SOUTHERN LEYTE STATE UNIVERSITY BONTOC

San Ramon, Bontoc, Southern Leyte, 6604 Philippines
Telefax No.: (053) 382-3121; E-mail: slsubc@yahoo.com
Website: www.slsuonline.edu.ph

ADVANCED DATABASE

LEARNING GUIDE 2020

REXAL S. TOLEDO

SOUTHERN LEYTE
STATE UNIVERSITY-BONTOC

This learning module is developed for instructional purposes only. Any form of reproduction or distribution is strictly prohibited.

Southern Leyte State University

Vision

A high quality corporate University of Science, Technology and Innovation.

Mission

SLSU will:

a) *Develop Science, Technology, and Innovation leaders and professionals;*
b) *Produce high-impact technologies from research and innovations;*
c) *Contribute to sustainable development through responsive community engagement programs;*
d) *Generate revenues to be self-sufficient and financially-viable.*

Quality Policy

We at Southern Leyte State University commit enthusiastically to satisfy our stakeholders' needs and expectations by adhering to good governance, relevance and innovations of our instruction, research and development, extension and other support services and to continually improve the effectiveness of our Quality Management System in compliance to ethical standards and applicable statutory, regulatory, industry and stakeholders' requirements.

The management commits to establish, maintain and monitor our quality management system and ensure that adequate resources are available.

COURSE OVERVIEW

Course No.	IT301/IT301L
Course Code	
Descriptive Title	Advanced Database Systems
Credit Units	2 units (Lecture) / 1 unit (Laboratory)
School Year/Term	2020-2021 / 1st semester
Mode of Delivery	
Name of Instructor	Rexal S. Toledo
Course Description	This course covers modern database and information system as well as research issues in the field. It will cover selected topics on NoSQL, object-oriented, active, deductive, spatial, temporal and multimedia databases. The course includes advanced issues of object-oriented database, XML, advanced client server architecture, Information Retrieval and Web Search and distributed database techniques.
Course Outcomes	1. Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions. 2. Design, implement, and evaluate a computing-based solution to meet a given set of computing requirements in the context of the program's discipline. 3. Apply software development fundamentals to produce computing-based solutions. 4. Communicate effectively in a variety of professional contexts.

TABLE OF CONTENTS

9	Retrieval Models and Types of Queries in IR Systems
10	Text Preprocessing and Inverted Indexing
11	Evaluation Measures of Search Relevance, Web Search and Analysis

MODULE 3

Introduction to Information Retrieval and Web Search

LESSON

PRE-TEST

MODULE 3:

1. What is structured data and what is unstructured data? Give an example of each from your experience.
2. Give a general definition of information retrieval (IR). What does information retrieval involve when we consider information on the Web?
3. Discuss the types of data and the types of users in today's information retrieval systems.
4. What is meant by navigational, informational, and transformational search?
5. What are the two main modes of interaction with an IR system? Describe and provide examples.
6. Explain the main differences between the database and IR systems.
7. Describe the main components of the IR system.
8. What are digital libraries? What types of data are typically found in them?
9. Name some digital libraries that you have accessed. What do they contain and how far back does the data go?
10. Give a brief history of IR and mention the landmark developments in this field.
11. What is the Boolean model of IR? What are its limitations?
12. What is the vector space model of IR? How does a vector get constructed to represent a document?
13. Define the TF-IDF scheme of determining the weight of a keyword in a document. Why is it necessary to include IDF in the weight of a term?
14. What are probabilistic and semantic models of IR?
15. Define recall and precision in IR systems.
16. How is an F-score defined as a metric of information retrieval? In what way does it account for both precision and recall?
17. What are the different types of queries in an IR system? Describe each with an example.
18. What are the approaches to processing phrase and proximity queries?
19. Describe the detailed IR process
20. What is stopword removal and stemming? Why are these processes necessary for better information retrieval?
21. What is a thesaurus? How is it beneficial to IR?
22. What is information extraction? What are the different types of information extraction from structured text?
23. What are vocabularies in IR systems? What role do they play in the indexing of documents?
24. Describe the process of constructing the result of a search request using an inverted index.
25. Define relevance feedback.
26. Describe the three types of Web analyses.
27. What are the three categories of agent-based Web content analyses mentioned in this chapter?
28. What is the database-based approach to analyzing Web content? What are Web query systems?
29. What algorithms are popular in ranking or determining the importance of Web pages? Which algorithm was proposed by the founders of Google?
30. What can you learn from Web usage analysis? What data does it generate?
31. What mining operations are commonly performed on Web usage data? Give an

example of each.

32. What are the applications of Web usage mining?
33. What is search relevance? How is it determined?
34. Define and explain conversational search.
35. Define topic modeling.

LESSON 8

Information Retrieval (IR) Concepts and Trends

LEARNING OUTCOMES

After studying this lesson, you should be able to:

1. Discuss the different kinds and levels of search that IR encompasses.
2. Compare IR and learn database technologies.
3. Present the different modes of user interaction with IR systems
4. Describe the typical IR process with a detailed set of tasks and with a simplified process flow.

Information Retrieval (IR)

- Process of retrieving documents from a collection in response to a query (search request) by a user.
- Deals mainly with unstructured data.
  Example: home buying contract documents

In Historical

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.

In academic programs

The field of IR has long been a part of Library and Information Science programs. 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, e-mails, instant messages, and collections from digital libraries.

Therefore, all loosely represented (unstructured) or semi structured information is also part of the IR discipline.

To see how information retrieval differs from structured data management. Consider a relation (or table) called HOUSES with the attributes:

```
HOUSES (Lot#, Address, Square_footage, Listed_price)
```

This is an example of `structured data`. We can compare this relation with homebuying contract documents, which are examples of unstructured data. These types of documents can vary from city to city, and even county to county.

- **Unstructured information**
- Information that does not have a well-defined formal model and corresponding formal language for representation and reasoning, but rather is based on understanding of natural language.
- Stored in a wide variety of standard formats

IR systems

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.

Free-form search request (sometimes called a keyword search query, or just query)

- IR systems use a user's information need expressed for interpretation by the system.

An IR system can be characterized at different levels by:

1. **Types of Users** - This ability depends on a multitude of factors, such as education, culture, and past exposure to computational environments. 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, understands the scope and the structure of the available repository, and forms relevant queries for the task, or a **layperson** user with a generic information need.
2. **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.

3. **Types of Information Need –** In the context of Web search, users' information needs may be defined as navigational, informational, or transactional.

Navigational search - refers to finding a particular piece of information (such as the University's Web site) that a user needs quickly.
Informational search - to find current information about a topic.
Transactional search - to reach a site where further interaction happens resulting in some transactional event (such as joining a social network, shopping for products, making online reservations, accessing databases, and so on).

IR search engines can be limited in level to more specific collections of documents.

Enterprise search systems offer IR solutions for searching different entities in an enterprise's intranet, which consists of the network of computers within that enterprise. The searchable entities include e-mails, corporate documents, manuals, charts, and presentations, as well as reports related to people, meetings, and projects.

On a smaller scale, there are personal information systems such as those on desktops and laptops, called:

Desktop search engines – for retrieving files, folders, and different kinds of entities stored on the computer.

Databases and IR Systems: A Comparison

- Databases and IR systems are closely related fields.

  Databases - deal with structured information retrieval through well-defined formal languages for representation and manipulation based on the theoretically founded data models and have fixed schemas defined in some data model such as the relational model.

  IR - deals with unstructured search with possibly vague query or search semantics and without a well-defined logical schematic representation. No fixed data model.

A Comparison of Databases and IR Systems

Databases	IR Systems
■ Structured data	■ Unstructured data
■ Schema driven	■ No fixed schema; various data models (e.g., vector space model)
■ Relational (or object, hierarchical, and network) model is predominant	■ Free-form query models
■ Structured query model	■ Rich data operations
■ Rich metadata operations	■ Search request returns list or pointers to documents
■ Query returns data	
■ Results are based on exact matching (always correct)	■ Results are based on approximate matching and measures of effectiveness (may be imprecise and ranked)

A Brief History of IR

- Stone tablets and papyrus scrolls
- Printing press
- Public libraries
- Computers and automated storage systems
    - Inverted file organization based on keywords and their weights as indexing method
- Search engine
- Crawler
- Challenge: provide high quality, pertinent, timely information

Modes of Interactions in IR Systems

Primary modes of interaction

Retrieval - Extract relevant information from document repository.
Browsing - Exploratory activity based on user's assessment of relevance

Web search combines both aspects and is one of the main applications of information retrieval today. The rank of a Web page in a retrieved set is the measure of its relevance to the query that generated the result set.

Generic IR Pipeline

Statistical approach

Documents analyzed and broken down into chunks of text.
Each word or phrase is counted, weighted, and measured for relevance or importance.

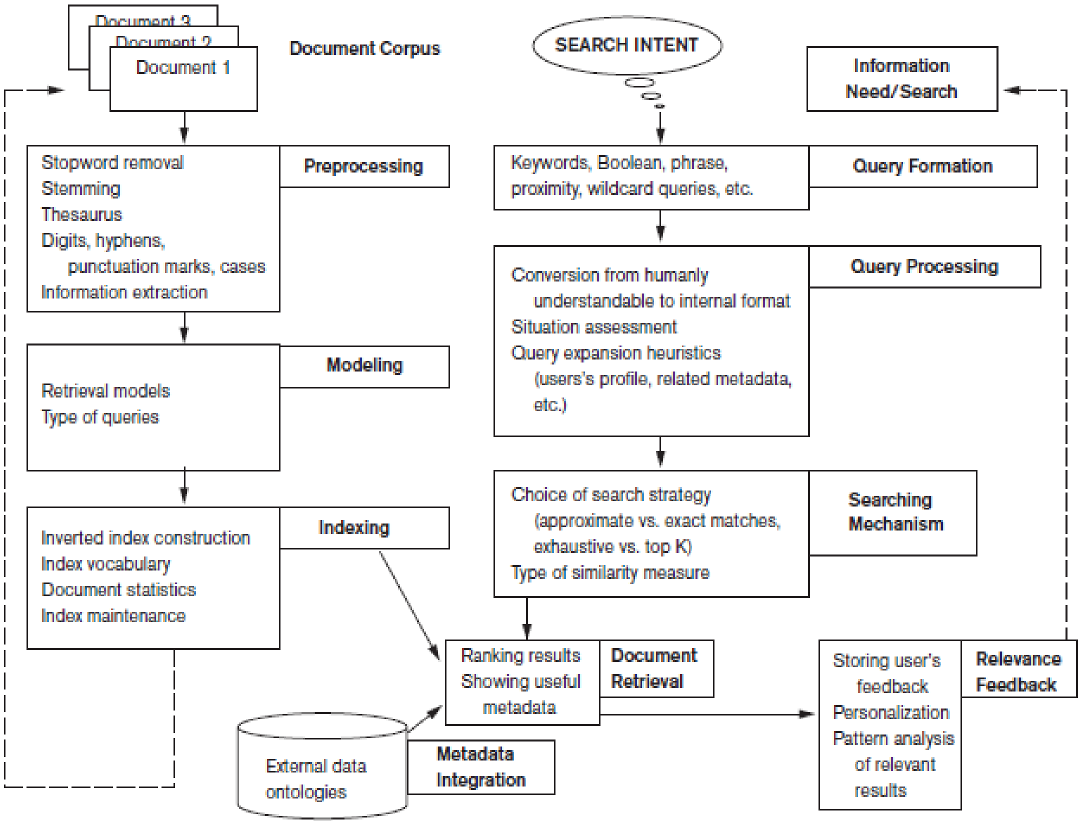
Types of statistical approaches:

- Boolean
- Vector space
- Probabilistic

Semantic approaches

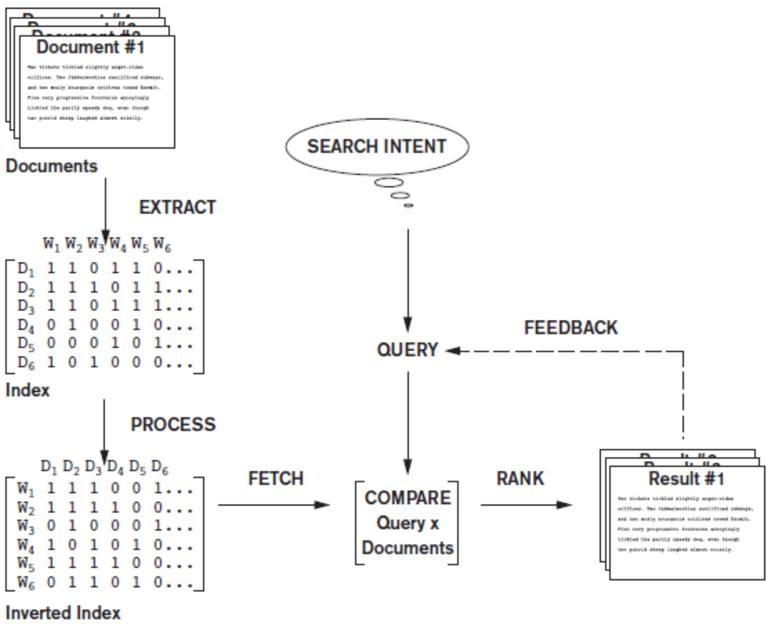
- Use knowledge-based retrieval techniques
- Rely on syntactic, lexical, sentential, discourse-based, and pragmatic levels of knowledge understanding
- Also apply some form of statistical analysis

Generic IR framework



Legend: Dashed lines indicate next iteration

Simplified IR process pipeline



Trends in Information Retrieval

Faceted search

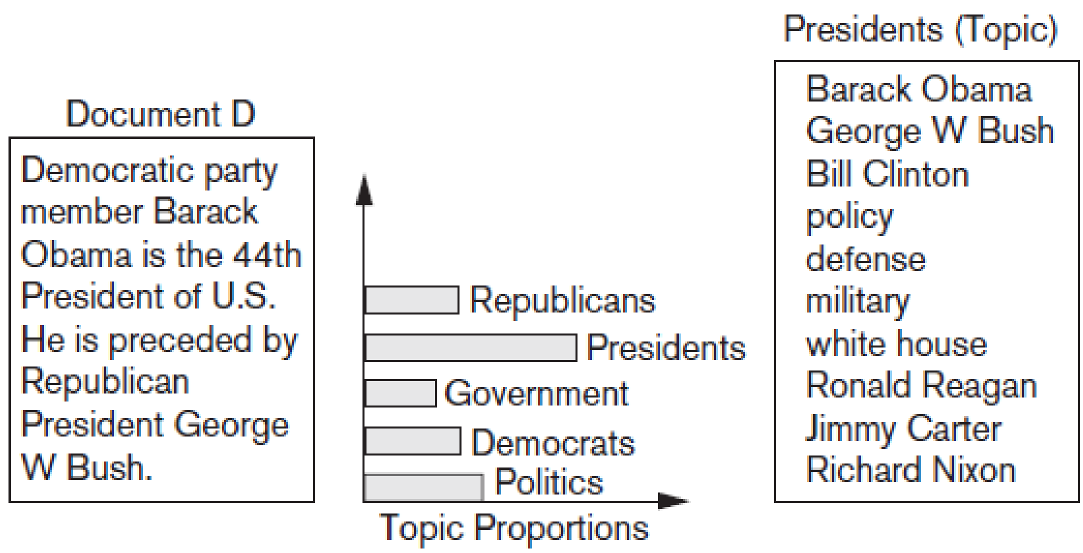
- Classifying content

Social search
   - Collaborative social search

Conversational information access
   - Intelligent agents perform intent extraction to provide information relevant to a conversation

Probabilistic topic modeling
   - Automatically organize large collections of documents into relevant themes

A document D and its topic proportions



Question-answering systems
- Factoid questions
- List questions
- Definition questions
- Opinion questions
- Composed of question analysis, query generation, search, candidate answer generation, and answer scoring

LESSON 9

Retrieval Models and Types of Queries in IR Systems

LEARNING OUTCOMES

After studying this lesson, you should be able to:

1. Describe the important models of IR. These are the three main statistical models—Boolean, vector space, and probabilistic—and the semantic model.
2. Learn the different types of Queries in IR Systems

Retrieval Models

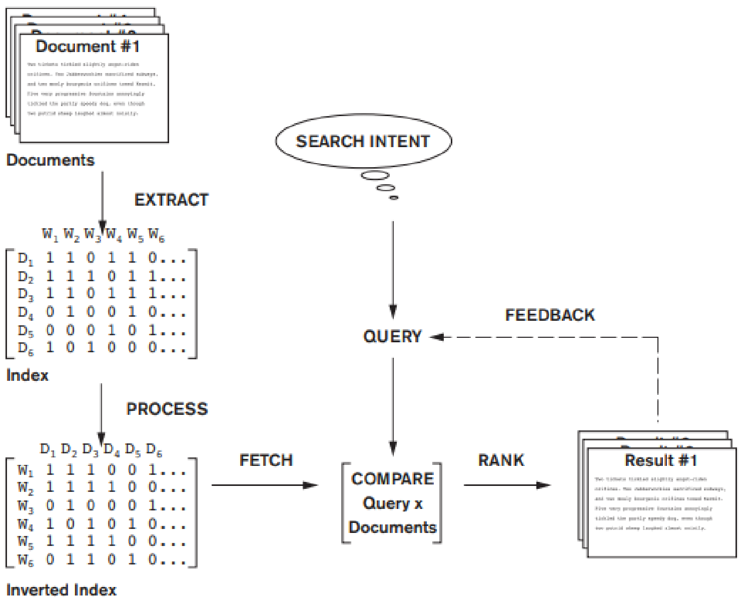


Figure 27.2
Simplified IR process pipeline.

Boolean model

- One of earliest and simplest IR models
- Documents represented as a set of terms
- Queries formulated using AND, OR, and NOT
- Retrieved documents are an exact match
    - No notion of ranking of documents
- Easy to associate metadata information and write queries that match contents of documents

Vector space model

- Weighting, ranking, and determining relevance are possible
- Uses individual terms as dimensions
- Each document represented by an n-dimensional vector of values
- Features
    - Subset of terms in a document set that are deemed most relevant to an IR search for the document set
- Different similarity assessment functions can be used
- Term frequency-inverse document frequency (TF-IDF)
    - Statistical weight measure used to evaluate the importance of a document word in a collection of documents
    - A discriminating term must occur in only a few documents in the general population

Probabilistic model

- Involves ranking documents by their estimated probability of relevance with respect to the query and the document
- IR system must decide whether a document belongs to the relevant set or nonrelevant set for a query
    - Calculate probability that document belongs to the relevant set
- BM25: a popular ranking algorithm

Semantic model

- Morphological analysis
    - Analyze roots and affixes to determine parts of speech of search words
- Syntactic analysis
    - Parse and analyze complete phrases in documents
- Semantic analysis
    - Resolve word ambiguities and generate relevant synonyms based on semantic relationships
- Uses techniques from artificial intelligence and expert systems

Types of Queries in IR Systems

Keyword queries

- Simplest and most commonly used
- Keyword terms implicitly connected by logical AND

Boolean queries

- Allow use of AND, OR, NOT, and other operators
- Exact matches returned
  - No ranking possible

Phrase queries

- Sequence of words that make up a phrase
- Phrase enclosed in double quotes
- Each retrieved document must contain at least one instance of the exact phrase

Proximity queries

- How close within a record multiple search terms are to each other
- Phrase search is most commonly used proximity query
- Specify order of search terms
- NEAR, ADJ (adjacent), or AFTER operators
- Sequence of words with maximum allowed distance between them
- Computationally expensive
  - Suitable for smaller document collections rather than the Web

- **Wildcard queries**
  - Supports regular expressions and pattern-based matching

    Example ‘data*’ would retrieve data, database, dataset, etc.

  - Not generally implemented by Web search engines

- **Natural language queries**
  - Definitions of textual terms or common facts
  - Semantic models can support

LESSON 10

Text Preprocessing and Inverted Indexing

LEARNING OUTCOMES

After studying this lesson, you should be able to:

1. Learn text preprocessing techniques.
2. Learn the construction and use of inverted indexes
3. Understand the various activities like stopword removal, stemming, and the use of thesauruses.

Stopword Removal

Stopword removal must be performed before indexing

Words that are expected to occur in 80% or more of the documents in a collection are typically referred to as **stopwords**, and they are rendered potentially useless. Because of the commonness and function of these words, they do not contribute much to the relevance of a document for a query search.

Examples include words such as the, of, to, a, and, in, said, for, that, was, on, he, is, with, at, by, and it.

Removal of stopwords results in elimination of possible spurious indexes, thereby reducing the size of an index structure by about 40% or more.

For example, a search for the phrase 'To be or not to be', where removal of stopwords makes the query inappropriate, as all the words in the phrase are stopwords.

Many search engines do not employ query stopword removal for this reason.

Stemming

A stem of a word is defined as the word obtained after trimming the suffix and prefix of an original word.

Example, 'comput' is the stem word for computer, computing, computable, and computation.

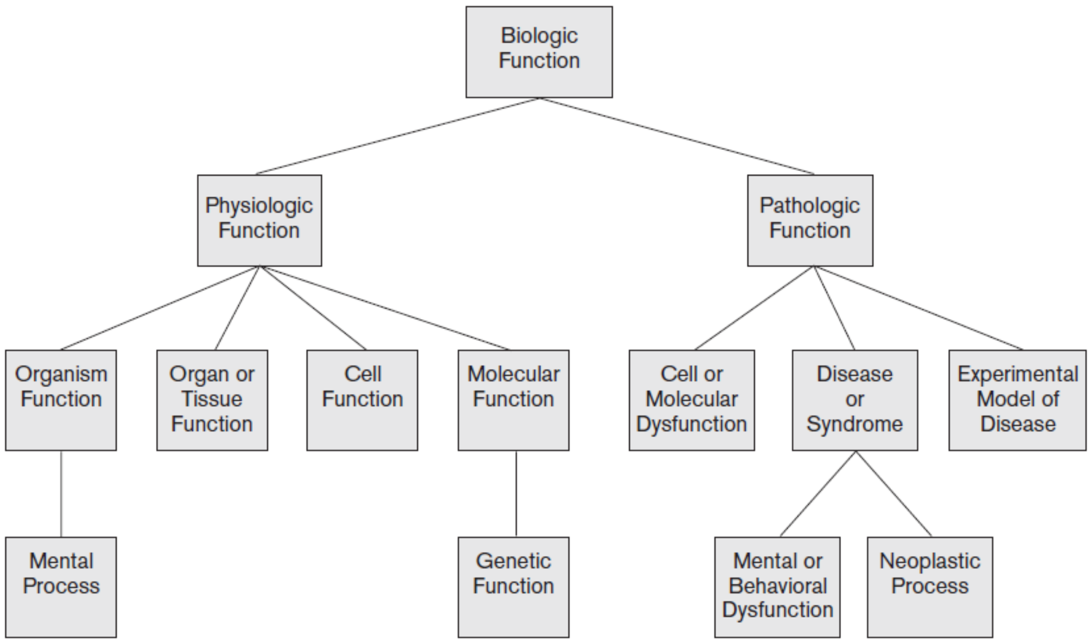
Stemming algorithm – can be applied to reduce any word to its stem. Martin Porter's stemming algorithm

Utilizing a thesaurus

A precompiled list of important concepts and the main word that describes each concept for a particular domain of knowledge.
Usage of a thesaurus, also known as a collection of synonyms

UMLS – a large biomedical thesaurus of millions of concepts (called the metathesaurus) and a semantic network of Meta concepts and relationships that organize the metathesaurus.

A portion of the UMLS Semantic Network: "Biologic Function" Hierarchy



Other preprocessing steps

Digits, dates, phone numbers, e-mail addresses, URLs, and other standard types of text may or may not be removed during preprocessing. Web search engines, however, index them in order to use this type of information in the document metadata to improve precision and recall.

Hyphens and punctuation marks may be handled in different ways. Either the entire phrase with the hyphens/punctuation marks may be used, or they may be eliminated. In some systems, the character representing the hyphen/punctuation mark may be removed, or may be replaced with a space.

Most information retrieval systems perform case-insensitive search, converting all the letters of the text to uppercase or lowercase.

Information extraction tasks

- Identifying noun phrases, facts, events, people, places, and relationships.

Inverted Indexing

The simplest way to search for occurrences of query terms in text collections can be performed by sequentially scanning the text. This kind of online searching is only appropriate when text collections are small. Most information retrieval systems process the text collections to create indexes and operate upon the inverted index data structure.

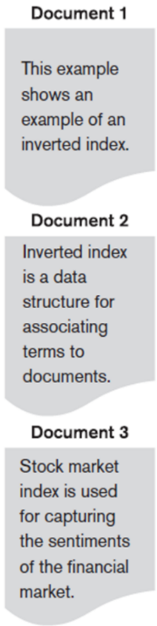
An inverted index structure comprises vocabulary and document information.

Vocabulary

Set of distinct query terms in the document set. Each term in a vocabulary set has an associated collection of information about the documents that contain the term, such as document id, occurrence count, and offsets within the document where the term occurs.

An inverted index of a document collection is a data structure that attaches distinct terms with a list of all documents that contains the term. The process of inverted index construction involves the extraction and processing steps.

Example of an inverted index.



ID	Term	Document: position
1.	example	1:2, 1:5
2.	inverted	1:8, 2:1
3.	index	1:9, 2:2, 3:3
4.	market	3:2, 3:13

Searching for relevant documents from the inverted index, given a set of query terms, is generally a three-step process.

1. **Vocabulary search**
2. **Document information retrieval**
3. **Manipulation of retrieved information**

Introduction to Lucene

- Lucene: open source indexing/search engine
- Indexing is primary focus
- Document composed of set of fields
    - Chunks of untokenized text
    - Series of processed lexical units called token streams
      Created by tokenization and filtering algorithms.
- Highly-configurable search API
- Ease of indexing large, unstructured document collections

LESSON 11

Evaluation Measures of Search Relevance, Web Search and Analysis

LEARNING OUTCOMES

After studying this lesson, you should be able to:

1. Learn the various evaluation metrics, such as recall precision and F-score, to measure the goodness of the results of IR queries.
2. Describe the various evaluation metrics for IR systems performance.
3. Learn Web analysis and its relationship to information retrieval

Evaluation Measures of Search Relevance

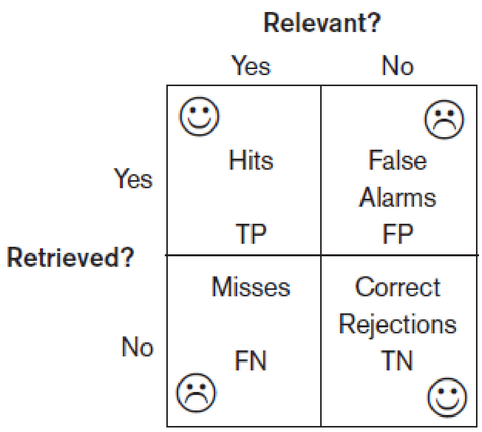
- Topical relevance
    - Measures result topic match to query topic
- User relevance
    - Describes 'goodness' of retrieved result with regard to user's information need
- Web information retrieval
    - No binary classification made for relevance or non-relevance
    - Ranking of documents
- Recall
    - Number of relevant documents retrieved by a search divided by the total number of actually relevant documents existing in the database
- Precision
    - Number of relevant documents retrieved by a search divided by total number of documents retrieved by that search

Retrieved Versus Relevant Search Results

A pictorial representation of the terms retrieved versus relevant and shows how search results relate to four different sets of documents.

Notation:

TP: true positive
FP: false positive
FN: false negative
TN: true negative



Recall can be increased by presenting more results to the user. May decrease the precision.

Doc. No.	Rank Position *i*	Relevant	Precision(*i*)	Recall(*i*)
10	1	Yes	1/1 = 100%	1/10 = 10%
2	2	Yes	2/2 = 100%	2/10 = 20%
3	3	Yes	3/3 = 100%	3/10 = 30%
5	4	No	3/4 = 75%	3/10 = 30%
17	5	No	3/5 = 60%	3/10 = 30%
34	6	No	3/6 = 50%	3/10 = 30%
215	7	Yes	4/7 = 57.1%	4/10 = 40%
33	8	Yes	5/8 = 62.5%	5/10 = 50%
45	9	No	5/9 = 55.5%	5/10 = 50%
16	10	Yes	6/10 = 60%	6/10 = 60%

Average precision

Computed based on the precision at each relevant document in the ranking

Recall/precision curve

Based on the recall and precision values at each rank position

x-axis is recall and y-axis is precision

F-score

Harmonic mean of the precision (p) and recall (r) values

Web Search and Analysis

The emergence of the Web has brought millions of users to search for information, which is stored in a very large number of active sites. To make this information accessible, search engines such as Google, bing and Yahoo! must **crawl** and index these sites and document collections in their index databases.

There are other types of search engines besides the ones that regularly **crawl** the Web and create automatic indexes: these are **human-powered**, **vertical search engines** or **metasearch engines**.These search engines are developed with the help of computer-assisted systems to aid the curators with the process of assigning indexes.

Vertical search engines

- Customized topic-specific search engines that crawl and index a specific collection of documents on the Web and provide search results from that specific collection

Metasearch engines are built on top of search engines: they query different search engines simultaneously and aggregate and provide search results from these sources.

Digital libraries
- Collections of electronic resources and services for the delivery of materials in a variety of formats

Web analysis
- Applies data analysis techniques to discover and analyze useful information from the Web

  Goals of Web analysis:

  - Finding relevant information
  - Personalization of the information
  - Finding information of social value

  Categories of Web analysis

  - Web structure analysis
  - Web content analysis
  - Web usage analysis

Web structure analysis
- Hyperlink
- Destination page
- Anchor text
- Hub
- Authority

PageRank ranking algorithm

Used by Google

Analyzes forward links and backlinks
- Highly linked pages are more important

Web content analysis
- Structured data extraction
  - Wrapper

- Web information integration
  - Web query interface integration
  - Schema matching
  - Ontology-based information integration
- Building concept hierarchies
- Segmenting web pages and detecting noise

Approaches to Web content analysis

The two main approaches to Web content analysis are (1) agent based (IR view) and (2) database based (DB view).

Agent-based
- Intelligent Web agents
- Personalized Web agents
- Information filtering/categorization

Database-based
- Attempts to organize a Web site as a database
- Object Exchange Model
- Multilevel database
- Web query system

Web Usage Analysis

- Application of data analysis techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications.

Web usage analysis attempts to discover usage patterns from Web data

Preprocessing
Usage, content, structure
Pattern discovery
Statistical analysis, association rules, clustering, classification, sequential patterns, dependency modeling
Pattern analysis
Filter out patterns not of interest

Practical applications of Web analysis:

1. **Web analytics**
   Understand and optimize the performance of Web usage
2. **Web spamming**
   Deliberate activity to promote a page by manipulating search engine results
3. **Web security**
   Allow design of more robust Web sites
4. **Web crawlers**

POST-TEST

Module 3:

1. What is structured data and what is unstructured data? Give an example of each from your experience.
2. Give a general definition of information retrieval (IR). What does information retrieval involve when we consider information on the Web?
3. Discuss the types of data and the types of users in today's information retrieval systems.
4. What is meant by navigational, informational, and transformational search?
5. What are the two main modes of interaction with an IR system? Describe and provide examples.
6. Explain the main differences between the database and IR systems.
7. Describe the main components of the IR system.
8. What are digital libraries? What types of data are typically found in them?
9. Name some digital libraries that you have accessed. What do they contain and how far back does the data go?
10. Give a brief history of IR and mention the landmark developments in this field.
11. What is the Boolean model of IR? What are its limitations?
12. What is the vector space model of IR? How does a vector get constructed to represent a document?
13. Define the TF-IDF scheme of determining the weight of a keyword in a document. Why is it necessary to include IDF in the weight of a term?
14. What are probabilistic and semantic models of IR?
15. Define recall and precision in IR systems.
16. How is an F-score defined as a metric of information retrieval? In what way does it account for both precision and recall?
17. What are the different types of queries in an IR system? Describe each with an example.
18. What are the approaches to processing phrase and proximity queries?
19. Describe the detailed IR process
20. What is stopword removal and stemming? Why are these processes necessary for better information retrieval?
21. What is a thesaurus? How is it beneficial to IR?
22. What is information extraction? What are the different types of information extraction from structured text?
23. What are vocabularies in IR systems? What role do they play in the indexing of documents?
24. Describe the process of constructing the result of a search request using an inverted index.
25. Define relevance feedback.
26. Describe the three types of Web analyses.
27. What are the three categories of agent-based Web content analyses mentioned in this chapter?
28. What is the database-based approach to analyzing Web content? What are Web query systems?
29. What algorithms are popular in ranking or determining the importance of Web pages? Which algorithm was proposed by the founders of Google?
30. What can you learn from Web usage analysis? What data does it generate?

31. What mining operations are commonly performed on Web usage data? Give an example of each.
32. What are the applications of Web usage mining?
33. What is search relevance? How is it determined?
34. Define and explain conversational search.
35. Define topic modeling.

REFERE NCES

1. Database Systems: Design, Implementation, and Management
2. 4th Edition, Peter Rob & Carlos Coronel
3. Database Systems: Design, Implementation, and Management, Fifth Edition, Rob and Coronel
4. Database System Concepts Seventh Edition, Avi Silberschatz ET. Al.
5. Fundamentals of Database Systems Seventh Edition, by Ramez Elmasri and Shamkant B. Navathe.

Internet:

6. https://www.geeksforgeeks.org/