1.7 Exam Pack (Review Questions)

Syllabus Topic : Introduction

Q. 1 Write a short note on Information Retrieval. (Refer Section 1.1) **(5 Marks)**

Q. 2 Explain the Interaction of user with retrieval system using diagram. (Refer Section 1.1.2) **(5 Marks)**

Q. 3 Write a short note on Past, Present and Future of Information Retrieval. (Refer Section1.1.3) **(5 Marks)**

Syllabus Topic : History of IR

Q. 4 Explain the History of Information Retrieval in detail. (Refer Section 1.2) **(5 Marks)**

Syllabus Topic : Components of IR

Q. 5 Explain the components of Information Retrieval in detail using diagram. (Refer Section 1.3) **(5 Marks)**

Syllabus Topic : Issues Related to IR

Q. 6 Explain the issues of Information Retrieval in detail. (Refer Section 1.4) **(5 Marks)**

Syllabus Topic : Boolean Retrieval

Q. 7 Explain the basic terms in Boolean Retrieval. (Refer Section 1.5) **(5 Marks)**

Q. 8 Explain with an example the concept of information Retrieval problem in detail. (Refer Section 1.5.1) **(5 Marks)**

Q. 9 Write a short note on Inverted Index and explain with example. (Refer Section 1.5.2) **(5 Marks)**

Syllabus Topic : Dictionaries and Tolerant Retrieval

Q. 10 Write a short note on : Wildcard Queries. (Refer Section 1.6.2) **(5 Marks)**

Q. 11 Write a note on Permuterm indexes. (Refer Section 1.6.2.1) **(5 Marks)**

Q. 12 Explain k-Gram indexes for wildcard queries in detail. (Refer Section 1.6.2.2) **(5 Marks)**

Q. 13 Write a short note on Spelling Correction. (Refer Section 1.6.3) **(2 Marks)**

Q. 14 Write the two forms of Spelling Correction and explain any one in detail. (Refer Section 1.6.3.2) **(5 Marks)**

Q. 15 Write a short note on Phonetic Correction. (Refer Section 1.6.4) **(5 Marks)**

Chapter Ends

Unit II

CHAPTER 2

Link Analysis and Specialized Search

Syllabus

Link Analysis and Specialized Search : Link Analysis, hubs and authorities, Page Rank and HITS algorithms, Similarity, Hadoop and Map Reduce, Evaluation, Personalized search, Collaborative filtering and content-based recommendation of documents and products, handling invisible Web, Snippet generation, Summarization, Question Answering, Cross- Lingual Retrieval.

Syllabus Topic : Link Analysis

2.1 Link Analysis

Q. 2.1.1 Define Link Analysis. (Ref. Sec. 2.1) **(2 Marks)**

Q. 2.1.2 Write a note on Link Analysis. (Ref. Sec. 2.1) **(5 Marks)**

- For the development of web search the analysis of hyperlinks and the graph structure of the Web are the basics. Link analysis for web search is the predecessors in the field of citation analysis, aspects of which overlap with an area that is called as bibliometrics.
- It quantifies the influence of scholarly articles by analyzing the pattern of citations. Citations represent the conferral of authority from a scholarly article to others.
- For example, One may contrive to set up multiple web pages pointing to a target web page, with the intent of artificially boosting the latter's tally of in-links which is called as link spam.

2.1.1 The Web as a Graph

The two instincts on which link analysis is built on are,

1. The anchor text pointing to page B is a good description of page B.
2. The hyperlink from A to B represents an acceptance of page B, by the creator of page A.

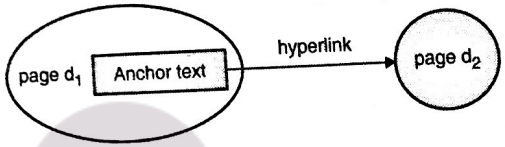

Fig. 2.1.1 : Web as a directed graph

2.1.1.1 Anchor Text and the Web Graph

Q. 2.1.3 Explain Anchor text and the web graph. (Ref. Sec. 2.1.1.1) **(5 Marks)**

- The below HTML code is from a web page which shows hyperlink pointing to the home page of the Journal of the ACM.

```
<a href="http://www.acm.org/jacm/"> Journal of ACM. </a>
```

- Here, the link points to the page http://www.acm.org/jacm/ and the anchor text is Journal of the ACM. There is a gap between the terms in a web page, and how web users would describe that web page.
- For example, the anchor text terms are included as terms under which to index the target web page.
- The use of anchor text has some side effects also, i.e. when we search for big blue we get on most of search engines the home page of IBM as big blue is the nickname of IBM whereas sometimes when we search for some other text some other result is shown which sometimes be misleading.

Syllabus Topic : Page Rank

2.1.2 Page Rank

Q. 2.1.4 Write a short note on Page Rank. (Ref. Sec. 2.1.2) **(5 Marks)**

- Page Rank is the first technique for link analysis which assigns to every node in the web graph a numerical score between 0 and 1 where the Page Rank of node depends on the link structure of the web graph.
- For example, when we consider a surfer who walks the web pages and executes a random walk such as, the surfer at a node A, out of which there are three hyperlinks to nodes B, C and D; the surfer proceeds at the next time step to one of these three nodes, with equal probabilities 1/3.
- The idea behind Page-Rank is that pages visited more often in this walk are more important. Page Rank is equal to long term visit rate which is further equal to steady state probability.

2.1.2.1 Markov Chains

Q. 2.1.5 Explain Markov Chains in detail. (Ref. Sec. 2.1.2.1) **(5 Marks)**

- A process that occurs in a series of time-steps in each of which a random choice is made is called as Markov chain which is a discrete-time stochastic process.
- This consists of N states. It is characterized by an N × N transition probability matrix Peach of whose entries is in the interval [0, 1] and the entries in each row of P add upto 1.
- The Markov chain can be in one of the N states at any given time step and then, the entry P_{ij} tells us the probability that the state at the next time step is j, conditioned on the current state being i.
- Each entry P_{ij} is known as a transition probability and depends only on the current state i; which is known as the Markov property.

- Thus, by the Markov property,

$$\forall i, j, P_{ij} \in [0, 1]$$

and

$$\forall i, \sum_{J=1}^{N} P_{ij} = 1 \qquad ...(2.1.1)$$

- A matrix with non-negative entries that satisfies the above Equation is known as a stochastic matrix which has a principal left eigenvector corresponding to its largest eigen value, which is 1. Fig. 2.1.2 shows a simple Markov chain.

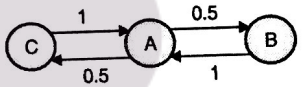

Fig. 2.1.2 : A simple Markov Chain

2.1.2.2 The Page Rank Computation

The left eigenvectors of the transition probability matrix P are N-vectors $\vec{\pi}$ are,

$$\vec{\pi} P = \lambda \vec{\pi} \qquad ...(2.1.2)$$

The N entries in the principal eigenvector $\vec{\pi}$ are the steady-state probabilities of the random walk with teleporting, and thus the Page Rank values for the corresponding web pages.

Syllabus Topic : Hubs and Authorities, HITS Algorithms

2.1.3 Hubs and Authorities

Q. 2.1.6 Write a note on Hubs and Authorities. (Ref. Sec. 2.1.3) **(5 Marks)**

- For a query, every web page is assigned two scores. One is called its hub score and the other is authority score. For any query, we compute two ranked lists of results rather than one.

- The ranking of one list is prompted by the hub scores and that of the other by the authority scores. Broad-topic searches are those that mean an informational query such as "I wish to learn about cancer" which we call authorities; in the computation we are about to describe, they are the pages that will emerge with high authority scores.

- A good hub page is one that points to many good authorities such as a good authority page is one that is pointed to by many good hub pages.

- For a web page v in subset of the web, we use h(v) to denote its hub score and a(v) its authority score. Initially, we set h(v) = a(v) = 1 for all nodes v and $v \rightarrow y$ the existence of a hyperlink from v toy.

- The below equation captures the intuitive notions that good hubs point to good authorities and that good authorities are pointed to by good hubs.

$$\left.\begin{array}{ll} v \rightarrow y & h(v) \leftarrow \Sigma\, a(y) \\ y \rightarrow v & a(v) \leftarrow \Sigma\, h(y) \end{array}\right\} \qquad ...(2.1.3)$$

- As shown in Equation (2.1.3) the first line sets the hub score of page v to the sum of the authority scores of the pages it links to i.e. if v links to pages with high authority scores, its hub score increases and the second line indicates that if page v is linked to by good hubs, its authority score increases.

- Now let $\vec{h}$ and $\vec{a}$, denote the vectors of all hub and all authority scores, for the pages in our subset of the web graph and let A be the adjacency matrix of the subset of the web graph where A is a square matrix with one row and one column for each page in the subset. The entry A_{ij} is 1 if there is a hyperlink from page i to page j, and else 0.

$$\left.\begin{array}{l} \vec{h} \leftarrow A\vec{a} \\ \vec{a} \leftarrow A^T\vec{h} \end{array}\right\} \qquad ...(2.1.4)$$

- Where A^T is the transpose of the matrix A. Substituting into one another the equation may be,

$$\left.\begin{aligned} \vec{h} &\leftarrow AA^T \vec{h} \\ \vec{a} &\leftarrow A^T A \vec{a} \end{aligned}\right\} \quad ...(2.1.5)$$

- If we replace the ← symbols by = symbols and introduce the eigen value.

$$\left.\begin{aligned} \vec{h} &= (1/\lambda_h) AA^T \vec{h} \\ \vec{a} &\leftarrow (1/\lambda_a) A^T A \vec{a} \end{aligned}\right\} \quad ...(2.1.6)$$

Here λ_h denotes the eigen value of AA^T and λ_a denotes the eigen value of A^TA.

Following are some of the effects,

1. The iterative updates in Equation (2.1.3), if scaled by the appropriate eigen values, are equivalent to the power iteration method for computing the eigenvectors of AA^T and A^TA.
2. In computing eigenvector entries, we are not restricted to using the power iteration method.

The resulting computation takes the following form :

1. Assemble the target subset of web pages, form the graph induced by their hyperlinks and compute AA^T and A^TA.
2. Compute the principal eigenvectors of AA^T and A^TA to form the vector of hub scores $\vec{h}$ and authority scores $\vec{a}$.
3. Output the top-scoring hubs and the top-scoring authorities.

This method of link analysis is known as HITS, i.e. Hyperlink-Induced Topic Search.

Page Rank can be precomputed, HITS has to be computed at query time. HITS is too expensive in most application scenarios. Page Rank and HITS make two different design choices concerning.

(i) The eigen problem formalization.

(ii) The set of pages to apply the formalization to.

These two are orthogonal. We can also apply HITS to the entire web and Page Rank to a small base set.

Syllabus Topic : Similarity

2.2 Similarity

Q. 2.2.1 Explain the concept of Similarity in detail. (Ref. Sec. 2.2) **(5 Marks)**

- Two documents are similar if they contain some of the same terms. There are some measures of similarity.
  1. The lengths of the documents.
  2. The number of terms in common.
  3. Whether the terms are common or unusual.
  4. How many times each term appears.
- The measures used to examine how documents are similar and the output is a numerical measure of similarity are shared word count, word count and bonus and cosine similarity.

2.2.1 Shared Word Count

- It counts the shared word between documents. For each document in the collection count how many of these words appear. Here no weighting are used, just a simple count is there. Each document is represented as a vector of key words (zeros and ones).
- The similarity of two documents is the product of the two vectors. The performance of this measure depends mainly on the dictionary used.

2.2.2 Word Count and Bonus

- Term frequency is the number of times a term occurs in a document. Document frequency is the number of documents that contain the term. Inversed document frequency is equal to log (N/df) where N is the total number of documents. Document similarity is defines as,

$$(D(i)) = \sum_{j-1}^{K} w(j),$$

$$w(j) = \begin{cases} 1 + \frac{1}{df(j)} \\ 0 \end{cases}$$

Where, k is the number of words.

The bonus 1/df(j) is a variant of inversed document frequency. Thus, if the word occurs in many documents, the bonus is small.

2.2.3 Cosine Similarity

A document is represented as a vector : $(V_1, V_2, \ldots \ldots, V_n)$. To compensate for the effect of document length, the standard way of quantifying the similarity between two documents d_1 and d_2 is to compute the cosine similarity of their vector representations $\vec{V}(d_1)$ and $\vec{V}(d_2)$.

$$sim(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}$$

Syllabus Topic : Hadoop and MapReduce

2.3 Hadoop and Map Reduce

Q. 2.3.1 Write a note on Hadoop and MapReduce. (Ref. Sec. 2.3) **(5 Marks)**

Q. 2.3.2 Explain MapReduce in detail with diagram. (Ref. Sec. 2.3) **(5 Marks)**

- Big data is large amount of files as we say in terabytes. Big data are processes and procedures tools that allow an organization to create and manipulate, and also manage huge data sets and storage facilities.

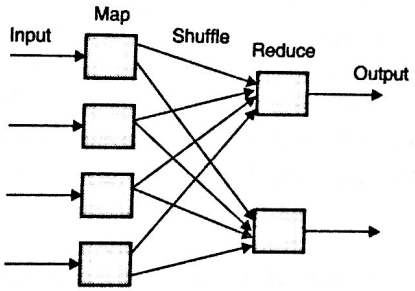

Fig. 2.3.1 : MapReduce

- MapReduce is a distributed programming framework which focuses on data placement and distribution. MapReduce can unlock the parallelism and make it easier to process large amount of data.
- MapReduce consist of two pieces of code that a user needs to write in order to use the framework: the Mapper and the Reducer.
- The MapReduce automatically launches many Mapper and Reducer tasks on a cluster of machines. The functions map and reduce are found in functional languages where the map function transforms a list of items into another list of items of the same length and the reduce function transforms a list of items into a single item.
- Following are the steps performed in MapReduce, which is also shown in Fig. 2.3.1.

1. Data comes in a set of records where records are sent to the Mapper, which transforms these records into pairs, each with a key and a value.
2. The next step is shuffle, which the library is performing by itself which uses a hash function so that all pairs with the same key end up next to each other and on the same machine.

3. Next step is reduce stage where the records are processed again, in batches, meaning all pairs with the same key are processed at once.

- The Mapper and Reducer are idempotent which means that if the Mapper or Reducer is called multiple times on the same input, the output will always be the same.
- Map Reduce uses one open source implementation i.e. Hadoop, which is a powerful tool designed for transformation of large data sets, deep analysis and explores complex data, which is being distributed across the cluster servers having high degree of fault tolerance and data availability.
- Hadoop Distributed File System (HDFS) are used by Hadoop which provides parallel processing, high throughput, and solve problem the problems where you have lots of data like,
  - Metadata such as which blocks are contained in which files is being provided by central "name node".
  - Due to block replication, if machine crashes data is neither lost nor unavailable.
  - File blocks are being partitioned across many machines.
- Hadoop can not only execute any program but also can dispense data across a cluster. It restricts the total number of node communication, which can be done by the unusual process, each and every individual record is processed by a task in isolated from one to another and a major limitation at initially.
- Hadoop avoids the expensive transmission when working with very bulky data sets and it is a flexible tool which allows new users to access the benefits of distributed computing, storage and transferring code instead of data.

Syllabus Topic : Evaluation

2.4 Evaluation

Q. 2.4.1 Write a short note on Evaluation. (Ref. Sec. 2.4) **(5 Marks)**

- To measure ad hoc information retrieval effectiveness three things are required,
  1. Document collection.
  2. Test suite of information needs which are expressible as queries.
  3. A set of relevance judgments, standardly a binary assessment of either relevant or nonrelevant for each query-document pair.
- As per need of user information, a document in the test collection is given a binary classification as either relevant or nonrelevant, which is referred to as the gold standard or ground truth judgment of relevance.

2.4.1 Standard Test Collections

Q. 2.4.2 Explain Standard Test Collections in Evaluation. (Ref. Sec. 2.4.1) **(5 Marks)**

Below are some test collections and evaluation series.

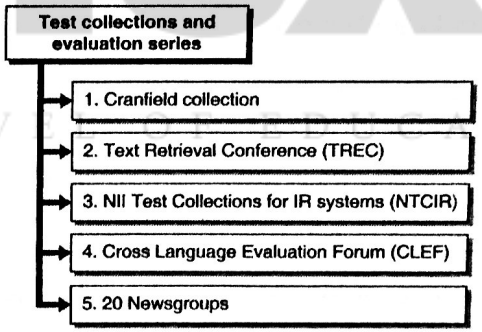

Fig. 2.4.1 : Test collections and evaluation series

➜ **1. Cranfield collection**

It is the pioneering test collection which allows precise quantitative measures of information retrieval effectiveness, but is nowadays too small for anything but the most elementary pilot experiments.

2. Text Retrieval Conference (TREC)

These test collections comprise 6 CDs containing 1.89 million documents and relevance judgments for 450 information needs, which are called topics and specified in detailed text passages.

3. NII Test Collections for IR systems (NTCIR)

It mainly focuses on East Asian language and cross-language information retrieval, where queries are made in one language over a document collection containing documents in one or more other languages.

4. Cross Language Evaluation Forum (CLEF)

It concentrates mainly on European languages and cross-language information retrieval.

5. 20 Newsgroups

It is collected by Ken Lang which consists of 1000 articles from each of 20 Usenet newsgroups.

2.4.2 Evaluation of Unranked Retrieval Sets

Q. 2.4.3 Explain Evaluation of unranked and ranked retrieval sets. **(5 Marks)**
(Ref. Secs. 2.4.2 and 2.4.3)

The two most frequent measures for information retrieval effectiveness are precision and recall.

- Precision (P) is the fraction of retrieved documents PRECISION that are relevant.

$$\text{Precision} = \frac{\#\text{ (relevant items retrieved)}}{\#\text{ (retrieved items)}} = P(\text{relevant} \mid \text{retrieved})$$

- Recall (R) is the fraction of relevant documents that are retrieved.

$$\text{Recall} = \frac{\#\text{ (relevant items retrieved)}}{\#\text{ (relevant items)}} = P(\text{retrieved} \mid \text{relevant})$$

Following is the contingency table :

Table 2.4.1

	Relevant	Nonrelevant
Retrieved	True positives (tp)	False positives (fp)
Non retrieved	False negatives (fn)	True negatives (tn)

Then :

$$P = tp/(tp+ fp)$$

$$R = tp/(tp+ fn)$$

A single measure that trades off precision versus recall is the F measure, which is the weighted harmonic mean of precision and recall.

2.4.3 Evaluation of Ranked Retrieval Sets

Q. 2.4.4 Explain Evaluation of unranked and ranked retrieval sets. **(5 Marks)**
(Ref. Secs. 2.4.2 and 2.4.3)

- Precision, recall, and the F measure are set-based measures which are computed using unordered sets of documents.
- For ranked retrieval context, appropriate sets of retrieved documents are naturally given by the top k retrieved documents and for each such set, precision and recall values can be plotted to give a precision-recall curve.
- The interpolated precision p_{interp} at a certain recall level r is defined as the highest precision found for any recall level $r' \geq r$:

$$p_{interp}(r) = \max p(r')$$

$$r' \geq r$$

Syllabus Topic : Personalized Search

2.5 Personalized Search

Q. 2.5.1 Explain Personalized Search in detail. (Ref. Sec. 2.5) **(5 Marks)**

Web search experiences that are tailored specifically to an individual's interests by incorporating information about the individual beyond specific query provided is known as personalized search. There are two approaches to personalizing search results,

1. Modify user's query
2. Re-ranking search results.

2.5.1 History

Generic web search engine cannot identify the different needs of different customers, so personalized concept was introduced. Personalized search was introduced by Google in 2004 and was implemented in 2005 to Google search. Google uses user language, location and web history for personalize searches. Search engines have two degrees of expertise,

1. Shallow Expert
2. Deep Expert

➜ **1. Shallow Expert :** It serves as a witness who knows some specific information on a given event.

➜ **2. Deep Expert :** It is the capacity to deliver unique information that is relevant to each individual inquirer.

2.5.2 Advantages

Q. 2.5.2 Explain the advantages, disadvantages and applications of Personalized Search. (Ref. Secs. 2.5.2, 2.5.3 and 2.5.4) **(5 Marks)**

1. To improve the quality of decisions consumers make.
2. The internet has made the transaction cost of obtaining information significantly lower than ever.
3. Improves consumers' decision quality and reduced the number of products inspected.

2.5.3 Disadvantages

Q. 2.5.3 Explain the advantages, disadvantages and applications of Personalized Search. (Ref. Secs. 2.5.2, 2.5.3 and 2.5.4) **(5 Marks)**

1. It limits the users' ability to become exposed to material that would be relevant to the user's search query.
2. Search personalization adds a bias to user's search queries.
3. Internet companies such as Google are gathering and potentially selling their users' internet interests and histories to other companies.

2.5.4 Applications of Personalized Search

Q. 2.5.4 Explain the advantages, disadvantages and applications of Personalized Search. (Ref. Secs. 2.5.2, 2.5.3 and 2.5.4) **(5 Marks)**

The applications are,

- Google Custom Search Engine
- Alpha Search Engine
- Google Web History Tool
- iGoogle and My Yahoo!

2.6 Recommender System

Q. 2.6.1 Write a short note on Recommender System. (Ref. Sec. 2.6) **(5 Marks)**

- A subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item is a recommendation system. Recommender

systems are used in different areas including movies, music, news, books, research articles, search queries etc.

- This systems produce a list of recommendations in two ways, collaborative filtering and through content-based filtering. This systems help users discover items they might not have found otherwise and is useful alternative to search algorithms. This system was first mentioned in a technical report as a "digital bookshelf" in 1990 by Jussi Karlgren at Columbia University.

Syllabus Topic : Collaborative Filtering

2.6.1 Collaborative Filtering

Q. 2.6.2 Distinguish between Collaborative and content based filtering. (Ref. Secs. 2.6.1 and 2.6.2) **(5 Marks)**

Q. 2.6.3 Explain Collaborative Filtering in detail. (Ref. Sec. 2.6.1) **(5 Marks)**

- It is based on collecting and analyzing a large amount of information on users behaviors, activities or preferences and predicting what users will like based on their similarity to other users.
- It does not rely on machine analyzable content and so it is capable of accurately recommending complex items such as movies without requiring an "understanding" of the item itself which is an advantage.
- The recommender system compares the collected data to similar and dissimilar data collected from others and calculates a list of recommended items for the user.
- Item-to-item collaborative filtering is an example of collaborative filtering. Facebook other social networks use collaborative filtering to recommend new friends, groups etc Collaborative filtering is categorized into memory-based and model based collaborative filtering.

Collaborative Filtering has three problems

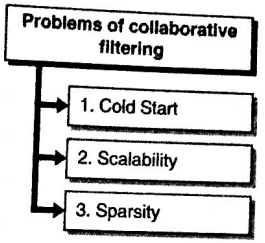

Fig. 2.6.1 : Problems of collaborative filtering

1. **Cold Start :** It requires a large amount of existing data on a user in order to make accurate recommendations.
2. **Scalability :** A large amount of computation power is often necessary to calculate recommendations.
3. **Sparsity :** The number of items sold on major e-commerce sites is extremely large.

Syllabus Topic : Content based Recommendation of Documents and Products

2.6.2 Content based Filtering

Q. 2.6.4 Distinguish between Collaborative and content based filtering. (Ref. Secs. 2.6.1 and 2.6.2) **(5 Marks)**

- It utilizes a series of discrete characteristics of an item in order to recommend additional items with similar properties. It is based on a description of the item and a profile of the user's preferences.
- Here, a user profile is built to indicate the type of item this user likes and keywords to describe the items. To create a user profile, the system mostly focuses on two types of information :
  1. A model of the user's preference.
  2. A history of the user's interaction with the recommender system.

- The system creates a content-based profile of users based on a weighted vector of item features. The issue is whether the system is able to learn user preferences from users actions regarding one content source and use them across other content types. For example, Pandora Radio is a Content based filtering which plays music with similar characteristics to that of a song provided by the user as an initial seed.

Syllabus Topic : Handling "Invisible" Web

2.7 Handling Invisible Web

Q. 2.7.1 Write a short note on Invisible Web. (Ref. Sec. 2.7) **(5 Marks)**

- Information which can be read with a web browser, but will not be found by a standard crawler for various reasons. On the Web there are innumerable information sources which contains vast amounts of authoritative and very useful information that cannot be indexed by general-purpose search engines and are hence invisible to most common users who do not know where these sources are, such information is known as Invisible Web. It is also called as Hidden Web or Deep Web. Some of features are listed below,
  - Static page are accessible only through web site search engine; no hyperlinks and no browsing.
  - Information is not suited to keyword search.
  - Internal page only meaningful in context of specific database.
  - Sites are not linked from anywhere.
  - Unsearchable or un-indexable file formats.
- Invisible Web sources incline to be more current, more comprehensive, Searchable, more specific/targeted, deeper breadth and often better quality. The Invisible Web is big, and it is getting bigger. The term "invisible" represents those resources that, because of their exclusion by general-purpose search engines, are not so easily found. A popular image

that shows the relationship between the visible and Invisible Web is one of a fishing trawler with its nets out in the middle of the ocean.

- The characteristics which make the Invisible Web Important are,
  - Size and Quality
  - Fastest Growing Part of the Web.

Syllabus Topic : Snippet Generation

2.8 Snippet Generation

Q. 2.8.1 Explain Snippet Generation. (Ref. Sec. 2.8) **(5 Marks)**

- Snippets is a short summary of documents that are designed to allow to decide its relevance. It contains document title and a short summary that is automatically extracted. There are two summaries,
  1. **Static :** This are always same regardless of the query.
  2. **Dynamic :** This are customized according to the user's information need as deduced from a query.
- A static summary is composed of a subset of the document and metadata associated with the document. The simplest form of summary takes the first two sentences or 50 words of a document, or extracts particular zones of a document, like title and author. There has been extensive work within Natural Language Processing (NLP) on better ways to do text summarization.
- Dynamic summaries display one or more "windows" on the document. This window consists of many query terms which are often called as keyword in context snippet. They are generated in concurrence with scoring. A dynamic summary cannot be precomputed, but, on the other hand, if a system has only a positional index, then it cannot easily reconstruct the context surrounding search engine hits in order to generate such a dynamic summary, hence for this reason we use static summaries.

Syllabus Topic : Summarization

2.9 Summarization

Q. 2.9.1 Explain the concept of Summarization in detail. (Ref. Sec. 2.9) **(5 Marks)**

Q. 2.9.2 Explain the approaches and techniques of Summarization in detail. (Ref. Sec. 2.9) **(5 Marks)**

Summarization means to find a subset of data which contains the "information" of the entire set. A text which is produced from one or more text and which consist important portion of information in the original text is called as summary. The schematic summary processing model consist of source text, interpretation, source representation, transformation, summary representation, generation and summary text. The factors of summarization are given below,

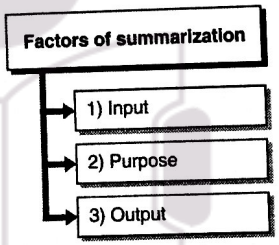

Fig. 2.9.1 : Factors of summarization

➜ **1. Input**

Subject type : domain

Form : regular text structure, free form

Source size : single and multiple doc

➜ **2. Purpose**

Situation : embedded in larger system

Usage : IR, sorting

➜ **3. Output**

Format : paragraph, table

Style : informative, indicative

Extraction and abstraction are two approaches for summarization.

☛ **Extraction based summarization**

It selects a subset of existing words, phrases or sentences in original text to form summary. Most of the summarization research is done here. Two types of task are there,

1. Generic summarization
2. Query-based summarization

➜ **1. Generic summarization :** obtains a generic summary or abstract of the collection.

➜ **2. Query-based summarization :** It summarizes objects specific to a query.

☛ **Abstractive based summarization**

It builds an internal semantic representation which then uses natural language generation techniques to create a summary. Due to problems like semantic representation, inference and natural language generation abstractive summarization has not reached a mature stage.

Summarization techniques can be classified into supervised or unsupervised.

1. Supervised
2. Unsupervised

➜ **1. Supervised :** It uses a collection of documents and human-generated summaries for them to train a classifier for the given text. Disadvantage is training data is expensive to produce and relatively sparse.

➜ **2. Unsupervised :** It models the document as a graph and uses an algorithm similar to Google's PageRank algorithm to find top-ranked sentences.

Syllabus Topic : Question Answering

2.10 Question Answering

Q. 2.10.1 Write a short note on Question Answering. (Ref. Sec. 2.10) **(5 Marks)**

Question answering builds systems that automatically answer questions posed by humans in a natural language. It is a computer program, which constructs its answers by querying a structured database of knowledge or information, usually a knowledge base. Examples of natural language document collections used for Question Answer systems consist of :

- A local collection of reference texts.
- Internal organization documents and web pages.
- Compiled newswire reports.
- A set of Wikipedia pages.
- A subset of World Wide Web pages.

It deals with fact, list, definition, How, Why, hypothetical, semantically constrained, and cross-lingual questions.

1. Closed-domain question answering
2. Open-domain question answering

➜ **1. Closed-domain question answering :** It deals with questions under a specific domain, Natural Language Processing systems can exploit domain-specific knowledge frequently formalized in ontologies.

➜ **2. Open-domain question answering :** It deals with questions about nearly anything, and can only rely on general ontologies and world knowledge.

Question Answer systems includes a question classifier module that determines the type of question and the type of answer. The idea of data redundancy in massive collections, such as

the web, means that nuggets of information are likely to be phrased in many different ways in differing contexts and documents which lead to two benefits:

1. By having the right information appear in many forms, the burden on the Question Answer system to perform complex NLP techniques to understand the text is lessened.
2. Correct answers can be filtered from false positives by relying on the correct answer to appear more times in the documents than instances of incorrect ones.

For example, systems have been developed to automatically answer temporal and geospatial questions, questions of definition and terminology, biographical questions, multilingual questions, and questions about the content of audio, images, and video.

Syllabus Topic : Cross-Lingual Retrieval

2.11 Cross-Lingual Retrieval

Q. 2.11.1 Write a note on Cross-Lingual Retrieval. (Ref. Sec. 2.11) **(5 Marks)**

Q. 2.11.2 Explain the challenges in Cross-Lingual Retrieval. (Ref. Sec. 2.11) **(5 Marks)**

- It is a subfield of information retrieval which deals with retrieving information written in a language different from the language of the user's query. It also has names like cross-language information retrieval, translingual information retrieval, multilingual information retrieval.
- Cross-lingual information retrieval refers more specifically to the use case where users formulate their information need in one language and the system retrieves relevant documents in another. For this, most CLIR systems use various translation techniques like,
  - Dictionary-based CLIR techniques.
  - Parallel corpora based CLIR techniques.
  - Comparable corpora based CLIR techniques.
  - Machine translator based CLIR techniques.

– Users with poor to moderate competence in the target language have benefited more from this cross lingual retrieval. Cross lingual information retrieval is important for countries like India where very large fraction of people are not conversant with English and thus don't have access to the large store of information on the web.

2.11.1 Challenges in Cross Lingual Retrieval

Following are some challenges in cross lingual retrieval,

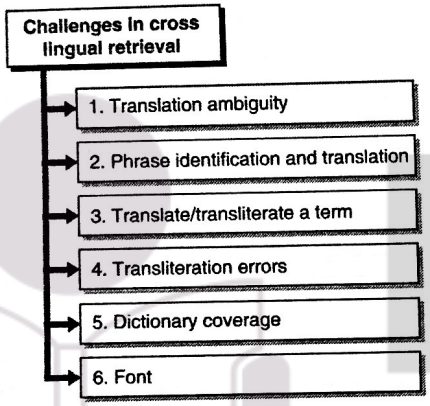

Fig. 2.11.1 : Challenges in cross lingual retrieval

➜ **1. Translation ambiguity**

While translating from source language to target language, more than one translation may be possible. Selecting appropriate translation is a challenge. For example, the word मान (maan, respect/neck) has two meanings neck and respect.

➜ **2. Phrase identification and translation**

Identifying phrases in limited context and translating them as a whole entity rather than individual word translation is difficult.

➜ **3. Translate/transliterate a term**

There are ambiguous names which need to be transliterated instead of translation. For example, भास्कर (Bhaskar, Sun) in Marathi refers to a person's name as well as sun.

➜ **4. Transliteration errors**

Errors while transliteration might end up fetching the wrong word in target language.

➜ **5. Dictionary coverage**

For translations using bi-lingual dictionary, the exhaustiveness of the dictionary is important criteria for performance on system.

➜ **6. Font**

Many documents on web are not in Unicode format. These documents need to be converted in Unicode format for further processing and storage.

2.12 Exam Pack (Review Questions)

☛ **Syllabus Topic : Link Analysis**

Q. 1 Define Link Analysis. (Refer Section 2.1) **(2 Marks)**

Q. 2 Write a note on Link Analysis. (Refer Section 2.1) **(5 Marks)**

Q. 3 Explain Anchor text and the web graph. (Refer Section 2.1.1.1) **(5 Marks)**

☛ **Syllabus Topic : Page Rank**

Q. 4 Write a short note on Page Rank. (Refer Section 2.1.2) **(5 Marks)**

Q. 5 Explain Markov Chains in detail. (Refer Section 2.1.2.1) **(5 Marks)**

☛ **Syllabus Topic : Hubs and Authorities, HITS Algorithms**

Q. 6 Write a note on Hubs and Authorities. (Refer Section 2.1.3) **(5 Marks)**

☛ **Syllabus Topic : Similarity**

Q. 7 Explain the concept of Similarity in detail. (Refer Section 2.2) **(5 Marks)**

Syllabus Topic : Hadoop and MapReduce

Q. 8 Write a note on Hadoop and MapReduce. (Refer Section 2.3) **(5 Marks)**

Q. 9 Explain MapReduce in detail with diagram. (Refer Section 2.3) **(5 Marks)**

Syllabus Topic : Evaluation

Q. 10 Write a short note on Evaluation. (Refer Section 2.4) **(5 Marks)**

Q. 11 Explain Standard Test Collections in Evaluation. (Refer Section 2.4.1) **(5 Marks)**

Q. 12 Explain Evaluation of unranked and ranked retrieval sets. (Ref. Secs. 2.4.2 and 2.4.3) **(5 Marks)**

Syllabus Topic : Personalized Search

Q. 13 Explain Personalized Search in detail. (Refer Section 2.5) **(5 Marks)**

Q. 14 Explain the advantages, disadvantages and applications of Personalized Search. (Refer Sections 2.5.2, 2.5.3 and 2.5.4) **(5 Marks)**

Q. 15 Write a short note on Recommender System. (Refer Section 2.6) **(5 Marks)**

Q. 16 Explain Collaborative Filtering in detail. (Refer Section 2.6.1) **(5 Marks)**

Syllabus Topic : Content based Recommendation of Documents and Products

Q. 17 Distinguish between Collaborative and content based filtering. (Refer Sections 2.6.1 and 2.6.2) **(5 Marks)**

Syllabus Topic : Handling "Invisible" Web

Q. 18 Write a short note on Invisible Web. (Refer Section 2.7) **(5 Marks)**

Syllabus Topic : Snippet Generation

Q. 19 Explain Snippet Generation. (Refer Section 2.8) **(5 Marks)**

Syllabus Topic : Summarization

Q. 20 Explain the concept of Summarization in detail. (Refer Section 2.9) **(5 Marks)**

Q. 21 Explain the approaches and techniques of Summarization in detail. (Refer Section 2.9) **(5 Marks)**

Syllabus Topic : Question Answering

Q. 22 Write a short note on Question Answering. (Refer Section 2.10) **(5 Marks)**

Syllabus Topic : Cross-Lingual Retrieval

Q. 23 Write a note on Cross-Lingual Retrieval. (Refer Section 2.11) **(5 Marks)**

Q. 24 Explain the challenges in Cross-Lingual Retrieval. (Refer Section 2.11) **(5 Marks)**

Chapter Ends...

❑❑❑