MODULE II

CHAPTER 2

IR Models

Syllabus

Modeling : Taxonomy of Information Retrieval Models, Retrieval: Formal Characteristics of IR models, Classic Information Retrieval, Alternative Set Theoretic models, Probabilistic Models, Structured text retrieval Models, models for Browsing;

Self-learning Topics : Terrier

2.1 INTRODUCTION

GQ. What do you mean information retrieval models?

- Information retrieval's (IR) objective is to give users the documents they need to satisfy their informational needs.
- We use the term "document" in a broad sense to refer to both textual and non-textual information, including multimedia items.
- Index terms are typically used by traditional information retrieval systems to index and retrieve documents. A keyword (or combination of related terms) with a distinct meaning is known as an index term (usually a noun)
- The semantics of the documents and of the user information need can be naturally expressed through sets of index terms.
- This method is simple to implement but retrieved documents are often irrelevant because a lot of semantics are lost when we replace its text with a set word.
- The main problem in information retrieval is judging relevant and non-relevant documents.

- Information retrieval systems use ranking algorithms to determine which documents are relevant and which are not.
- The predictions of what is relevant and what is not are based on the accepted IR mode

2.2 A TAXONOMY OF INFORMATION RETRIEVAL MODELS

GQ. What are the three classic models in information retrieval system?

GQ. Explain the taxonomy of information retrieval with a classification diagram.

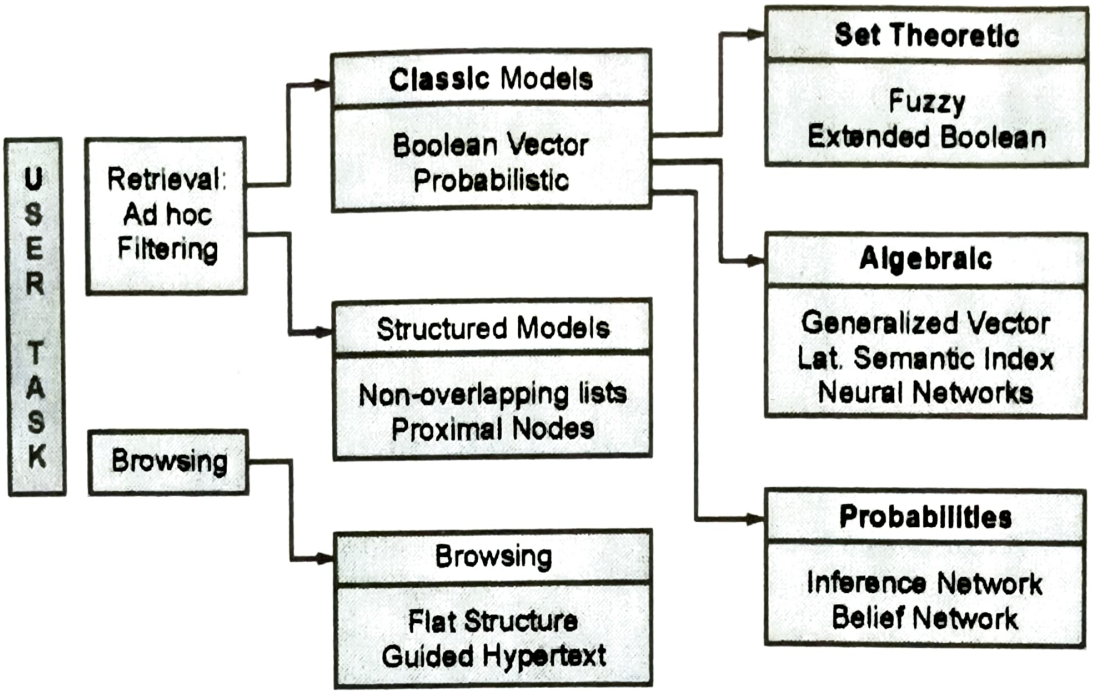
The three classic models in information retrieval :

(1) **Boolean :** Documents and queries are represented as set s of index terms in the Boolean model. As a result, we describe the model as set theoretic

(2) **Vector :** Documents and queries are represented as vectors in the vector model in a t-dimensional space. As a result, we define the model as algebraic.

(3) **Probabilistic :** The framework for modeling document and query representations in the probabilistic model is based on probability theory. As a result, we refer to the model as probabilistic, as its name suggests.

- For each sort of traditional model (i.e., set-theoretic, algebraic, and probabilistic), alternative modeling paradigms have been put out over the years.
- We make a distinction between the fuzzy and extended Boolean models when it comes to alternative set-theoretic models.
- We differentiate the generalized vector, latent semantic indexing, and neural network models as alternative algebraic models.
- We distinguish between the inference network and belief network models when referring to alternative probabilistic models. A taxonomy of these information retrieval models is shown in Fig. 2.2.1.
- We distinguish between the non-overlapping lists model and the proximal nodes model for structured text retrieval.



(181)**Fig. 2.2.1 : A taxonomy of Information Retrieval Models**

- As discussed in chapter 1, the logical view of the documents (whole text, collection of index words, etc.), the IR model (Boolean, vector, probabilistic, etc.), and the user tasks (retrieval, browsing) are orthogonal features of a retrieval system.
- Thus, even though some models are better suited for one user task than another, the same IR model can be utilized with various document logical views to carry out various user tasks as shown in Fig. 2.2.2.

Logical view of documents

USER TASK	Index terms	Full text	Full Text + Structure
Retrieval	Classical set theoretic algebraic probabilistic	Classical set theoretic algebraic probabilistic	Structured
Browsing	Flat	Flat hypertext	Structure guided hypertext

Fig. 2.2.2 : Retrieval models most frequently associated with distinct combinations of a document logical view and a user task

2.3 RETRIEVAL : AD HOC AND FILTERING

GQ. Define Ad hoc retrieval and Filtering.

- **Ad hoc retrieval :** When new queries are entered into a traditional information retrieval system, the collection of documents remains largely static.
- **Filtering :** Queries are relatively static as new documents are added to the system (and leave). In filtering user profile is created according to the user's preferences.
- The incoming documents are then compared to this profile in an effort to identify any that might be of interest to this specific user.
- This method can be used, for instance, to choose a news article from among the many that are broadcast each day.
- Ranking of the filtered documents is not provided.
- A set of keywords are used to create user profile.

2.4 A FORMAL CHARACTERIZATION OF IR MODELS

GQ. Illustrate formal characterization of IR Model.

The formal characterization of IR Model is as follows :

Definition : An information retrieval model is a quadruple [D, Q, F, R $\{q_i, d_j\}$] where

- D is a set composed of logical views (or representations) for the documents in the collection.
- Q is a set composed of logical views (or representations) for the user information needs. Such representations are called queries,
- F is a framework for modeling document representations, queries, and their relationships.
- R $\{q_i, d_j\}$ is a ranking function which associates a real number with a query $q_i \in Q$ and a document representation $d_j \in D$. Such ranking defines an ordering among the documents with regard to the query Q_i.

2.5 CLASSIC INFORMATION RETRIEVAL

GQ. Explain Classic Information Retrieval.

In this section, we briefly present the three classic models in information retrieval namely, the Boolean, the vector, and the probabilistic models.

Basic Concepts

- Each document is described by a group of representative keywords called index terms.
- An index term is only a word from the document whose semantics makes it easier to recall its core ideas.
- Index terms are used to index and summarise the contents of the document. Nouns are preferred as index terms.
- When used to describe the contents of a document, various index terms have differing degrees of importance.
- Each index term in a document is given a numerical weight in order to represent this effect.
- Let k_i be an index term, d_j be a document, and $W_{ij} > 0$ be a *weight* associated with the pair (k_i, d_j). This weight quantifies the importance of the index term for describing the document semantic contents.

Definition : Let t be the number of index terms in the system and k_i be a generic index term. $K = \{ k_1, \ldots k_t \}$ is the set of all index terms. A weight $W_{i,j} > 0$ is associated with each index term k_i of a document d_j. For an index term which does *not* appear in the document text, $W_{i,j} = 0$. With the document d_j is associated an index term vector d_j represented by $d_j = (w_{1,j}, w_{2,j}, \ldots w_{t,j})$ Further, let g_i be a function that returns the weight associated with the index term k_i in any t-dimensional vector (i.e., $g_i(d_j) = W_{i,j}$).

2.5.1 Boolean Model

GQ. What is the basis for the Boolean model?

GQ. What are the advantages and disadvantages of the Boolean model?

- The Boolean model is a simple retrieval model based on set theory and

Boolean algebra. Retrieval is based on whether or not the documents contain the query terms.

- The Boolean model is interested only in the presence or absence of a term in the document.
- In the exact match, a query specifies precise criteria. Each document either matches or fails to match the query. The results retrieved in the exact match are a set of documents (without ranking).
- In the best match, a query describes good or best matching documents. In this case, the result is a ranked list of documents. The Boolean model here I'm going to deal with is the most common exact match model.

Basic Assumption of Boolean Model

- An index term is either present(1) or absent(0) in the document
- All index terms provide equal evidence with respect to information needs.
- Queries are Boolean combinations of index terms.
- Each query term specifies a set of documents containing the term
  - AND ($\wedge$): the intersection of two sets
  - OR ($\vee$): the union of two sets
  - NOT ($\neg$): set inverse, or really set difference
  - X AND Y: represents doc that contains both X and Y
  - X OR Y: represents doc that contains either X or Y
  - NOT X: represents the doc that does not contain X

GQ. Explain Boolean model with example.

The Boolean Model Example

- Consider the terms: K1, K2, K3, ...,K8.
- 6 documents containing different terms:

  D1 = {K1, K2, K3, K4, K5}

  D2 = {K1, K2, K3, K4}

  D3 = {K2, K4, K6, K8}

  D4 = {K1, K3, K5, K7}

D5 = {K4, K5, K6, K7, K8}

D6 = {K1, K2, K3, K4}

- **Query :** K1 ∧ (K2 ∨ ¬K3) e.g documents containing K1 and (K2 or (not K3)
- **Answer :**

{D1, D2, D4, D6} ∩ ({D1, D2, D3, D6} ∪ {D3, D5}) = {D1, D2, D6}

Definition : For the Boolean model, the index term weight variables are all binary i.e., $w_{i,j} \in \{0, 1\}$ A query q is a conventional Boolean expression. Let $\vec{q}_{dnf}$ be the disjunctive normal form for the query q. Further, let $\vec{q}_{cc}$ be any of the conjunctive components of $\vec{q}_{dnf}$. The similarity of a document d_j to the query q is defined as

$$sim(d_j, q) = \begin{cases} \text{if } \exists \vec{q}_{cc} \mid \left(\vec{q}_{cc} \in \vec{q}_{dnf}\right) \wedge \left(\forall k_i\, g_i\left(\vec{d}_j\right)\right) \\ 0 \quad \text{otherwise} \end{cases}$$

$$= g_i\left(\vec{q}_{cc}\right)$$

If sim(dj, q) = 1 then the Boolean model predicts that the document dj is relevant to the query q (it might not be). Otherwise, the prediction is that the document is not relevant.

☞ Advantages of the Boolean Mode

(1) The simplest model is based on sets.

(2) Easy to understand and implement.

(3) It only retrieves exact matches

(4) It gives the user, a sense of control over the system.

(5) Boolean retrieval was adopted by many commercial bibliographic systems.

(6) Boolean queries are akin to database queries.

☞ Disadvantages of the Boolean Model

(1) The model's similarity function is Boolean. Hence, there would be no partial matches. This can be annoying for the users.

(2) Information need has to be translated into Boolean expressions which most users find awkward.

(3) In this model, the Boolean operator usage has much more influence than a critical word.

(4) The Boolean queries formulated by the users are most often too simplistic.

(5) As a result, the Boolean model frequently returns either too few or too many documents in response to the user query.

(6) The query language is expressive, but it is complicated too.

(7) No ranking for retrieved documents (absence of grading scale).

(8) It is not possible to assign a degree of relevance.

2.5.2 Vector Model

Also called term vector models, the vector space model is an algebraic model for representing text documents (or also many kinds of multimedia objects in general) as vectors of identifiers such as index terms.

The vector space model is based on the notion of similarity between the search document and the representative query prepared by the user which should be similar to the documents needed for information retrieval.

GQ. Define the Vector Model with relevant mathematical equations.

GQ. What are the assumptions of vector space model?

GQ. What are the Parameters in calculating a weight for a document term or query term?

GQ. How can you calculate tf and idf in the vector model?

- The vector model suggests a framework that allows for partial matching because it acknowledges that using binary weights is too restrictive.
- It assigns non-binary weights to index terms in queries and documents.
- The degree of similarity between each document stored in the system and the user query is calculated using these term weights.
- The vector model considers documents that match the query terms only partially by ordering the retrieved documents in decreasing order of this degree of similarity.
- In comparison to the Boolean model, the ranked document answer set is significantly more precise (in the sense that it better satisfies the user's information need).

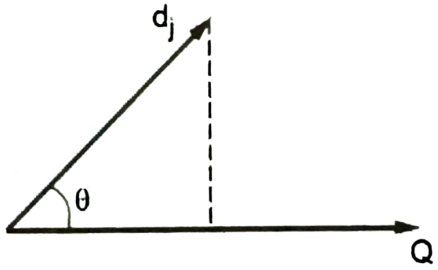
Definition : For the vector model, the weight $W_{i,j}$ associated with a pair (k_i, d_j) is positive and non-binary. Further, the index terms in the query are also weighted. Let be the weight associated with the pair $[k_i, q]$, where $w_{i\,q} > 0$. Then, the query vector q is defined as $q = (w_{1,q}, w_{2,q}, \ldots w_{t,q})$ where t is the total number of index terms in the system. As before, the vector for a document d_j is represented by $d_j = (w_{i,j}, W_{2,j}, \ldots, w_{t,j})$.

GQ. What is cosine similarity?
GQ. Define term frequency.
GQ. Define inverse term frequency.

- The vector model proposes to evaluate the degree of similarity of the document d_j with regard to the query q as the correlation between the vectors d_j and q.
- For instance, this correlation can be quantified by the cosine of the angle between these two vectors as shown in Fig. 2.5.1. That is,

$$sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

where $|d_j|$ and $|q|$ are the norms of the document and query vectors. The factor $|q|$ does not affect the ranking (i.e., the ordering of the documents) because it is the same for all documents. The factor $|d_j|$ provides a normalization in the space of the documents.



(1B2)**Fig. 2.5.1 : The cosine of Q is adopted as *sim* (d_j, q).**

- By calculating the raw frequency of a phrase (k_i) within a document (d_j), the vector model measures the intra-clustering similarity.
- Such term frequency is usually referred to as the tf factor and provides one measure of how well that term describes the document contents (i.e., intra-document characterization).
- The inverse of the frequency of a phrase k_i among the documents in the collection is used to calculate the inter-cluster dissimilarity. This factor is known as the inverse document frequency or the idf factor.

If the words of the documents are chosen to be the terms in the above representation, the dimensionality of the vectors is the number of distinct words occurring in the vocabulary of the corpus.

The value of the vector is non-zero if a term or word occurs in the document. Many different schemes can be taken for example count of occurrences, normalized counts, etc. but the most popular and efficient representation is tf-idf term weighing. The vector space model represents the documents and queries as vectors in a multidimensional space whose dimensions are the terms further used to build an index to represent the documents

Definition : Let N be the total number of documents in the system and n_i be the number of documents in which the index term k_i appears. Let $freq_{i,j}$ be the raw frequency of term k_i in the document d_j (i.e., the number of times the term k_i is mentioned in the text of the document d_j). Then, the normalized frequency $f_{i,j}$ of term k_i in document d_j is given by $f_{i,j} = \frac{freq_{i,j}}{max_i\ freq_{i,j}}$ where the maximum is computed over all terms which are mentioned in the text of the document d_j. If the term k_i does not appear in the document d_j then $f_{i,j} = 0$. Further, let idf_i inverse document frequency for k_i be given by $idf_i = \log \frac{N}{n_i}$

Weights are given by $w_{i,j} = f_{i,j} \times \log \frac{N}{n_i}$

Such term weighting schemes are called tf-idf schemes.

The Vector Model Example

- Let's consider that a collection includes 10,000 documents
- The term A appears 20 times in a particular document
- The maximum appearance of any term in this document is 50
- The term A appears in 2,000 of the collection documents.

$f(i,j) = freq(i,j) / max(freq(i,j)) = 20/5 = 0.4$

$idf(i) = \log(N/n_i) = \log(10{,}000/2{,}000) = \log(5) = 2.32$

$w_{i,j} = f(i,j) * \log(N/n_i) = 0.4 * 2.32 = 0.93$

GQ. What are the advantages and disadvantages of the Vector Model?

☞ Advantages of Vector Space Model

(1) Its term-weighting scheme improves the quality of answer set and retrieval performance.

(2) Its partial matching strategy allows retrieval of documents that approximate the query conditions.

(3) Its cosine ranking formula sorts the documents according to their degree of similarity to the query.

☞ Disadvantage of Vector Space Model

(1) The assumption of mutual independence between index terms

2.5.3 Probabilistic Model

GQ. What are the Fundamental assumptions for probabilistic principle?
GQ. Write the advantages and disadvantages of probabilistic model.

Probabilistic models provide the foundation for reasoning under uncertainty in the realm of information retrieval.

- The probabilistic model is an effort to frame the information retrieval problem within a probabilistic framework.
- The probabilistic model tries to estimate the probability that the user will find the document d_j relevant with ratio

$$\mathbf{P\ (d_j\ relevant\ to\ q)\ /\ P\ (d_j\ non\ relevant\ to\ q)}$$

- It is useful to derive ranking functions used by search engines and web search engines in order to rank matching documents according to their relevance to a given search query
- This model is used to calculate the probability that a document, d_j, will be relevant to a given query, q.
- The model makes the assumption that the query and document representations influence this probability of relevance.
- *Given a query q, there exists a subset of the documents R which are relevant to q But membership of R is uncertain*
- Users give with information needs, which they translate into query representations. Similarly, there are documents, which are converted into document representations. Given only a query, an IR system has an uncertain understanding of the information needed.
- So IR is an uncertain process, because,
  - Information need to query
  - Documents to index terms
  - Query terms and index terms mismatch
- Probability theory provides a principled foundation for such reasoning under uncertainty. This model provides how likely a document is relevant to an information need.
- Documents can be relevant and non-relevant, we can estimate the probability of a term t appearing in a relevant document P(t | R=1).

Probabilistic methods are one of the oldest but also one of the currently hottest topics in Information Retrieval.

For Probabilistic model

GQ. How can you find the similarity between doc and query in probabilistic principle Using Bayes' rule?

- All index term weights are all binary i.e., $w_{i,j} \in \{0,1\}$, $w_{i,q} \in \{0,1\}$
- Let R be the set of documents known to be relevant to query q
- Let R' be the complement of R.
- Let (Rld) be the probability that the document d_j is relevant to the query q
- Let P(R'ldj) be the probability that the document d_j is non-relevant to the query q
- The similarity $sim(d_j,q)$ of the document d_j to the query q is defined as the ratio

$$sim(d_j, q) = \frac{P(R|\vec{d}_j)}{P(R|\vec{d}_j)}$$

using Bayes' rule,

$$sim(d_j, q) = \frac{P(\vec{d}_j|R) \times P(R)}{P(\vec{d}_j|\vec{R}) \times P(\vec{R})}$$

- $P(d_j|R)$ stands for the probability of randomly selecting the document dj from the set R of relevant documents.
- P(R) stands for the probability that a document randomly selected from the entire collection is relevant

☞ **Advantage of Probabilistic Model**

(1) Documents are ranked in decreasing order of probability of relevance.

☞ **Disadvantages of Probabilistic Model**

(1) Need to guess initial estimates for $P(K_i | R)$

2.6 ALTERNATIVE SET THEORETIC MODELS

GQ. Discuss alternative set theoretic models.

In this section, we discuss two alternative set theoretic models, namely the fuzzy set model and the extended Boolean model.

2.6.1 Fuzzy Set Model

GQ. Explain fuzzy set model
GQ. Write basics of fuzzy set theory.

- When documents and queries are represented by sets of keywords, descriptions that are only loosely related to the actual semantic contents of the corresponding documents and queries are produced.
- As a result, there is only a rough match between a document and the search terms (or vague).
- This can be represented mathematically by assuming that each query phrase defines a fuzzy set and that each page has a degree of membership (often smaller than 1) in this set.
- This interpretation provides the foundation for many models of IR based on fuzzy theory.

Basics of Fuzzy Set Theory

- Fuzzy sets theory is an extension of classical set theory.
- Elements have a varying degree of membership. A logic based on two truth values,
- True and False are sometimes insufficient when describing human reasoning.
- Fuzzy Logic uses the whole interval between 0 (false) and 1 (true) to describe human reasoning.
- A Fuzzy Set is any set that allows its members to have different degree of membership, called membership function, having interval [0, 1].
- Fuzzy Logic is derived from fuzzy set theory
- Many degrees of membership (between 0 to 1) are allowed.
- Thus a membership function $\mu A(x)$ is associated with a fuzzy sets $\tilde{A}$

such that the function maps every element of the universe of discourse X to the interval [0, 1].

- The mapping is written as: $\mu\tilde{A}$ (x): X –> [0, 1].
- Fuzzy Logic is capable of handling inherently imprecise (vague or inexact or rough or inaccurate) concepts
- A fuzzy set is defined as follows: If X is a universe of discourse and x is a particular element of X, then a fuzzy set A defined on X and can be written as a collection of ordered pairs A = { (x, $\mu\tilde{A}$ (x)), x ∈ X }

GQ. Define membership function.
GQ. Explain fuzzy information retrieval.

Example

- Let X = {g1, g2, g3, g4, g5} be the reference set of students.
- Let $\tilde{A}$ be the fuzzy set of "smart" students, where "smart" is a fuzzy term.

  $\tilde{A}$ = {(g1 ,0.4) (g2 ,0.5) (g3 ,1) (g4 ,0.9) (g5 ,0.8)}
- Here $\tilde{A}$ indicates that the smartness of g1 is 0.4 and so on
- **Membership Function :** The membership function fully defines the fuzzy set. A membership function provides a measure of the degree of similarity of an element to a fuzzy set

Fuzzy Information Retrieval

- The main idea is to supplement the query's index terms with related terms (obtained from a thesaurus) so that the user query can acquire more relevant pages
- By creating a term-term correlation matrix (referred to as a keyword connection matrix in whose rows and columns are connected to the index terms in the document collection, a thesaurus can be created. In this matrix C, a normalized correlation factor $C_{i,l}$ between two terms k_i and k_l can be defined by

$$C_{i,\,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$$

Where n_i is the number of documents which contain the term k_i, n_l is the number of documents which contain the term k_l, *and* $n_{i,l}$ is the

number of documents which contain both terms.

- In this fuzzy set, a document d_j has a degree of membership $u_{i,j}$ computed as

$$u_{i,j} = 1 - \prod_{k_l \in d_j} (1 - C_{i,l})$$

which computes algebraic sum over all terms in document d_j

2.6.2 Extended Boolean Model

GQ. Discuss extended Boolean model.

- In the Boolean model, no provision for term weighting and no ranking of the answer set is generated.
- As a result, the size of the output might be too large or too small
- However, an alternative strategy is to add the capabilities of term weighting and partial matching to the Boolean model. With this method, it's possible to integrate vector model properties with Boolean query constructions.
- The *extended Boolean model*, was introduced in 1983 by Salton, Fox, and Wu.

2.7 STRUCTURED TEXT RETRIEVAL MODELS

GQ. Explain Structured text retrieval models.

- Think about a user who has a strong visual memory. A user of this type would then remember that the particular document in which he is interested has a page where the phrase "*Nuclear Blast*" occurs in italics in the text around a Figure whose label contains the word "earth."
- This query may be phrased as ['*Nuclear Blast*' and 'earth'] in a traditional information retrieval approach, which would return all pages containing both strings. But it's clear that this customer didn't want as many documents as this answer provides.
- In this scenario, the user wants to make his inquiry clearer by using a richer expression, like

same-page (near *Nuclear Blast*,' Figure (label ('earth')))

which conveys the details in his visual recollection

- Structured text retrieval models are types of retrieval models that incorporate information on both the text content and the document structure.
- Structured text retrieval models consider both the text's content and document structure.
- A structured text retrieval system looks for all the documents that match the search criteria, that's why the retrieval job is not associated with any idea of relevance.
- The current models for structured text retrieval are data retrieval models rather than information retrieval models.
- The retrieval system could search for documents that match the query conditions only partially
- The position in the test of a string of words that matches the user query is referred to as the "term match point."

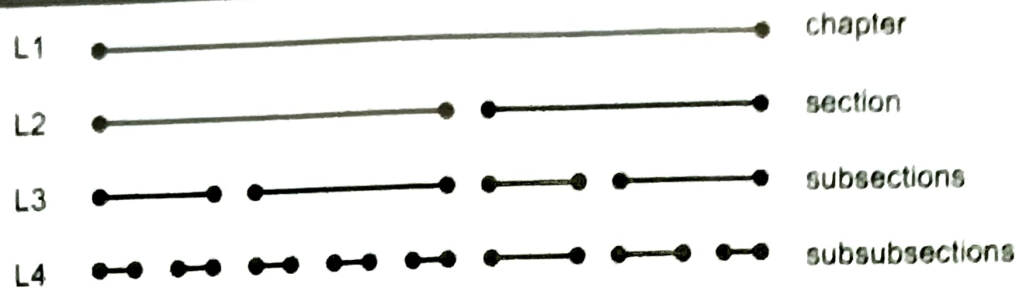
  e.g user query: ['information retrieval system']

  if this appears at 3 positions in document d_j , then match points are 3.

2.7.1 Model based on Non-overlapping lists

GQ. Explain non overlapping lists with the help of an example.

- Each document's whole text is divided into a list of non-overlapping text sections.
- Multiple lists are generated as there are various ways to break a text into non-overlapping sections. For example
  (1) A List for chapters
  (2) A List for sections
  (3) A List for subsections
- These lists are kept as separate and distinct data structures.
- A single inverted file is built with each structural element to allow searching for both index terms and text areas. Fig. 2.7.1 shows an example of different lists.



(183)**Fig. 2.7.1 : Structure of text documents through different indexing list**

Implementation

- A single inverted file is built, in which each structural component stands as an entry in the index
- Each entry has a list of text regions as a list of occurrences
- Such a list could be easily merged with the traditional inverted file

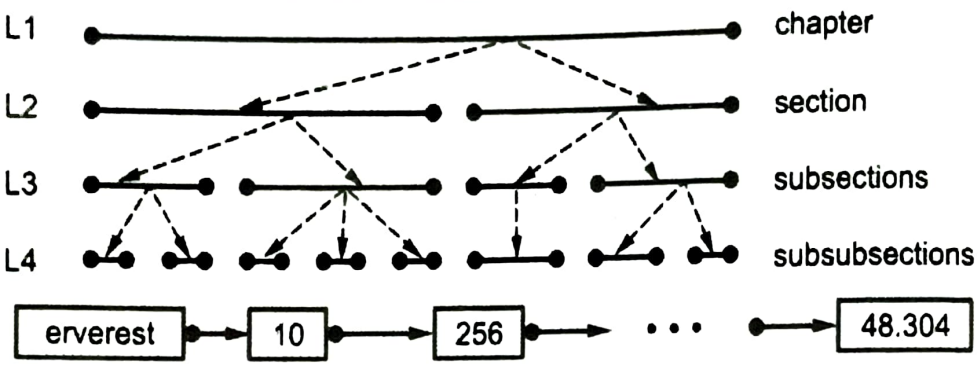
Example types of queries

- Select a region that contains a given word
- Select a region A which does not contain any other region B
- Select a region not contained within any other region

2.7.2 Model Based on Proximal Nodes

GQ. Discuss model based on proximal nodes.

- This model was proposed by Navarro and Baeza-Yates
- Basic idea is to define a strict hierarchical index over the text. This enriches the previous model that uses flat.
- It allows the definition of independent hierarchical (non-flat) indexing structures over the same text of the document.
- Every indexing system is made up of nodes, which are chapters, sections, paragraphs, pages, and lines.
- Each node is associated with a text region.
- If user query refers to different hierarchies, compiles answer is formed by nodes which all come from only one of them.
- This type of models allow us to formulate more complex queries than the model based on non-overlapping lists.
- Only nearby (proximal) nodes are looked for faster query processing. Fig. 2.7.2 shows the hierarchical indexing structure of four levels and an inverted list for the word 'Everest'.



(1B4)**Fig. 2.7.2 : Hierarchical indexing structure**

Features

- One node might be contained within another node.
- But two nodes of the same hierarchy cannot overlap.
- The inverted list for words complements the hierarchical index.
- Query language in regular expression
  - (1) Searches for string
  - (2) Reference to structural components by name
  - (3) Combination of these
- An example query [(*section) with ("Everest")]
- Searches for the sections, the subsections, and the sub-subsections that contain the word "Everest"
- Model is a compromise between expressiveness and efficiency

2.8 MODELS FOR BROWSING

Sometimes the user is interested to spend some time in exploring the document, looking for interesting references instead of searching for a specific query.

- Users have goals to pursue in both cases
- But the searching task's goal is more clear than a browsing task's goal in the user's mind.

☞ Types of Browsing

GQ. What are different types of browsing.

(1) Flat Browsing

- Documents are represented as dots in a (two- dimensional) plan or as elements in a (single dimension) list.
- The user then glances here and there looking for information within the documents visited
- The user looks for correlations among neighbor documents or for keywords
- These keywords could be added to the original query for query expansion and this process is called relevance feedback. this helps in the retrieval of more relevant documents.
- Users can also explore a single document in a flat manner (like a web page)

☞ Drawback

On a given page user may not have an indication about the context where the user is. For example, if a user opens a book on a random page, he might not know in which chapter that page is.

(2) Structure Guided Browsing

- Documents are organized in a structure as a directory to help users in browsing.
- Directories are hierarchies of classes that group documents covering related topics
- These hierarchies of classes have been used to classify document collections. E.g.: "Yahoo!" provides a hierarchical directory
- The user performs a structured guided type of browsing.
- The same idea applied to a single document
  - Chapter level, section level, etc.
  - The last level is the text itself (flat!)
  - A good UI is needed for keeping track of the context in a focused manner.

- e.g. the "adobe acrobat pdf" files

- Additional facilities are provided when searching such as
  - A history map to identify classes recently visited
  - Display occurrences (of terms) by showing the structures in a global context, in addition to the text positions

(3) The Hypertext Model

- The fundamental concept related to the task of writing is the notion of sequencing
- A sequenced organizational structure lies underneath the most written text
- The reader should not expect to fully understand the message conveyed by the writer by randomly reading pieces of text here and there
- Sometimes, we even can't capture the information through sequential reading of the whole text
- For example, a book about "the history of the wars" is organized chronologically, but the user might in interested in wars fought by any particular army or country, in such case user will have a tough time finding the information he is interested in.
- Because contents are organized sequentially
- in these situations, one of the possible solutions is to rewrite the book but there is no point in rewriting the book
- Another solution is to define a new structure to organize the contents which can be achieved through the design of hypertext.

Hypertext

- A high-level interactive navigational structure allows users to browse text non-sequentially
- Consist of nodes (text regions) correlated by directed links in a graph structure
  - A node could be a chapter in a book, a section in an article, or a web page
  - Links are attached to specific strings inside the nodes

- The process of navigating the hypertext can be understood as a traversal of a directed graph.
- Hypertexts provide the basis for HTML(Hyper Text Markup Language) and HTTP(Hypertext Transfer Protocol)

☞ Drawbacks of Hypertext

(1) Loose in hyperspace : the user will lose track of the organizational structure of the hypertext when it is large

A hypertext map shows where the user is at all times (graphical user interface design)

(2) But, the user is restricted to the intended flow of information previously convinced by the hypertext designer

Should take into account the needs of potential users

Analyzing the requirements before starting implementation of hypertext is required

(3) During the hypertext navigation, the user might find it difficult to orient himself Guiding tools can help in navigation (hypertext map)

Short Questions and Answers

Q. 1 What do you mean information retrieval models?

Ans. :

A retrieval model can be a description of either the computational process or the human process of retrieval: The process of choosing documents for retrieval; the process by which information needs are first articulated and then refined.

Q. 2 What is cosine similarity?

Ans. :

This metric is frequently used when trying to determine similarity between two documents. Since there are more words that are in common between two documents, it is useless to use the other methods of calculating similarities

Q. 3 What are the characteristics of relevance feedback?

Ans. :

(1) It shields the user from the details of the query reformulation process.

(2) It breaks down the whole searching task into a sequence of small steps which are easier to grasp.

(3) Provide a controlled process designed to emphasize some terms and de-emphasize others.

Q. 4 What are the assumptions of vector space model?

Ans. :

(1) Assumption of vector space model:

(2) The degree of matching can be used to rank-order documents

(3) This rank-ordering corresponds to how well a document satisfying a user's information needs

Q. 5 What are the disadvantages of Boolean model?

Ans. :

(1) It is not simple to translate an information need into a Boolean expression.

(2) Exact matching may lead to retrieval of too many documents.

(3) The retrieved documents are not ranked

(4) The model does not use term weights

Q. 6 Define term frequency.

Ans. :

Term frequency : Frequency of occurrence of query keyword in document

Q. 7 What are the three classic models in information retrieval system?

Ans. :

(1) Boolean model

(2) Vector Space model

(3) Probabilistic model

Q. 8 What is the basis for Boolean model?

Ans. :

Simple model based on set theory and Boolean algebra

(1) Documents are sets of terms

(2) Queries are specified as Boolean expressions on terms.

Q. 9 What are the disadvantages of Boolean model?

Ans. :

Exact matching may retrieve too few or too many documents

(1) Difficult to rank output, some documents are more important than others.

(2) Hard to translate a query into a Boolean expression

(3) All terms are equally weighted

(4) More like data retrieval than information retrieval

(5) No notion for partial matching

Q. 10 What are the Fundamental assumptions for probabilistic principle?

Ans. :

- q- user query,dj – doc in the collections
- Model assumes, relevance depends on the query and the doc representation only
- R – ideal answer set, relevant to the query
- R - ideal answer set, non-relevant to the query
- Similarity to the query ratio is, i.e. probabilistic ranking computed as
- Ratio = P(dj relevant-to q) / P(dj non-relevant-to q)
- The rank minimizes the probability of the erroneous judgment

Q. 11 Write the advantages and disadvantages of probabilistic model:

Ans. :

Advantages

(1) Doc's are ranked in decreasing order of their probability of relevant

Disadvantages

(1) Need to guess the initial separation of doc's into relevant and non-relevant sets.

(2) All weights are binary

(3) The adoption of the independence assumption for index terms

(4) Need to guess initial estimates for P(ki | R)

(5) Method does not take into account tf and idf factors

Q. 12 Why Classic IR might lead to poor retrieval ?

Ans. :

(1) The user information need is more related to concepts and ideas than to index terms but in classic IR.

(2) Unrelated documents might be included in the answer set.

(3) Relevant documents that do not contain at least one index term are not retrieved.

(4) Reasoning : retrieval based on index terms is vague and noisy.

Chapter Ends...

❑❑❑