 Plovdiv University “Paisii Hilendarski” – Faculty of Mathematics and Informatics

ESSAY

by

COMPUTATIONAL LINGUISTICS

on topic

Search in large text arrays (Information Retrieval). Inverted index. Boolean query processing. Ordering the results.

Name: Tony Todorov

Faculty number: 2001261067

1. ***Definition for Information Retrieval***

Information Retrieval (IR) can be defined as a software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories, particularly textual information. Information Retrieval is the activity of obtaining material that can usually be documented on an unstructured nature i.e. usually text which satisfies an information need from within large collections which is stored on computers. For example, Information Retrieval can be when a user enters a query into the system.

Not only librarians, professional searchers etc, engage themselves in the activity of information retrieval but nowadays hundreds of millions of people engage in IR every day when they use web search engines. Information Retrieval is believed to be the dominant form of Information access. The IR system assists the users in finding the information they require but it does not explicitly return the answers to the question. It notifies regarding the existence and location of documents that might consist of the required information. Information retrieval also extends support to users in browsing or filtering document collection or processing a set of retrieved documents. The system searches over billions of documents stored on millions of computers. A spam filter, manual or automatic means are provided by Email program for classifying the mails so that it can be placed directly into particular folders.

An IR system has the ability to represent, store, organize, and access information items. A set of keywords are required to search. Keywords are what people are searching for in search engines. These keywords summarize the description of the information.

1. ***History***

The idea of using computers to search for relevant pieces of information was popularized in the article As We May Think by Vannevar Bush in 1945. It would appear that Bush was inspired by patents for a 'statistical machine' – filed by Emanuel Goldberg in the 1920s and 1930s – that searched for documents stored on film. The first description of a computer searching for information was described by Holmstrom in 1948, detailing an early mention of the Univac computer. Automated information retrieval systems were introduced in the 1950s: one even featured in the 1957 romantic comedy, Desk Set. In the 1960s, the first large information retrieval research group was formed by Gerard Salton at Cornell. By the 1970s several different retrieval techniques had been shown to perform well on small text corpora such as the Cranfield collection (several thousand documents). Large-scale retrieval systems, such as the Lockheed Dialog system, came into use early in the 1970s.

In 1992, the US Department of Defense along with the National Institute of Standards and Technology (NIST), cosponsored the Text Retrieval Conference (TREC) as part of the TIPSTER text program. The aim of this was to look into the information retrieval community by supplying the infrastructure that was needed for evaluation of text retrieval methodologies on a very large text collection. This catalyzed research on methods that scale to huge corpora. The introduction of web search engines has boosted the need for very large scale retrieval systems even further.

there is ... a machine called the Univac ... whereby

letters and figures are coded as a pattern of

magnetic spots on a long steel tape. By this

means the text of a document, preceded by

its subject code symbol, can be recorded ...

the machine ... automatically selects and

types out those references which have been

coded in any desired way at a rate of 120 words a minute

- *J. E. Holmstrom, 1948*

1. ***Types of information retrieval systems***

Information retrieval systems can be categorized into several types based on their functionality, purpose, and the nature of the content they retrieve. Here are some common types of information retrieval systems:

* **Web Search Engines:** Web search engines like Google, Bing, and Yahoo are perhaps the most widely used information retrieval systems. They retrieve web pages and other online content in response to user queries, providing relevant search results based on algorithms considering factors like relevance and authority.
* **Digital Libraries:** Digital library retrieval systems focus on organizing and retrieving digital versions of books, academic papers, journals, and other scholarly resources. These systems are used primarily for research and education purposes.
* **Enterprise Search:** Enterprise search systems are designed to help organizations retrieve and manage internal documents and data. They enable employees to find information within the organization’s databases, content management systems, emails, and other repositories.
* **Multimedia Retrieval Systems:** These systems retrieve multimedia content such as images, audio, and video. Applications include image search engines, music recommendation systems, and video content retrieval platforms.
* **Content Recommendation Systems:** Recommendation systems, like those used by streaming services (e.g., Netflix and Spotify), retrieve and recommend content based on user preferences, viewing history, and behavior.
* **Question-Answering Systems:** Question-answering systems like chatbots and virtual assistants retrieve specific answers to user questions, often by searching through a knowledge base or a predefined set of documents.
* **Geographic Information Systems (GIS):** GIS systems retrieve and display geographic information, maps, and spatial data. They are used in urban planning, environmental management, and navigation.
* **Cross-Language Information Retrieval (CLIR):** CLIR systems retrieve information in one language in response to queries expressed in another. They are crucial for multilingual information access.
* **Personalized Information Retrieval Systems:** These systems tailor search results and recommendations to individual users based on their preferences, behavior, and history. They are common in e-commerce and content recommendation.
* **Vertical Search Engines:** Vertical search engines focus on specific niches or industries. Examples include job search engines, real estate search engines, and medical literature search engines.
* **Meta-Search Engines:** Meta-search engines aggregate results from multiple search engines and present them to users. They aim to provide a more comprehensive view of search results.
* **Social Media Search:** Social media platforms have their search functionality, allowing users to search for posts, images, videos, and other content.
* **Desktop Search:** Desktop search tools help users find files and documents on their local computer or networked drives. Examples include the search functionality in Windows and macOS.
* **Legal and Patent Search:** Specialized search systems are used in the legal field and for patent searches to retrieve specific legal documents and patent information.
* **Image Retrieval Systems:** These systems enable users to search for images based on visual content, such as color, shape, and texture, rather than textual keywords.
* **Medical Information Retrieval:** Information retrieval systems for healthcare professionals help retrieve medical literature, patient records, and clinical guidelines.
* **News Aggregation:** News aggregation systems gather and retrieve news articles and updates from various sources to provide users with a comprehensive view of current events.

1. ***Model types***

Information retrieval models are mathematical and conceptual frameworks used in information retrieval to represent and describe the process of retrieving relevant documents or information from a collection in response to a user’s query. These models help search engines and other retrieval systems rank and retrieve documents based on their relevance to a query. Here are some standard information retrieval models:

* **Boolean Model:** The Boolean model is based on set theory and uses Boolean operators (AND, OR, NOT) to combine query terms and retrieve documents that exactly match the query. It’s a straightforward but rigid model.
* **Vector Space Model (VSM):** The VSM represents documents and queries as vectors in a multi-dimensional space, with each dimension corresponding to a term in the vocabulary. The similarity between a query vector and a document vector is used to rank documents. Cosine similarity is often used as the similarity metric.
* **Probabilistic Model:** Probabilistic models treat the retrieval process as a probabilistic event. They calculate the probability that a document is relevant to a query based on term frequency and document length. Models like the Okapi BM25 and Language Models for Information Retrieval (LMIR) fall into this category.
* **Term Frequency-Inverse Document Frequency (TF-IDF) Model:** TF-IDF is a statistical measure that evaluates the importance of a term within a document relative to a collection of documents. Documents are ranked based on TF-IDF scores.
* **Language Models:** Language models, such as the BM25 language and Dirichlet Prior models, estimate the probability of observing a query given a document. These models are based on probabilistic concepts.
* **Latent Semantic Indexing (LSI) Model:** LSI analyzes the latent structure within a collection of documents to discover relationships between terms and documents. It reduces the dimensionality of the term-document matrix using singular value decomposition (SVD) and captures semantic similarities.
* **Latent Dirichlet Allocation (LDA):** LDA is a topic modelling technique for document retrieval. It assigns documents to topics and allows users to search for related documents.
* **Neural Information Retrieval Models:** With the rise of deep learning, neural network-based models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers have been applied to information retrieval tasks. Models like BERT and its variants are used for contextual understanding and ranking.
* **Fuzzy Retrieval Models:** Fuzzy retrieval models consider approximate matches to query terms, allowing for documents with terms similar to the query terms to be retrieved.
* **Feedback Models:** Feedback models use user feedback (relevance judgments) to improve retrieval results in subsequent searches. Relevance feedback and pseudo-relevance feedback are examples of this approach.
* **Distributed Representation Models:** These models represent words and documents as dense vectors in a continuous space, enabling the capture of semantic relationships. Word2Vec and Doc2Vec are examples of such models.
* **Deep Reinforcement Learning for IR:** Recent research has explored using reinforcement learning techniques to optimize information retrieval processes, learning to rank documents based on user interactions.

1. ***Standard Boolean model***

The (standard) Boolean model of information retrieval (BIR) is a classical information retrieval (IR) model and, at the same time, the first and most-adopted one. The BIR is based on Boolean logic and classical set theory in that both the documents to be searched and the user's query are conceived as sets of terms (a bag-of-words model). Retrieval is based on whether or not the documents contain the query terms and whether they satisfy the boolean conditions described by the query.

* **Advantages**
  + Clean formalism
  + Easy to implement
  + Intuitive concept
  + If the resulting document set is either too small or too big, it is directly clear which operators will produce respectively a bigger or smaller set.
  + It gives (expert) users a sense of control over the system. It is immediately clear why a document has been retrieved given a query.
* **Disadvantages**
  + Exact matching may retrieve too few or too many documents.
  + Hard to translate a query into a Boolean expression.
  + All terms are equally weighted.
  + More like data retrieval than information retrieval.
  + Retrieval based on binary decision criteria with no notion of partial matching.
  + No ranking of the documents is provided (absence of a grading scale).
  + Information need has to be translated into a Boolean expression, which most users find awkward.
  + The Boolean queries formulated by the users are most often too simplistic.
  + The model frequently returns either too few or too many documents in response to a user query.

From a pure formal mathematical point of view, the BIR is straightforward. From a practical point of view, however, several further problems should be solved that relate to algorithms and data structures, such as, for example, the choice of terms (manual or automatic selection or both), stemming, hash tables, inverted file structure, and so on.

* **Hash sets:** Each document is represented by a hash table which contains every single term of that document. Since hash table size increases and decreases in real time with the addition and removal of terms, each document will occupy much less space in memory. However, it will have a slowdown in performance because the operations are more complex than with bit vectors. On the worst-case performance can degrade from O(n) to O(n2). On the average case, the performance slowdown will not be that much worse than bit vectors and the space usage is much more efficient.
* **Signature file:** Each document can be summarized by Bloom filter representing the set of words in that document, stored in a fixed-length bitstring, called a signature. The signature file contains one such superimposed code bitstring for every document in the collection. Each query can also be summarized by a Bloom filter representing the set of words in the query, stored in a bitstring of the same fixed length. The query bitstring is tested against each signature.
* **Inverted** **file:** An inverted index file contains two parts: a vocabulary containing all the terms used in the collection, and for each distinct term an inverted index that lists every document that mentions that term.

1. ***Extended Boolean model***

The Extended Boolean model is an expansion of the traditional Boolean model used in information retrieval systems. In the traditional Boolean model, documents are represented as sets of terms, and queries are expressed using Boolean operators such as AND, OR, and NOT to retrieve relevant documents. The Extended Boolean model extends the traditional Boolean model by incorporating additional features such as term weighting and relevance feedback. Here are some key aspects of the Extended Boolean model:

* **Term Weighting:** Unlike the traditional Boolean model, where all terms are considered equally important, the Extended Boolean model assigns weights to terms based on their importance in the document collection. This allows for more sophisticated ranking of documents based on the relevance of terms.
* **Relevance Feedback:** The Extended Boolean model incorporates feedback from users to improve the relevance of search results. Users can provide feedback on the relevance of retrieved documents, which can then be used to refine subsequent searches.
* **Proximity Searching:** The Extended Boolean model supports proximity searching, allowing users to specify that terms should occur within a certain distance of each other in the retrieved documents. This is useful for capturing the context of terms within a document.
* **Query Expansion:** Another feature of the Extended Boolean model is query expansion, where additional terms related to the original query terms are automatically added to the search to improve recall and precision.

1. ***Reverse index***

Reverse index (also known as inverted) is a data structure used in information retrieval systems to efficiently retrieve documents or web pages containing a specific term or set of terms. In an inverted index, the index is organized by terms (words), and each term points to a list of documents or web pages that contain that term.

Inverted indexes are widely used in search engines, database systems, and other applications where efficient text search is required. They are especially useful for large collections of documents, where searching through all the documents would be prohibitively slow.

An inverted index is an index data structure storing a mapping from content, such as words or numbers, to its locations in a document or a set of documents. In simple words, it is a hashmap-like data structure that directs you from a word to a document or a web page.

* **Advantages**
  + The inverted index is to allow fast full-text searches, at a cost of increased processing when a document is added to the database.
  + It is easy to develop.
  + It is the most popular data structure used in document retrieval systems, used on a large scale for example in search engines.
* **Disadvantages**
  + Large storage overhead and high maintenance costs on updating, deleting, and inserting.
  + Instead of retrieving the data in decreasing order of expected usefulness, the records are retrieved in the order in which they occur in the inverted lists.
* **Features**
  + **Efficient search:** Inverted indexes allow for efficient searching of large volumes of text-based data. By indexing every term in every document, the index can quickly identify all documents that contain a given search term or phrase, significantly reducing search time.
  + **Fast updates:** Inverted indexes can be updated quickly and efficiently as new content is added to the system. This allows for near-real-time indexing and searching for new content.
  + **Flexibility:** Inverted indexes can be customized to suit the needs of different types of information retrieval systems. For example, they can be configured to handle different types of queries, such as Boolean queries or proximity queries.
  + **Compression:** Inverted indexes can be compressed to reduce storage requirements. Various techniques such as delta encoding, gamma encoding, variable byte encoding, etc. can be used to compress the posting list efficiently.
  + **Support for stemming and synonym expansion:** Inverted indexes can be configured to support stemming and synonym expansion, which can improve the accuracy and relevance of search results. Stemming is the process of reducing words to their base or root form, while synonym expansion involves mapping different words that have similar meanings to a common term.
  + **Support for multiple languages:** Inverted indexes can support multiple languages, allowing users to search for content in different languages using the same system.

To search for documents containing a particular term or set of terms, the search engine queries the inverted index for those terms and retrieves the list of documents associated with each term. The search engine can then use this information to rank the documents based on relevance to the query and present them to the user in order of importance. There are two types of inverted indexes:

* **Record-Level Inverted Index:**

Record Level Inverted Index contains a list of references to documents for each word.

* **Word-Level Inverted Index:**

Word Level Inverted Index additionally contains the positions of each word within a document. The latter form offers more functionality but needs more processing power and space to be created.

1. ***What is document retrieval and how does it work?***

Document retrieval is a fundamental component of information retrieval, and it involves finding and retrieving specific documents or pieces of information from a collection or database of documents. This process is used in various contexts, including search engines, digital libraries, content management systems, and enterprise document management. Here are the key steps and considerations in document retrieval:

* **Document Collection:** Document retrieval begins with a collection of documents. These documents can be in various formats, including text documents, web pages, PDFs, images, videos, or any other type of digital content.
* **Indexing:** The documents are typically preprocessed and indexed to make retrieval efficient. Indexing involves extracting key information from documents, such as keywords, metadata, and structural information. This information creates a data structure that allows for fast and efficient retrieval.
* **Query:** Users submit queries to the retrieval system. A query can be a single keyword, a phrase, or a complex Boolean expression. In some cases, users may enter natural language queries.
* **Query Processing:** The retrieval system processes the user’s query, which may involve tasks like tokenization (breaking the query into words or phrases), stemming (reducing words to their root form), and removing stop words (common words like “and” or “the” that are not useful for retrieval).
* **Matching:** The system matches the query terms to the indexed documents to identify potential matches. Different retrieval models and algorithms may be used for this step, such as the Vector Space Model or Boolean retrieval.
* **Ranking and Scoring:** If the system is designed to rank documents by relevance, each matching document is assigned a score. Standard scoring methods include TF-IDF (Term Frequency-Inverse Document Frequency) and cosine similarity. Documents are then ranked by their scores.
* **Retrieval:** Based on the ranking, the system retrieves a set of documents considered the most relevant to the user’s query. The number of retrieved documents may vary depending on the design and user preferences.
* **Presentation:** The retrieved documents are presented to the user, typically in a list format. The user interface may include additional features like text snippets from each document, filters, and sorting options.
* **Relevance Feedback:** Some retrieval systems allow users to provide feedback on the relevance of the retrieved documents. This feedback can be used to refine subsequent searches and improve the accuracy of the retrieval system.
* **Evaluation Metrics:** To assess the performance of the document retrieval system, various metrics are used, including precision, recall, F1 score, and Mean Average Precision (MAP). These metrics measure the system’s ability to retrieve relevant documents.
* **Personalization:** In some systems, user preferences and search history are considered to personalize the retrieval results, providing users with content that is more relevant to their interests and needs.

Document retrieval is a critical aspect of modern information systems. It significantly improves access to relevant information in various domains, from web search engines to research databases and digital libraries. Advances in natural language processing and machine learning have also contributed to refining and personalizing document retrieval systems.

*Content:*

Definition for Information Retrieval …………………………………………………………. 2

History ………………………………………………………………………………………... 3

Types of information retrieval systems ……………………………………………………….. 5

Model types …………………………………………………………………………………… 7

Standard Boolean model ……………………………………………………………………… 9

Extended Boolean model ……………………………………………………………………. 10

Inverted index ……………………………………………………………………………….. 12

What is document retrieval and how does it work? …………………………………………. 14

*Sources:*

<https://en.wikipedia.org/wiki/Information_retrieval>

<https://en.wikipedia.org/wiki/Boolean_model_of_information_retrieval>

<https://en.wikipedia.org/wiki/Extended_Boolean_model>

<https://www.geeksforgeeks.org/inverted-index/>

<https://spotintelligence.com/2023/09/05/information-retrieval/>

<https://www.khoury.northeastern.edu/home/jaa/CSG339.06F/Lectures/boolean.pdf>

<https://medium.com/data-science-at-microsoft/search-and-ranking-for-information-retrieval-ir-5f9ca52dd056>

<https://www.pinecone.io/learn/offline-evaluation/>

<https://www.sciencedirect.com/topics/computer-science/retrieval-performance>