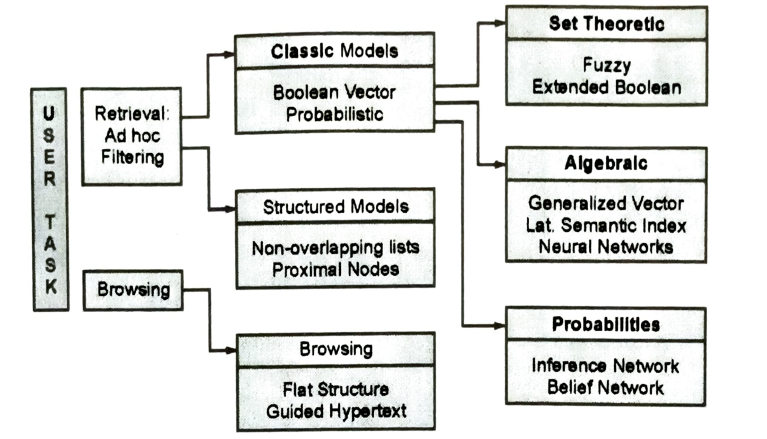
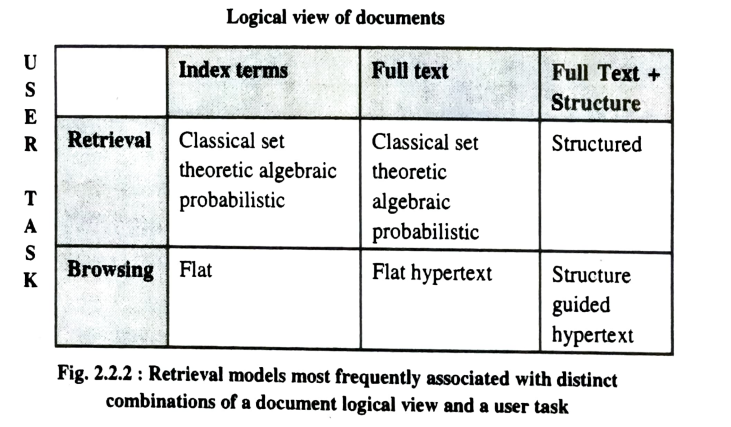
Taxonomy of IR Models





Classic Information Retrieval Models:

The classical IR systems are designed based on **mathematical concepts** and are the most widely used, simplest, and easy-to-implement systems for information retrieval models.

In this system, the retrieval of information depends on **documents** containing the defined set of queries and there is **no ranking** or grading of any kind.

Classic Information Retrieval models can be easily implemented and updated accordingly.

The different classical IR models are based on **mathematical knowledge** combined with concepts like Document Representation, Query representation, and Retrieval / Matching function taken into account.

The term **classic** in the name of classical IR systems denotes that they use **foundational techniques** for text documents without extra information about the structure or content of a document.

The different types of classical Information Retrieval models are: Boolean models, Vector space models, and Probabilistic IR models.

1. Boolean Model:

The Boolean model in information retrieval is based on the set theory and boolean algebra.

We can pose any query in the form of a Boolean expression, where the terms are logically combined using the Boolean operators AND, OR, and NOT in the Boolean retrieval model.

An index term is either present(1) or absent(0) in the document.

The Boolean operators, the terms in the query and the concerned documents can be combined to form a whole new set of documents.

* The Boolean AND of two logical statements x and y means that both x AND y must be satisfied and will be a set of documents that will smaller or equal to the document set.
* The Boolean OR of these same two statements means that at least one of these statements must be satisfied and will fetch a set of documents that will be greater or equal to the document set otherwise.
* The Boolean NOT of a statement x means that documents that satisfy x will be excluded from the result set. The resulting set will be smaller or equal to the original document set because it eliminates documents that meet the criteria specified by x.

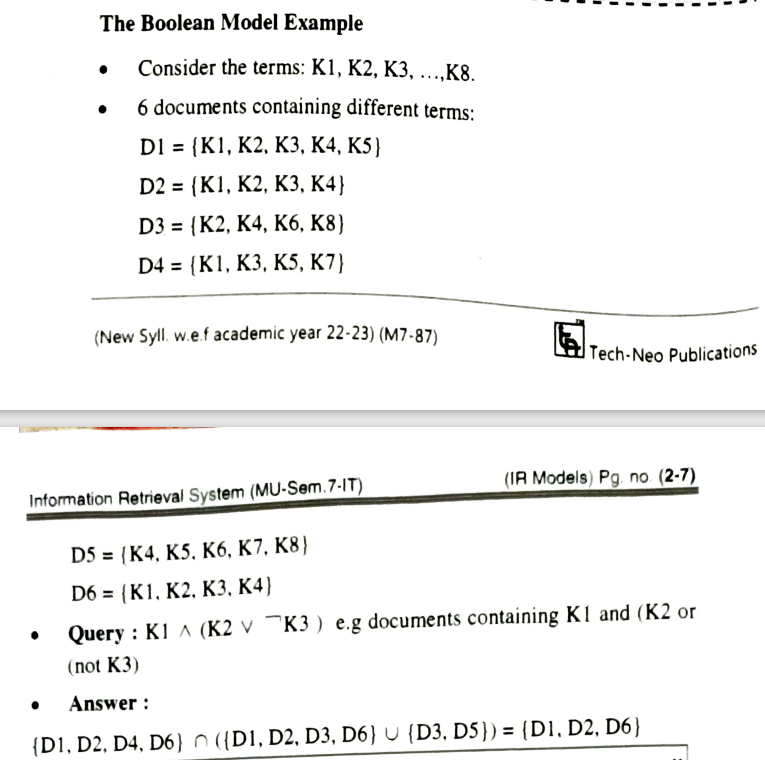
Any number of logical statements can be combined using the three Boolean operators.

The queries are designed as boolean expressions which have precise semantics and the retrieval strategy is based on **binary decision criterion.**

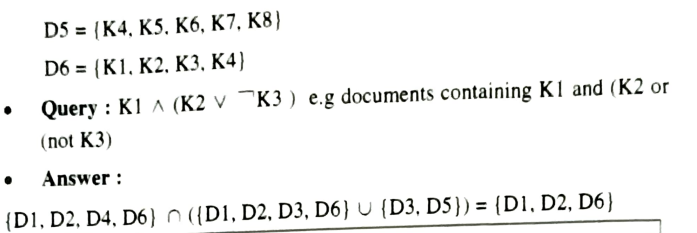
The Boolean model can also be explained well by **mapping the terms in the query with a set of documents.**

In the exact match, a query will specify a criteria. Each document either matches or fails to match the query. - not ranked

In the best match, a query describes good or best matching documents. - ranked



The most famous web search engine, Google also ranks the web page result set based on a two-stage system: In the first step, a Simple Boolean Retrieval model, returns matching documents in no particular order, and in the next step ranking is done according to some estimator of relevance.



Processing the data for Boolean retrieval model

We should strip unwanted characters/markup like HTML tags, punctuation marks, numbers, etc. before breaking the corpus into tokens on whitespace.

Stemming needs to be done and then common stopwords are to be removed depending on the application need.

The **document incidence matrix** or inverted index iis built.It consists of a mapping between keywords and the list of documents that contain those keywords.

Then the common queries/phrases may be detected using a domain-specific dictionary if needed.

1. VECTOR SPACE MODEL - algebraic - mathematical and geometric approach

**Notion of Similarity in Vector Space Model**

**Angle of deviation between query and document:** One way is to compare the deviation of angles between each document vector and the original query vector where the query is represented as some kind of vector as the documents.

**Cosine distance as similarity metric:** The most popular and easier method in practice is to calculate the cosine of the angle between the vectors - A cosine value of zero means that the query and document vector are orthogonal and have no match at all.

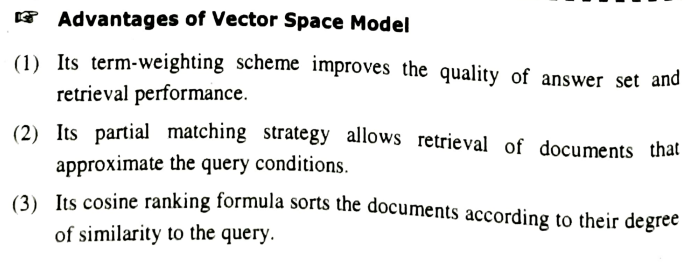
A zero cosine similarity value implies that the terms in the query do not exist in the document we are considering.

**Assumptions of the Vector Space Model**

The more similar a document vector is to a query vector, the more likely it is that the document is relevant to that query.

The words used to define the dimensions of the space are orthogonal or independent.

The similarity assumption is an approximation and realistic whereas the assumption that words are pairwise independent doesn't hold true in realistic scenarios.



**Disadvantages of Vector Space Model**

1. Long documents are poorly represented because they have poor similarity values due to a small scalar product and a large dimensionality of the terms in the model.
2. Search keywords must be precisely designed to match document terms and the word substrings might result in a false positive match.
3. Semantic sensitivity: Documents with similar context but different term vocabulary won't be associated resulting in false negative matches.
4. The order in which the terms appear in the document is lost in the vector space representation.
5. Weighting is intuitive but not represented formally in the model.
6. Adding a single new document changes the frequencies of terms that occur in the documents, which changes the vector lengths of every document that contains one or more of these terms.
7. **Probabilistic Model**

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

**Uncertainty in retrieval models:**

The probabilistic models in information retrieval are built on the idea that the process of retrieval is inherently uncertain from multiple standpoints:

1. There is uncertainty in the understanding of user’s information needs - We can not sure that the user mapped their needs into the query they have presented.
2. Even if the query represents the need well, there is uncertainty in the estimation of document relevance for the query which stems from either the uncertainty from the selection of the document representation or the uncertainty from matching the query and documents.

**Basis of probabilistic retrieval model:**

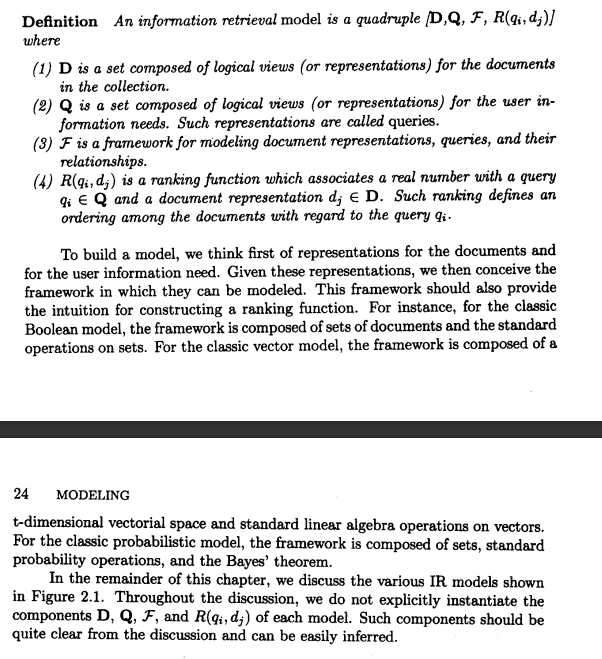
Probabilistic model is based on the Probability Ranking Principle which states that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query given all the other pieces of evidence available.

Probabilistic information retrieval models estimate how likely it is that a document is relevant for a query.

There may be a variety of sources of evidence that are used by the probabilistic retrieval methods and the most common one is the **statistical distribution of the terms in both the relevant and non-relevant documents.**

Probabilistic information models are also among the oldest and best performing and most widely used IR models.

Formal characterization of IR Models



Ad Hoc Retrieval: Ad hoc retrieval refers to the retrieval of documents from a collection based on user queries submitted at the moment (ad hoc) without any prior knowledge or filtering.

Web search engines like Google primarily use ad hoc retrieval. When you enter a search query, the search engine retrieves web pages that match your query, regardless of whether they were previously seen or indexed.

Filtering retrieval, also known as filtering-based information retrieval, focuses on the continuous monitoring and filtering of documents or information streams to deliver relevant content to users.

Email spam filters, news recommendation systems, and social media content feeds are examples of filtering retrieval. For instance, an email spam filter automatically filters out unwanted or irrelevant emails from your inbox based on predefined criteria.

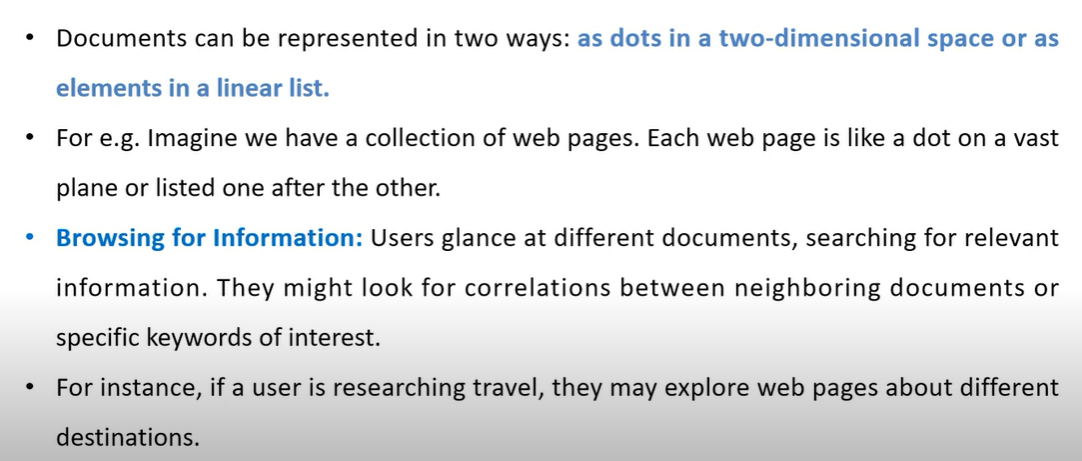
Extended Boolean Model:

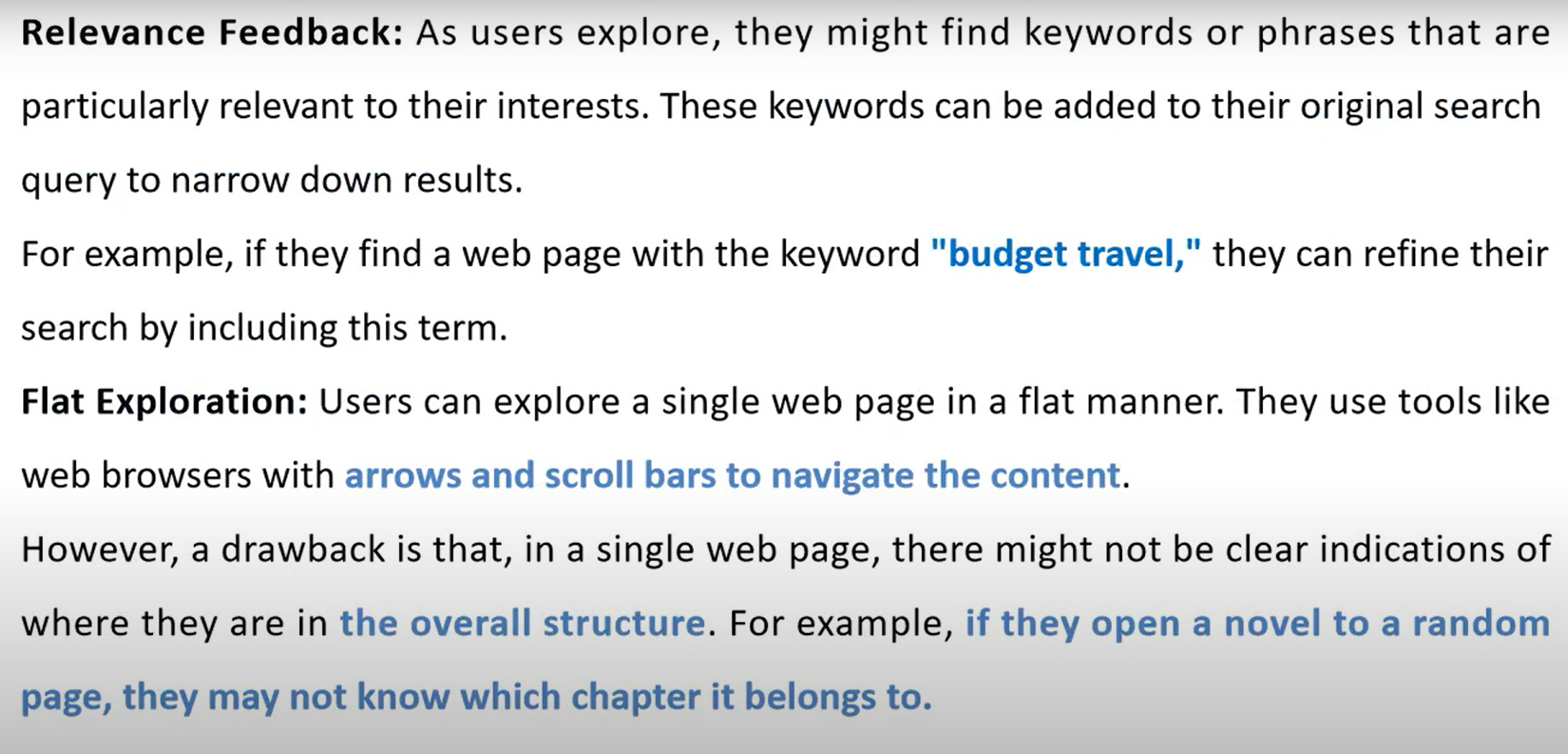
In the three models discussed below, a document has a weight associated with each index term. This document weight is a measure of the degree to which the document is characterized by that term. Without loss of generality, we assume that document weights for all index terms lie in the range [0, 1]. This is less restrictive than in the standard Boolean model, which limits the values to the extremes of the range, namely 0 and 1.

To retrieve documents relevant to a given query, we need to calculate the query-document similarity for documents in the collection. The query-document similarity is an attempt to predict the relevance of a document to the query. In the following subsections, we consider each model and its method for calculating similarity.

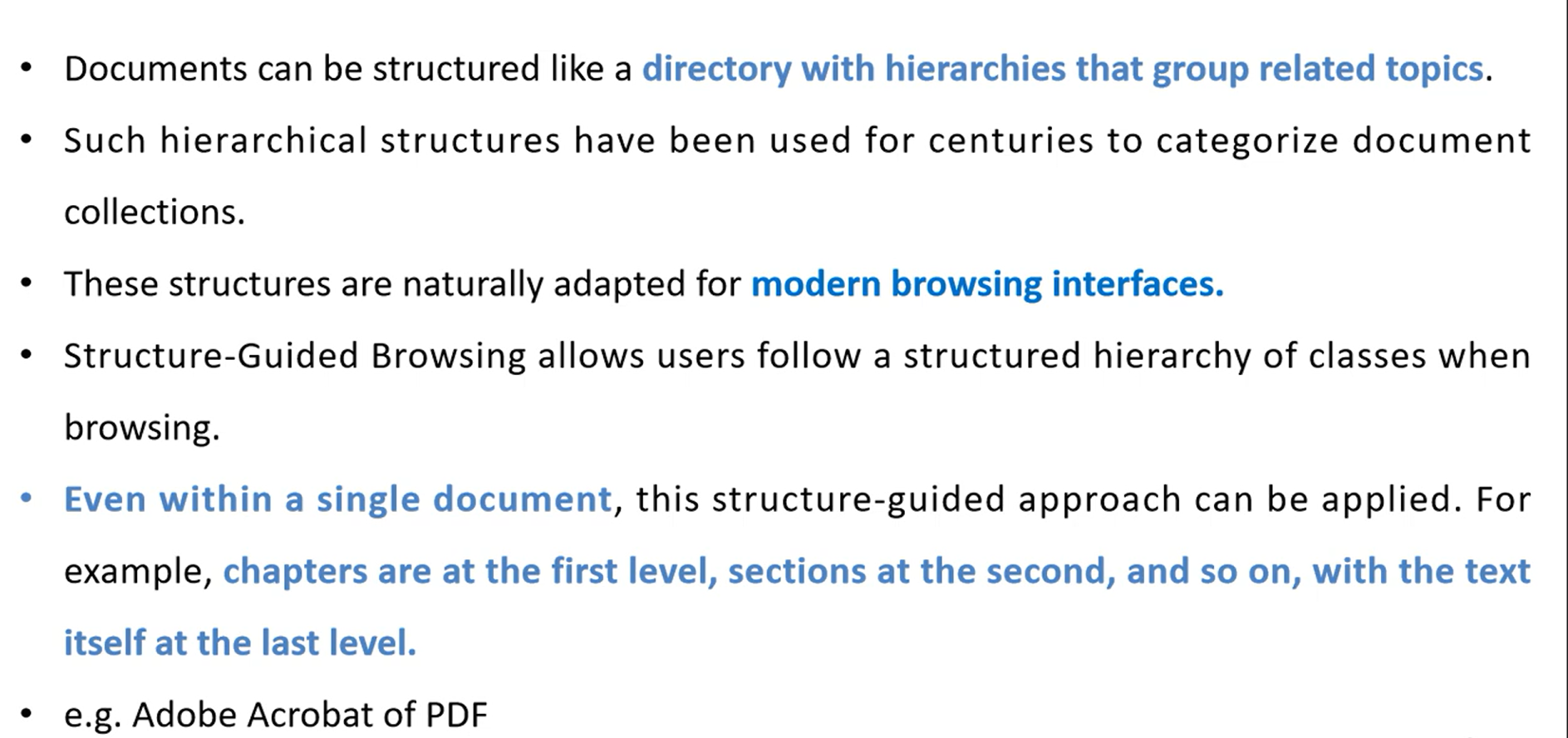
The Extended Boolean Model is really just a general term that refer to all IR model that improve upon the Standard Boolean Model. There had been proposed many such model, such as the MMM model, the Paice model and P-norm model.

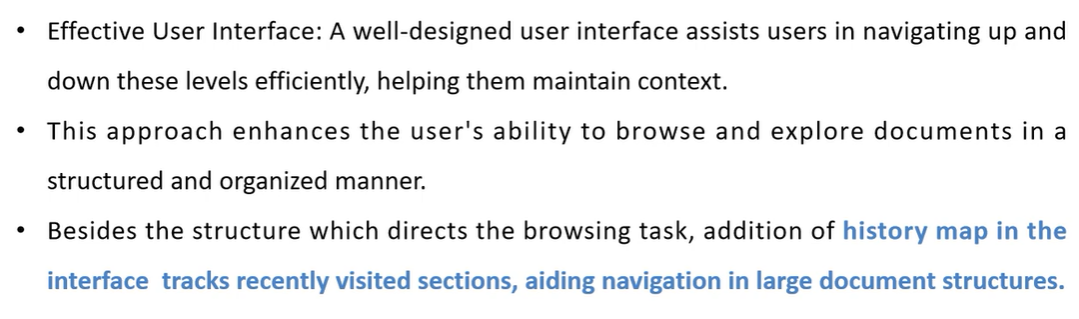
Flat browsing





Structure Guided Browsing





Hypertext browsing

