INFORMATION RETRIEVAL

IR CONCEPTS

Information retrieval is the process of retrieving documents from a collection in response to a query (or a search request) by a user. Information retrieval is —the discipline that deals with the structure, analysis, organization, storage, searching, and retrieval of information‖ as defined by Gerald Salton, an IR pioneer.

Information in the context of IR does not require machine-understandable structures, such as in relational database systems. Examples of such information include written texts, abstracts, documents, books, Web pages, emails, instant messages, and collections from digital libraries. Therefore, all loosely represented (unstructured) or semi structured information is also part of the IR discipline.

IR systems go beyond database systems in that they do not limit the user to a specific query language, nor do they expect the user to know the structure (schema) or content of a particular database. IR systems use a user's information need expressed as a free-form search request (sometimes called a keyword search query, or just query) for interpretation by the system.

An IR system can be characterized at different levels: by types of users, types of data, and the types of the information need, along with the size and scale of the information repository it addresses. Different IR systems are designed to address specific problems that require a combination of different characteristics. These characteristics can be briefly described as follows:

Types of Users

The user may be an expert user (for example, a curator or a librarian), who is searching for specific information that is clear in his/her mind and forms relevant queries for the task, or a layperson user with a generic information need.

Types of Data

Search systems can be tailored to specific types of data. For example, the problem of retrieving information about a specific topic may be handled more efficiently by customized search systems that are built to collect and retrieve only information related to that specific topic. The information repository could be hierarchically organized based on a concept

or topic hierarchy. These topical domain-specific or vertical IR systems are not as large as or as diverse as the generic World Wide Web, which contains information on all kinds of topics.

Types of Information Need

In the context of Web search, user's information needs may be defined as navigational, informational, or transactional.

Navigational search refers to finding a particular piece of information (such as the Georgia Tech University Website) that a user needs quickly.

The purpose of informational search is to find current information about a topic (such as research activities in the college of computing at Georgia Tech—this is the classic IR system task).

The goal of **transactional search** is to reach a site where further interaction happens (such as joining a social network, product shopping, online reservations, accessing databases, and so on).

RETRIEVAL MODELS

There are the three main statistical models—Boolean, vector space, and probabilistic—and the semantic model.

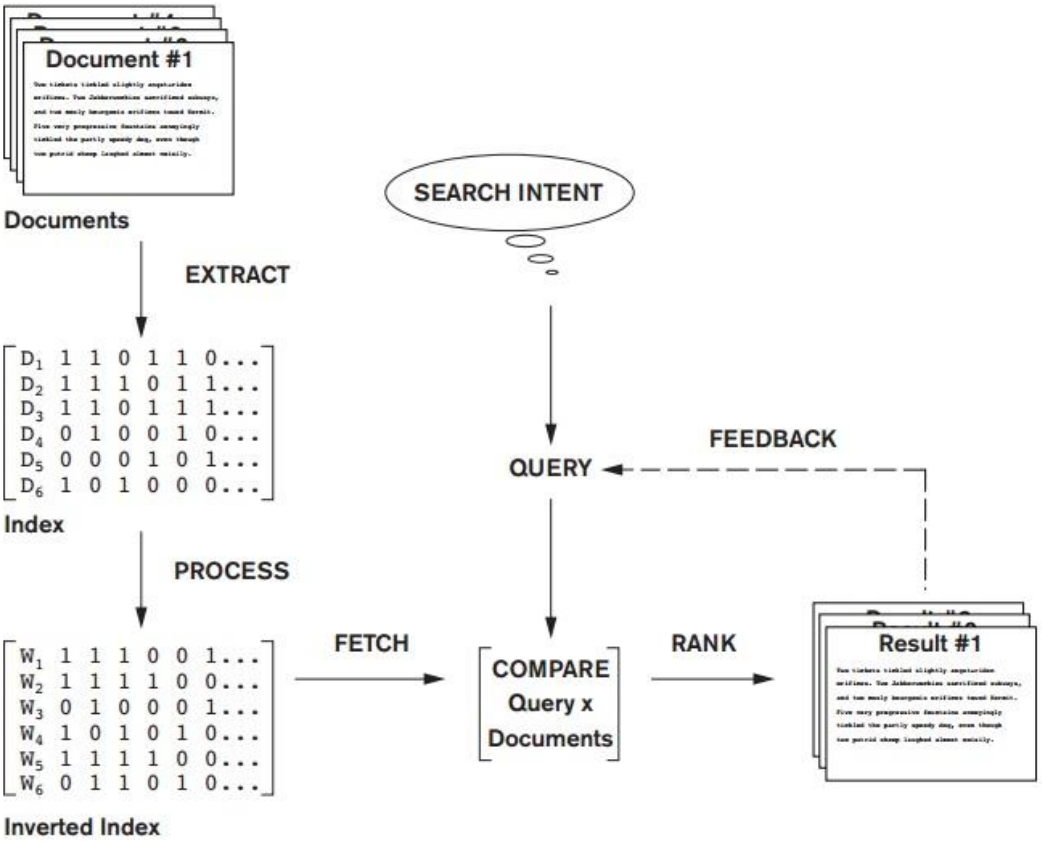

Figure 27.2
Simplified IR process pipeline.

1. **Boolean Model**

In this model, documents are represented as a set of terms. Queries are formulated as a combination of terms using the standard Boolean logic set-theoretic operators such as AND, OR and NOT. Retrieval and relevance are considered as binary concepts in this model, so the retrieved elements are an —exact match‖ retrieval of relevant documents.

Boolean retrieval models lack sophisticated ranking algorithms and are among the earliest and simplest information retrieval models. These models make it easy to associate metadata information and write queries that match the contents of the documents as well as other properties of documents, such as date of creation, author, and type of document.

2. **Vector Space Model**

The vector space model provides a framework in which term weighting, ranking of retrieved documents, and relevance feedback are possible. Documents are represented as features and weights of term features in an n dimensional vector space of terms. Features are a subset of the terms in a set of documents that are deemed most relevant to an IR search for this particular set of documents.

The process of selecting these important terms (features) and their properties as a sparse (limited) list out of the very large number of available terms (the vocabulary can contain hundreds of thousands of terms) is independent of the model specification. The query is also specified as a terms vector (vector of features), and this is compared to the document vectors for similarity/relevance assessment.

In the vector model, the document term weight wij (for term i in document j) is represented based on some variation of the TF (term frequency) or TF-IDF (term frequency-inverse document frequency) scheme (as we will describe below). TF-IDF is a statistical weight measure that is used to evaluate the importance of a document word in a collection of documents. The following formula is typically used:

$$\text{cosine}(d_j, q) = \frac{\langle d_j \times q \rangle}{\| d_j \| \times \| q \|} = \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|V|} w_{iq}^2}}$$

In the formula given above, we use the following symbols:

- dj is the document vector.
- q is the query vector.
- wij is the weight of term i in document j.
- wiq is the weight of term i in query vector q.
- |V| is the number of dimensions in the vector that is the total number of important keywords (or features).

3. Probabilistic Model

In the probabilistic framework, the IR system has to decide whether the documents belong to the **relevant set** or the **nonrelevant set** for a query. To make this decision, it is assumed that a predefined relevant set and nonrelevant set exist for the query, and the task is to calculate the probability that the document belongs to the relevant set and compare that with the probability that the document belongs to the nonrelevant set.

Given the document representation D of a document, estimating the relevance R and nonrelevance NR of that document involves computation of conditional probability P(R|D) and P(NR|D). These conditional probabilities can be calculated using Bayes' Rule

$$P(R|D) = P(D|R) \times P(R)/P(D)$$

$$P(NR|D) = P(D|NR) \times P(NR)/P(D)$$

A document D is classified as relevant if P(R|D) > P(NR|D). Discarding the constant P(D), this is equivalent to saying that a document is relevant if:

$$P(D|R) \times P(R) > P(D|NR) \times P(NR)$$

The likelihood ratio P(D|R)/P(D|NR) is used as a score to determine the likelihood of the document with representation D belonging to the relevant set.

4. Semantic Model

Semantic approaches include different levels of analysis, such as morphological, syntactic, and semantic analysis, to retrieve documents more effectively. In morphological analysis, roots and affixes are analyzed to determine the parts of speech (nouns, verbs, adjectives, and so on) of the words. The development of a sophisticated semantic system requires complex knowledge bases of semantic information as well as retrieval heuristics. These systems often require techniques from artificial intelligence and expert systems. Knowledge bases like Cyc15 and WordNet16 have been developed for use in knowledge-based IR systems based on semantic models.

QUERIES IN IR SYSTEMS

The queries formulated by users are compared to the set of index keywords. Most IR systems also allow the use of Boolean and other operators to build a complex query. The query language with these operators enriches the expressiveness of a user's information need.

1. Keyword Queries

Keyword-based queries are the simplest and most commonly used forms of IR queries: the user just enters keyword combinations to retrieve documents. The query keyword terms are implicitly connected by a logical AND operator.

A query such as _database concepts' retrieves documents that contain both the words _database' and _concepts' at the top of the retrieved results.

In addition, most systems also retrieve documents that contain only _database' or only _concepts' in their text. Some systems remove most commonly occurring words (such as

a, the, of, and so on, called stop words) as a preprocessing step before sending the filtered query keywords to the IR engine.

2. Boolean Queries

Some IR systems allow using the AND, OR, NOT, ( ), + , and – Boolean operators in combinations of keyword formulations. AND requires that both terms be found. OR lets either term be found.

NOT means any record containing the second term will be excluded. ‗( )' means the Boolean operators can be nested using parentheses. ‗+' is equivalent to AND, requiring the term; the '+' should be placed directly in front of the search term.'–' is equivalent to AND NOT and means to exclude the term; the '–' should be placed directly in front of the search term not wanted.

Complex Boolean queries can be built out of these operators and their combinations, and they are evaluated according to the classical rules of Boolean algebra.

3. Phrase Queries

When documents are represented using an inverted keyword index for searching, the relative order of the terms in the document is lost. In order to perform exact phrase retrieval, these phrases should be encoded in the inverted index or implemented differently (with relative positions of word occurrences in documents).

A phrase query consists of a sequence of words that makes up a phrase. The phrase is generally enclosed within double quotes. Each retrieved document must contain at least one instance of the exact phrase. Phrase searching is a more restricted and specific version of proximity searching.

4. Proximity Queries

Proximity search refers to a search that accounts for how close within a record multiple terms should be to each other. The most commonly used proximity search option is a phrase search that requires terms to be in the exact order.

5. Wildcard Queries

Wildcard searching is generally meant to support regular expressions and pattern matching-based searching in text. In IR systems, certain kinds of wildcard search support may be implemented—usually words with any trailing characters (for example, ‗data*' would retrieve data, database, datapoint, dataset, and so on).

6. Natural Language Queries

There are a few natural language search engines that aim to understand the structure and meaning of queries written in natural language text, generally as a question or narrative. This is an active area of research that employs techniques like shallow semantic parsing of text, or query reformulations based on natural language understanding.

The system tries to formulate answers for such queries from retrieved results. Some search systems are starting to provide natural language interfaces to provide answers to specific types of questions, such as definition and factoid questions, which ask for definitions of technical terms or common facts that can be retrieved from specialized databases.