination of basic logical transactions defined by George Bool, which are three (AND, OR, NOT) meaning intersection, addition and difference, and are used during formulation of the query [6] [7]. One major disadvantage is a Boolean system is not able to rank the returned list of documents. In this 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 [1][7]. The diagrammatic representation of AND, OR and NOT with simple Butter, Jam as example is follows:

Vector Space Model:

Gerard Salton and his colleagues suggested this model in 1983. This model is based on similarity criterion proposed by Hans Peter Luhn in 1957. This is a statistical model for searching of information based on the similarity criterion between inquiries and documents. The similarity criterion as follows:



Butter AND Jam

Butter OR Jam

NOT Butter
BUT Jam

"The more two representations agreed in given elements and their distribution, the higher wouldbe the probability of their representing similar information" [6].

Based on this criterion, Salton and his colleagues considered that both documents and queries could be represented as vectors in Euclidean space, so that each term is assigned an independent dimension, and then they calculated the similarity between vectors using the cosine between the vectors representing both the document and the query [1]. Vector Space Model introduced the 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 for measuring the inverse of the number of documents that contain a query or document term[1]. The biggest challenge facing this model is to set the appropriate value for the vector components, and this problem is Term Weighting [6].

Probabilistic Model:

It is based on the Probability Ranking Principle. In information retrieval system the documentsare ranked based on their probability of relevance to the query. The principle takes into account that there is uncertainty in the representation of the information need and the documents. There can be a variety of sources of evidence that are used by the probabilistic retrieval methods, and the most common one is the statistical distribution of the terms in both the relevant and non-relevant documents [8][6]. 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” [1][6].

Non-Classic IR model:

It is just opposite to the classic IR model. They are based on the principles other than similarity, probability and Boolean operations. They are built upon propositional logic to combine documents and queries in some representation and suitable logic. Propositional logic is also known as sentential logic isa branch of logic is a way of combining or altering statements or propositions to form more complicatedstatements or propositions [9]. Some of the Non-classical IR models include situation theory models, information logic models, and interaction models [5].

Alternative IR model:

It is an improvement to the traditional IR model that makes use of some unique approaches from other domains. Alternative IR models include fuzzy models, cluster models, and latent semantic indexing (LSI) models.

Fuzzy Model:

This model is a fuzzy generalization of the Boolean model. The fuzzy information retrieval model defines the fuzzy relationship between query language and the retrieved documents. It assumes that a set of fuzzy documents is associated with each word in the query language.

Each word in the query language defines a fuzzy set, and the elements in the sets are retrieved documents. Correspondingly, each document in the set has a degree of membership to correspond to eachword in the query language. As a retrieval result, the fuzzy set reflects how well each document matches the query [10].

Cluster Model:

Clustering is used in information retrieval systems to enhance the efficiency and effectiveness of the retrieval process. Clustering is achieved by partitioning the documents in a collection into classessuch that documents that are associated with each other are assigned to the same cluster. This association is generally determined by examining the index term representation of documents or by capturing user feedback on queries on the system. In cluster-oriented systems, the retrieval process can be enhanced by employing characterization of clusters [11].

Latent Semantic Indexing (LSI) Model:

Latent Semantic Indexing is a technique that projects queries and documents into a space with “latent” semantic dimensions. In the latent semantic space, a query and a document can have high cosine similarity even if they do not share any terms - as long as their terms are semantically similar. The latent semantic space that we project into has fewer dimensions than the original space (which has as many dimensions as terms). LSI is thus a method for dimensionality reduction. A dimensionality reduction technique takes a set of objects that exist in a high-dimensional space and represents them in a lowdimensional space, often in a two-dimensional or three-dimensional space for the purpose of visualization. Latent semantic indexing is the application of a particular mathematical technique, called Singular Value Decomposition or SVD, to a word-by-document matrix. SVD (and hence LSI) is a least- squares method [12].

3. Information Retrieval Extraction Process:

The main aim of the Information Retrieval System is to find relevant information or a document that satisfies user information needs. To achieve this the system has to follow various process like indexing, filtering and searching. The two basic measures for accessing the quality of information are Precision and Recall [1][2][4].

Precision is defined as the proportion of retrieved documents which are actually relevant to thequery derived from the user request [2].

Precision = Not Relevant / Total Retrieved

Recall is the proportion of documents known to be relevant to the query in the entire collectionthat have been retrieved in the retrieved document list for that query [2].

Recall = Not Relevant Retrieved / Total Known Relevant

Indexing:

A basic definition of indexing was given in 1988 by Salton as the facilitation of information retrieval accuracy by collecting, parsing and storing data. The common techniques are inverted files, suffix trees and signature files [14].

A. Inverted Files:

Inverted files are defined as central components of an indexing algorithm in a search engine. The engine that searches information has a goal of query speed optimization. This means finding documents where a certain word occurs. The next step is developing a forward index. The index that is developed plays a role of storing the lists of words in every document. The document is then inverted, leading to a developed inverted index. [1][14].

B. Suffix Trees:

A suffix tree is a Trie that is compressed and contains suffixes of the given texts as the keys thatbelong to them as well as their values as the positions present in the text. The idea of compressing tries makes suffix trees be referred to as tries. Consequently, the sub-trees are referred to as sub-tries. The concept of suffix trees was developed in the year 1973 by Weiner.The use of suffix trees is applied when solving multiple string problems occurring in free text search as well as text editing. Suffix trees are also used in computational biology as well as other areas of application [14].

C. Signature File:

In this method each document yields a bit string using hashing on its words and superimposed coding. The resulting document signatures are stored sequentially in a separate file called signature file, which is much smaller than the original file and can be searched much faster.

Filtering and Searching:

Filtering is a process that all the stop words and common words are removed. Searching is the core process of information retrieval techniques [1]. There are various searching algorithms, including linear search, binary search, brute force search etc [1] are used as searching techniques. Stemming, crawling, stop word elimination query and relevance feedback are mostly techniques in filtering and searching [15].

4. Information Retrieval Applications:

Information retrieval (IR) services are computer-based systems that allow users to search and retrieve documents, websites, and other types of information from a database or a collection of documents. These services are designed to help users to find relevant information quickly and efficiently[15]. The applications are,

Search Engines: These are the most common type of IR service, and they allow users to search the Internet for websites, documents, and other types of information. Desktop search, Enterprise search, Federated search, Mobile search, and social search are also popular searches [1][15].

Digital Library: These IR services allow users to search for books, journals, and other materials in a library's collection. The digital content may be stored locally, or accessed remotely via computer networks [1][15].

Document Databases: These IR services allow users to search for documents within a specific database or collection, such as a database of research papers or legal documents [15].

Specialized IR Services: These are IR services that are designed to search specific types of information, such as medical literature or patents [15].

IR services use various techniques to index and retrieve information, including keyword searches, natural language processing, and machine learning algorithms. They may also use metadata, such as author names, publication dates, and subject tags, to help users find relevant information.

5. Conclusion

At last, we conclude that, information retrieval is a process of searching and extracting information from collection of databases. This paper dealt with detailed study of various information retrieval models and its types. Then it includes the extraction process and various areas of information retrieval application areas.

References:

[1] Akram Roshdi and Akram Roohparvar, Review: Information Retrieval Techniques and Application, International Journal of Computer Networks and Communications Security VOL.3, NO. 9, SEPTEMBER 2015, 373–377

[2] Dr. M Hanumanthappa, Deepa T. Nagalavi, A survey on Information Retrieval System for Online Newspapers, International Journal of Engineering Research & Technology (IJERT),
NCSE'14 Conference Proceedings

[3] Christopher D. Manning, Prabhakar Raghavan, "Introduction to Information Retrieval" ,
University of Stuttgart, Cambridge University, 2008

[4] https://www.researchgate.net/publication/269819624_Survey_Paper_on_Information_Retrieval_

Algorithms_and_Personalized_Information_Retrieval_Concept

[5] https://www.engati.com/glossary/information-retrieval

[6] Manal Sheikh Oghli, Muhammad Mazen Almustafa, "Comparison of basic Information Retrieval Models", IJERT, volume 10, issue 9 sept 2021

[7] https://redirect.cs.umbc.edu/~ian/irF02/lectures/06Models-Boolean.pdf

[8] https://aspoerri.comminfo.rutgers.edu/InfoCrystal/ Ch_2.html

[9] https://www.scaler.com/topics/nlp/nlp-ir-models/

[10] Qiu D, Jiang H, Chen S. Fuzzy Information Retrieval Based on Continuous Bag-of-Words Model. Symmetry. 2020; 12(2):225. https://doi.org/10.3390/sym12020225

[11] Bhatia SK, Deogun JS. Conceptual clustering in information retrieval. IEEE Trans Syst Man Cybern B Cybern. 1998;28(3):427-36. doi: 10.1109/3477.678640. PMID: 18255959.

[12] Rosario, B. (2000). Latent semantic indexing: An overview. *Techn. rep. INFOSYS*, *240*, 1-16.

[13] 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

[14] Zohair Malki, "Comprehensive Study and Comparison of Information Retrieval Indexing Techniques", (IJACSA) International Journal of Advanced Computer Science and Applications,
Vol. 7, No. 1, 2016

[15]https://www.engati.com/glossary/information-retrieval