Databases and Information Systems CS303

Information Retrieval
10-11-2023

Information Retrieval

- Information retrieval generally refers to the querying of unstructured textual data.

- In its simplest form, an information-retrieval system locates and returns all documents that contain all the keywords in the query

- More-sophisticated systems estimate relevance of documents to a query so that the documents can be shown in order of estimated relevance
    - Use information about term occurrences, as well as hyperlink information, to estimate relevance.

Relevance Ranking using Terms

- The set of all documents that satisfy a query expression may be very large
- Irrelevant documents may be retrieved as a result
- Information-retrieval systems estimate relevance of documents to a query, and return only highly ranked documents as answers

Relevance Ranking using Terms

- Ranking using TF-IDF
  - Given a particular term t, how relevant is a particular document d to the term ?
  - Use the the number of occurrences of the term in the document as a measure of its relevance
    - Assuming that relevant terms are likely to be mentioned many times in a document
  - Number of occurrences depends on the length of the document
    - Document containing 10 occurrences of a term may not be 10 times as relevant as a document containing one occurrence.

Relevance Ranking using Terms

- Ranking using TF-IDF
  - Given a particular term t, how relevant is a particular document d to the term ?
  - Use the the number of occurrences of the term in the document as a measure of its relevance
    - Assuming that relevant terms are likely to be mentioned many times in a document
  - Number of occurrences depends on the length of the document
    - Document containing 10 occurrences of a term may not be 10 times as relevant as a document containing one occurrence.

$$TF(d,t) = \log\left(1 + \frac{n(d,t)}{n(d)}\right)$$

  - n(d) is the number of words in the document d
  - n(d,t) is the number of occurrences of the term t in the document d

Relevance Ranking using Terms

- Ranking using TF-IDF
  - The measure can be refined:
    - Term Frequency TF(d,t)
    - Inverse Document Frequency IDF(t)
      - n(t) is the number of documents where t occurs

$$IDF(t) = \frac{1}{n(t)}$$

$$r(d, Q) = \sum_{t \in Q} TF(d, t) * IDF(t)$$

Relevance Ranking using Terms

- Ranking using TF-IDF
  - Stop Words : a , is, the ...
    - Ignored from the query
- If query contains multiple terms :
  - Proximity of the terms in the document is considered
  - The formula for r (d, Q) can be modified to take proximity of the terms into account
- Information-retrieval system returns documents in descending order of their relevance to Q.
  - Since there may be a very large number of documents that are relevant, systems typically return only the first few documents with the highest degree of estimated relevance
  - Then allow users to interactively request further documents

Relevance Ranking using Terms

- Similarity Based Retrieval
  - User can give the system document A, and ask the system to retrieve documents that are similar to A

  - If the set of documents similar to a query document A is large, the system may present a few of the similar documents
    - Allow the user to choose the most relevant few
    - Start a new search based on similarity to A and to the chosen documents.
    - This approach is called relevance feedback

Relevance using Hyperlinks

- Relevance ranking of a page is influenced greatly by hyperlinks that point to the page

- Popularity Ranking
  - Find pages that are popular, and rank them higher than other pages that contain the specified keywords
  - Use hyperlinks to a page as a measure of its popularity

Relevance using Hyperlinks

- To estimate the popularity of a page, use the number of pages that link to the page as a measure of its popularity.
  - But many sites have a number of useful pages but external links often point only to the root page of the site.
  - Other pages would then be wrongly inferred to be not very popular
  - Use websites rather than pages (but what is a website?)

- Allow transfer of prestige from popular pages to pages to which they link
  - In contrast to the one-person one-vote principles of democracy
  - A link from a popular page x to a page y is treated as conferring more prestige to page y than a link from a not-so-popular page z

Relevance using Hyperlinks

- Page Rank :
  - Google introduced PageRank, which is a measure of popularity of a page based on the popularity of pages that link to the page
  - Uses a random walk model.
    - Suppose a person browsing the Web performs a random walk (traversal) on Web pages as follows:
      - the first step starts at a random Web page
      - in each step, the random walker does one of the following
        - With a probability p the walker jumps to a randomly chosen Web page
        - with a probability of 1 − p the walker randomly chooses one of the outlinks from the current Web page and follows the link
  - PageRank of a page is the probability that the random walker is visiting the page at any given point in time
  - Pages that are pointed to from many Web pages are more likely to be visited
  - Pages pointed to by Web pages with a high PageRank will also have a higher probability of being visited

Relevance using Hyperlinks

- Other Measures on popularity
  - PageRank algorithm does not take query keywords into account
    - Example :
      - The page google.com is likely to have a very high PageRank because many sites contain a link to it.
      - Suppose it contains a word mentioned in passing, such as “Dharwad” , search on the keyword “Dharwad” would then return google.com as the highest-ranked answer, ahead of a more relevant answers
    - Solution is to use the anchor text to keep track
    - Another solution is to first find pages containing key words then compute popularity measure among them

Relevance using Hyperlinks

- Search Engine Spamming:
  - Creating Web pages, or sets of Web pages, designed to get a high relevance rank for some queries, even though the sites are not actually popular sites.
  - Using sites instead of pages as the unit of ranking have been proposed to avoid some spamming
    - But not fully effective
    - Ever-growing battle

Synonyms and Homonyms

- Synonyms are different words with same meaning
  - Each word can have a set of synonyms defined
  - Occurrence of a word can be replaced by the “or” of all its synonyms (including the word)

- Homonyms are same words with different meaning
  - Table can mean coffee table or table in a database
  - System should understand the concept each word in a document and what concepts a user is looking for
  - Has to analyze each document to disambiguate each word in the document, and replace it with the concept that it represents
  - Disambiguation is usually done by looking at other surrounding words in the document.
    - Harder for user queries, since queries contain very few words

Ontologies

- Hierarchical structures that reflect relationships between concepts
  - Example : Lion is-a mammal Mammal is-a animal
    - Part-of
- WorkNetSystem and Cyc project aims to create ontology of concepts
- The largest effort in using ontologies for concept-based queries is the Semantic Web
  - Led by the WWW Consortium
  - Consists of a collection of tools, standards, and languages that permit data on the Web to be connected based on their semantics, or meaning.
  - Decentralized and distributed
  - Can integrate multiple, distributed ontologies.

Indexing of Documents

- Important for efficient processing of queries in an information-retrieval system
- Inverted Index that maps each keyword K_i to a list S_i of identifiers of the documents that contain K_i (ordered by popularity)
  - B^+ tree can be used for this
  - AND and OR becomes intersection and union of lists respectively

Measuring Retrieval Effectiveness

- Each keyword may be contained in a large number of documents
  - compact representation is needed to keep space usage of the index low
- Index is sometimes stored such that the retrieval is approximate
  - Some relevant documents may not be retrieved (false drops)
  - Good index will never have false drops
- Precision measures what percentage of the retrieved documents are actually relevant to the query.
- Recall measures what percentage of the documents relevant to the query were retrieved
- Ideally both should be 100 percent

Crawling and Indexing the Web

- Web crawlers are programs that locate and gather information on the Web
- Recursively follow hyperlinks present in known documents to find other documents
  - Start from an initial set of URLs, which may be created manually
  - Web crawler then locates all URL links in these pages, and adds them to the set of URLs to be crawled, if they have not already been fetched, or added to the to-be-crawled set
- Not possible to crawl the whole Web in a short period of time
  - All search engines cover only some portions of the Web
  - Crawlers may take weeks or months to perform a single crawl of all the pages they cover
  - A database stores a set of links (or sites) to be crawled
    - it assigns links from this set to each crawler process.
    - New links found during a crawl are added to the database,
  - Pages have to be refetched
    - Since webpages are updated frequently
- Pages fetched during a crawl are handed over to Ranking system

Crawling and Indexing the Web

- Many sites containing large collections of data may not make all the data available as hyperlinked pages
  - They provide search interfaces, where users can enter terms, or select menu options, and get results
    - Example: Searching Trains in IRCTC

- The information in such sites is called deep Web information

- Deep Web crawlers extract such information by guessing what terms would make sense to enter
  - The pages extracted by a deep Web crawl may be indexed just like regular Web pages.

Information Retrieval : Beyond Ranking of Pages

- Query results
  - Can be images, videos ..
  - Query results are displayed graphically, so that user does not need to click URL
  - Results carry diverse information on homonyms

- Information Extraction
  - Convert information from textual form to a more structured form
  - Several systems have been built for information extraction for specialized applications.
  - They use linguistic techniques, page structure, and user-defined rules for specific domains.
  - On the Web scale manual creation of such patterns is not feasible
    - Machine-learning techniques, which can learn such patterns given a set of training examples (ChatGPT)

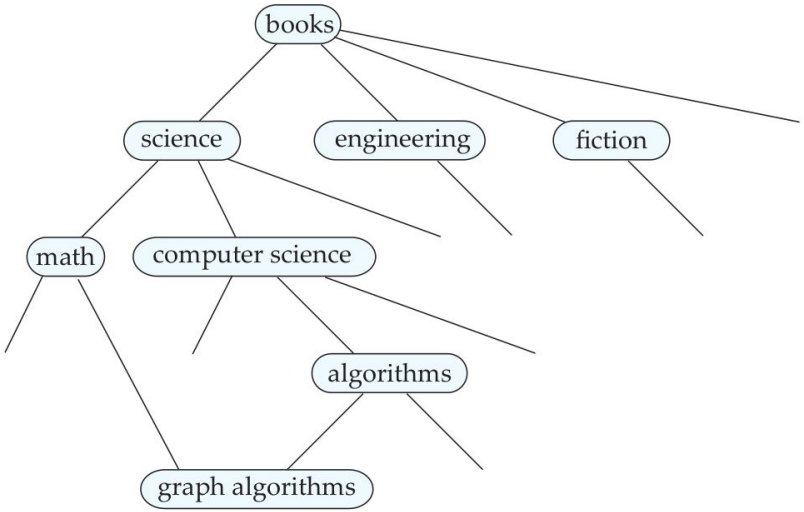
Information Retrieval : Beyond Ranking of Pages

- Question Answering
  - What is the population of India?
    - Google gives Direct Answer
  - Current-generation question answering systems are limited in power
    - Do not really understand either the question or the documents used to answer the question.
  - However, they are useful for a number of simple question answering tasks

- Querying Structured Data
  - User can specify partially related data in the query
    - Can happen if the user does not know the schema
  - Ways to rank data looking and how they are connected in the database
    - Like Foreign key relationship
    - Paths in a graph

Directories and Categories

- Directory is a classification DAG structure
  - Each leaf of the directory stores links to documents on the topic represented by the leaf
  - Internal nodes contain links to documents that cannot be classified under any of the child nodes.
    Example : Netflix, Marketing websites



- Concepts are classified into Categories to which they belong

Reference:

Database System Concepts by Silberschatz, Korth and Sudarshan
(6th edition)
Chapter 21

Current research Fields

- Non-standard query evaluation
- Probabilistic Databases
- Graph Query Optimization
- .....