Grenze International Journal of Engineering and Technology, January Issue

Information Retrieval & Search Engine

Archita Bhatnagar[1], Divya Bhatnagar[2], Gulbir Singh[3], Tejinder Kaur[4], Vivek Bhatnagar[5] and Vivek Kumar Shrivastava[6]
[1]Computer Science Engineering, Faculty of Engineering and Technology, Swami Vivekanand Subharti University, Meerut, UP
Email: architabhatnagar89@gmail.com
[2]Department of Computer Applications Poddar International College, Jaipur, Rajasthan
Email: divyabhatnagar30@gmail.com
[3]Department of Computer Science and Engineering, Graphic Era Hill University Haldwani Campus, Nainital Uttarakhand, India
Email: gulbir.rkgit@gmail.com
[4,5]Department of MMICTBM, Maharishi Markandeshwar (Deemed to Be University), Mullana, Ambala-Haryana 133207, India
Email: jattidhillon55@gmail.com, bhatnagar.78@gmail.com
[6]Computer Science and Engineering Department, MMEC, Mullana Ambala, Haryana
Email:viveks2001@gmail.com

***Abstract*— Data mining is the process of satisfying a user's information needs, which is represented by text queries. Search engines are mostly an online model for finding information. The Internet search problem faces many other challenges, such as the collection of Web resources, the ranking of these resources, and the use of communication tools for searches. While traditional data mining queries only use document content to obtain results, the dynamics of the Web requires stringent quality control methods. In addition, web documents contain valuable metadata and structured text, such as title, author, or text that can be used to improve search accuracy.**

***Index Terms*— Search engine, Retrieval, Information, System, Database, Machine Learning, Technique.**

I. INTRODUCTION

Information retrieval (IR) is usually defined as the process of retrieving relevant information from a set of sources to answer a specific question. It is an important aspect of information management, especially in the digital age where so much information is available online. Search engines are the most common type of IR system designed to help users find the information they need [1]

TABLE 1. SOME DIMENSIONS OF INFORMATION RETRIEVAL [1]

Examples of content	Examples of applications	Examples of tables
Text	Web Search	Adhoc Search
Images	Vertical Search	Filtering
Video	Enterprise Search	Classification
Scanned Documents	Desktop Search	Question Answering
Audio	Peer to Peer Search	Music

Grenze ID: 01.GIJET.11.1.522_1

Here is a description of data mining and search engines:

II. Information Retrieval (IR)

Purpose: The main purpose of an IR system is to provide users with relevant documents or information based on queries. This may include research databases, the Internet, or digital libraries [2]

A. Key Concepts

- Index: Organizing information (documents, web pages) into a structure that allows for quick searching. This involves creating an "inverted index" for textual data, where each word indicates its occurrence in the documents.
- Query: Users submit a query and the system iterates to find relevant documents or data [3]
- Ranking: The system ranks answers according to their relevance to the question. Ranking is influenced by a variety of factors, including term frequency, document frequency, and machine learning models.
- Relevance feedback: Modern IR systems improve based on user feedback, learning the most useful results and adjusting rankings over time [3]

III. Search engine

A search engine is a system that indexes large sets of data (such as the Internet) and allows users to find specific information using queries.

A. The main components of a search engine are [4]

- Web crawling: Search engines use robots to constantly crawl the web and index new or updated content.
- Indexing: Once you publish content, it is indexed using metadata and keywords for easier discovery.
- Ranking algorithms: Search engines rank results based on relevance, often using complex algorithms. For example, Google uses Page Rank and other algorithms to rank web pages.
- User query processing: Search engines analyze and refine user queries, often using natural language processing (NLP) techniques to understand user intent and context [5]

B. Major functions supported by search engines: Indexing process

- Query process. The indexing builds the structures, and the query process produce documents in ranked order [6]

C. Top search engines

- Google: The most widely used search engine in the world, known for its sophisticated algorithms, personalized search, and fast indexing.

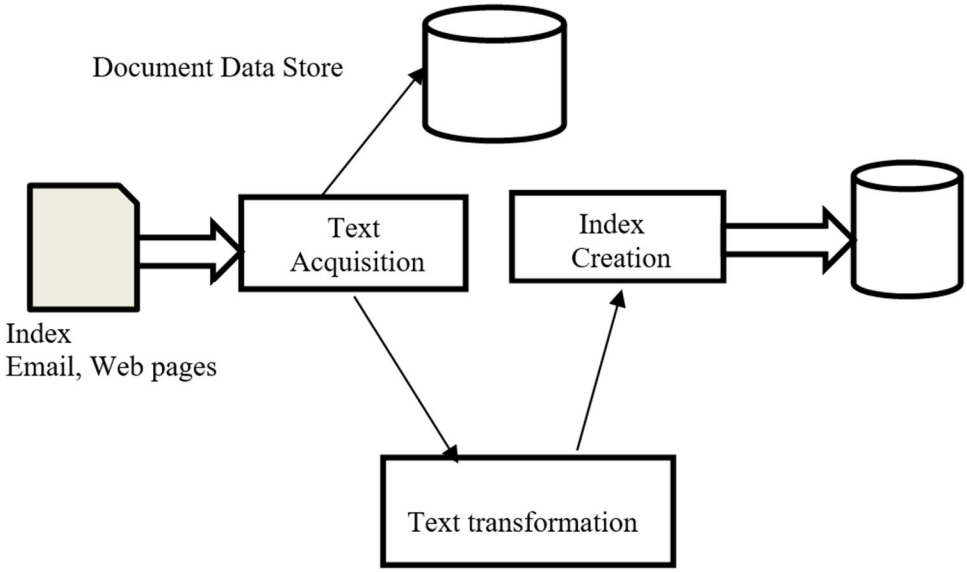

Fig.1.Index Process [1]

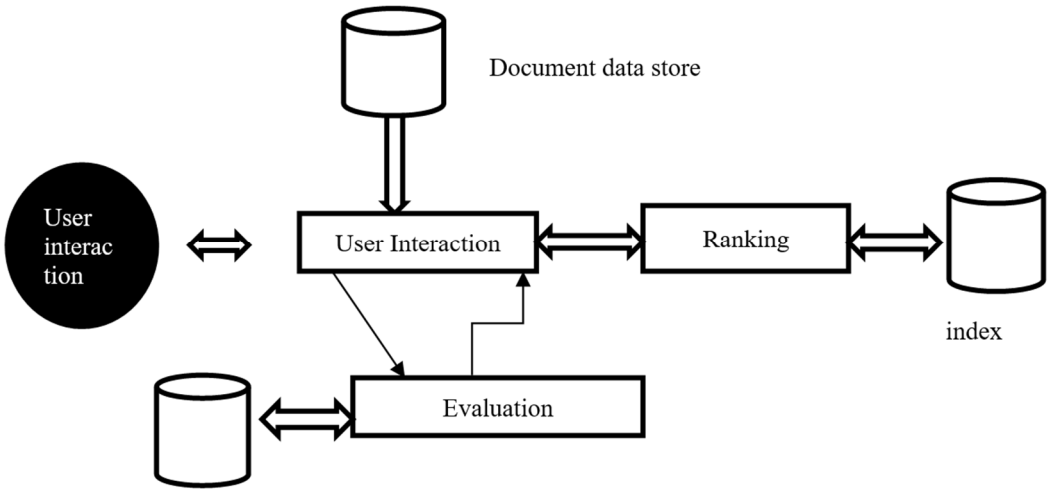

Fig.2. Query process [2]

- Bing: A Microsoft search engine that offers features similar to Google but with different ranking algorithms.
- Yahoo!: Once a dominant player, Yahoo now uses Bing for its search results.
- Duck Go: Focuses on privacy, does not track user behaviour, and provides non-personalized search results.
- Baidu: China's leading search engine optimized for Chinese content [7]

IV. Types of research Questions: [8]

Navigational: The user tries to find a specific website or page (e.g. "Open AI website").
Informative: The user wants to understand the topic (e.g. "What is AI?").
Transaction: The user wants to perform an action, such as making a purchase or downloading something (e.g. "buy a laptop").

A. Search Engine Optimization (SEO):[9]

SEO is the process of optimizing a website for search engine results by optimizing content, using keywords, optimizing website structure, and obtaining back links.

IR System Evaluation:

Precision: Number of documents cited.
Recall: Number of related documents successfully submitted.
F-score: Combination of precision and recall to measure overall system performance.

B. Challenges with IR and Search Engines: [10]

Understanding: A search query can have multiple meanings, making it difficult for the system to understand the user's intent.
Scale: As the amount of data available online grows, a search engine must index and retrieve information efficiently.
Spam and Low-Quality Content: Search engines always find it difficult to process irrelevant or low-quality content.

V. Advanced Research Methods

Boolean search: Allows users to combine keywords with operators such as AND, OR, and NOT for more refined searches.
Match search: Finds documents that contain two or more keywords.
Jokers and truncation: Use characters (such as * or?) to overwrite characters for flexible matching [11]

VI. Components of Information Retrieval System

An IR system consists of various components that extract important information from the collected documents.

A. Display the document: [12]

Word Processing: Documents need to be processed before they can be indexed. It includes:
Tokenization: Breaking text into words, terms, or tokens.
Remove words: Common words (such as "the" and "is") that are not useful for the search will be removed.
Stemming/lemmatization: reducing words to their original form to match different word forms (eg "write" to "write").
Vector representation: Documents are often represented as vectors in a large space (word box or TF-IDF), each shape corresponding to a word in the dictionary.

B. Inverse Index: [13]

Inverse Index: A feature of many IR systems, it lists terms in the documents that refer to them. This allows you to quickly retrieve documents with specific conditions.
Publication list: For each term, the inverse index contains a list of documents (and sometimes fields) that contain the term.

C. Ranking Model: [14]

Classification is a key component of IR. Common models include:
Boolean model: In this simple model, documents are based on the logical operators (and not) in question. They have no classification.
Vector Space Model (VSM): Represents queries and documents as vectors in a multidimensional space. Similarity between them is often calculated using cosine similarity.
Probabilistic models: These models estimate the probability that a document will be relevant when a query is received. Okapi BM25 is one of the most popular probability models used in modern search engines.
Language models: These models estimate the probability of generating a query from a given document. A common approach is the demand probability model (QLM) [15]
Neural networks: Recent advances in machine learning have enabled the use of deep learning models, particularly neural networks and transformer models such as BERT.

VII. Machine Learning in Information Retrieval

Modern IR systems incorporate machine learning techniques to improve classification, relevance, and personalization. The key areas are:

A. Training for Classification (LTR): [16]

Learning classification models are machine learning models designed to classify search results based on relevance. They rely on supervised learning and use features such as time frequency, document length, and click-through rate.
LTR Paper: Each paper is given a score and the papers are ranked according to the score.
Pair wise LTR: The model learns the ranking of each pair of documents and decides which one to rank higher.
List wise LTR: The model optimizes the entire sorted list [17]

B. Deep IR Training

Neural IR models train on large amounts of textual data to detect semantic relationships between words and phrases to make searches more accurate. These models are used for:
Word embeddings: Models like Word2Vec and Glove represent words as vectors capturing their semantic content.
Transformer-based models: Models like BERT (Bilateral Transformer Encoder) have revolutionized IR in understanding the context and meaning of user queries. BERT-based search models like Colbert perform permission-level search.

C. Customization

Personalized search involves customizing search results based on a user's past behaviour, location, preferences, and search history.
Collaborative filtering: This technique, often used in recommendation systems, predicts user preferences based on similar user behaviour.
User profile: Information about users' demographics, interests, and historical data is used to improve the relevance of search results [18]

VIII. Expand and Refine Requirements

Sometimes user requests are too broad or vague. IR systems use a variety of techniques to improve and enhance the following problems:

A. Response to Relevance

After initial responses, the system can prompt users to show relevant and irrelevant documents. Based on this feedback, the question has been refined.
Roche's Algorithm: A Classic Query Expansion Method Based on Fitness Answers. This updates the query vector, moving it toward relevant documents and away from unrelated documents.

B. False Significance Answer

If there is no open-ended answer, the system locates the most relevant documents and automatically determines the question based on them.

C. Question Details

Expand synonyms: Add synonyms or related words to the question to improve memory.
Terminology: Combining frequently occurring terms to expand queries with related terms [19]

IX. Evaluation of IR systems

Evaluation in IR systems is very important to determine how well the system reads relevant documents. Common metrics include:
Accuracy: Percentage of matching documents received.
Rejection: Percentage of valid documents received.
F-score: A harmonic mean of precision and recall that provides a stable measure.
Mean Accuracy (MAP): A measure of accuracy at each position in a sorted list.
Normalized Discounted Cumulative Gain (NDCG): takes into account the position of relevant documents in the ranking, giving more weight to documents ranked first in the results.

X. Modern Search Engine Architecture

Modern search engines like Google and Bing have very complex architectures with many features designed for speed, accuracy and scale.

A. Web Verification

Web crawlers, or spiders, crawl the web to find and download web pages to be indexed.
Courtesy: Searchers should avoid overloading the servers by limiting the number of requests to the site within a certain period of time.

B. Indexing and Compression

Depending on the scale of the data, search engines use compression techniques such as delta-encoding and Gloom-Rice encoding to reduce the size of the index.
Sharing and Replication: An index is often distributed across multiple servers (shading) and parts of it can be replicated for fault tolerance and faster scans.

C. Ranking Algorithm

Page Rank: Google's original algorithm that scores each page based on the number and quality of directing links.
Click patterns: Many modern search engines adjust their rankings based on user interactions, such as clicks and wait times.
Contextual and semantic detection: Search engines use techniques like NLP and machine learning to better understand the meaning and intent of a query.

XI. Challenges and Future Trends

A. Experiments

Big Data Processing: With the ever-increasing amount of online content, IR systems face the challenge of efficiently indexing and extracting relevant information from big data.

Real-time search: With the rise of social media and other fast-updating sources, search engines need to generate search results in real-time or real-time.
Privacy Concerns: Personal searches and data collection raise privacy concerns, leading to privacy-conscious search engines like Duck Go [20]

B. Future Trends

Voice Search and Conversational IR: With the rise of virtual assistants like Siri, Alexa, and Google Assistant, voice search and conversational IR are becoming increasingly important. These systems focus on understanding natural language and providing concise and relevant answers.
Multimodal search: The future of IR may involve combining multiple types of data for more comprehensive search, such as combining text, images and video.
AI-powered personalization: With the advancement of AI, search engines are able to understand users' needs, preferences and context, providing more personalized search results.

REFERENCES

[1] McBryan 94 Oliver A. McBryan. GENVL and WWWW: Tools for Taming the Web. First International Conference on the World Wide Web. CERN, Geneva (Switzerland), May 25-26- 27 1994. http://www.cs.colorado.edu/home/mcbryan/mypa pers/www94.ps
[2] Sunny lam," The Overview of Web Search Engines", February 2001
[3] J Pei, "Information retrieval and web search architecture", Lecture slides, 2017
[4] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, New York, NY, USA, 1999
[5] Comscore report, February 2016
[6] G. Chowdhury Introduction To Modern Information Retrieval, 3rd Ed Facet Publishing, 2010
[7] R. Baeza-Yates, B. Ribeiro-Neto, "Modern information retrieval, ACM Press, New York, USA, 1999.
[8] Alis J. Technologies, 'Web languages hit parade. http://babel.alis.com:8080/palmares.html, 1997.
[9] K. Andrews, "Visualizing cyberspace: information visualization in the harmony internet browser," in Proceedings '95 information visualization, Atlanta, USA, October 1995, pp. 97-104.
[10] K. S. Jones, P. Willett, "Readings in information retrieval, Morgan Kaufmann., 1997. [6] G. Kowalski, M.T. Maybury, Information storage and retrieval systems, Springer, 2005.
[11] P. Aniek, J. Brennan, R. Flynn, D. Hanssen, B. Alvey, and J. Robbins, "A direct manipulation interface for boolean information retrieval via natural language query," in Proc. of the 19th Annual International ACM/SIGIR Conference, Brussels, Belgium, 1990, pp. 135-150.
[12] A. Apostolico and Z. Galil, "Combinatorial algorithms on words. Springer-Verlag, New York, 1985.
[13] www.wikipedia.org "Information Retrieval" Retrieved 22-06- 2011
[14] P. Aniek, "Adapting a full-text information retrieval system to the computer troubleshooting domain," in Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 349- 358,.
[15] S. Amit, "Modern information retrieval: A brief overview, 2022.
[16] ANSI/NISO Standards, Z39.50, "Information retrieval: Application service definition and protocol specification, 2024.
[17] C.T. Meadow, B.R. Boyce, D.H. Kraft, C.L. Barry. Text information retrieval systems. Academic Press, 2019.
[18] C. J. Van Risjbergen, "The geometry of information retrieval, Cambridge UP, 2004.
[19] M. Levene, Search Engines: Information Retrieval in Practice, The Computer Journal , 2021.
[20] W. Bruce Croft et.al Search Engines Information Retrieval in Practice, Pearson education 2019.