UNIT IV – WEB SEARCH – LINK ANALYSIS AND SPECIALIZED SEARCH

Part A - Questions

1. **What is link analysis?**

   The goal of information retrieval is to find all documents relevance for a user query in a collection of documents. With the advent of the web new source of information became available, one of them being the hyperlink between documents and records of user behavior. Collections of documents connected by hyperlinks. Hyperlinks provide a valuable source of information for web information retrieval. This area of information retrieval is commonly link analysis.

2. **What is in query independent ranking?**

   In query-independent ranking a score is assigned to each page without a specific user query with the goal of measuring the intrinsic quality of a page. At query time this score is used with or without some query-dependent criteria to rank all documents matching the query.

3. **What is query dependent ranking?**

   In query-dependent ranking a score measuring the quality and the relevance of a page to a given user query is assigned to some of the pages.

4. **Define authorities?**

   Authorities are pages that are recognized as providing significant, trustworthy and useful information on a topic. In-degree is one simple measure of authority. However in-degree treats all links as equal.

5. **Define hubs.**

   Hubs are index pages that provide lots of useful links to relevant content pages. Hub pages for IR are included in the home page.

6. **What is Hadoop?**

   At Goggle MapReduce operation are run on a special file system called Google File System that is highly optimized for this purpose. GFS is not open source. Doug Cutting and Yahoo! reverse engineered the GFS and called it Hadoop Distributed File System. The software framework that supports HDFS, MapReduce and other related entities is called the project Hadoop or simply Hadoop

7. What are the Hadoop Distributed File System?

The Hadoop Distributed File System is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user application. HDFS stores file system metadata and application data separately. The HDFS namespace is a hierarchy of files and directories. Files and directories are represented on the NameNode by inodes, Which record attributes like permissions, modification and access times, namespace and disk space quatas.

8. Define MapReduce.

MapReduce is a programming model and software framework first developed by Google. Intended to facilitate and simplify the processing of vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner.

9. List the characteristics of MapReduce?

- Very large scale data: peta, exa bytes
- Write once and read many data. It allows for parallelism without mutexes
- Map and Reduce are the main operations: Simple code
- All the map should be completed before reduce operation starts.
- Map and reduce operations are typically performed by the same physical processor.
- Number of map tasks and reduce tasks are configurable.
- Operations are provisioned near the data.
- Commodity hardware and storage.

10. What are the limitation of Hadoop/MapReduce?

- Cannot control the order in which the maps or reductions are run.
- For maximum parallelism, you need Maps and Reduces to not depend on data generated in the same MapReduce job.
- A database with an index will always be faster than a MapReduce job on unindexed data.
- Reduce operations do not take place until all Maps are complete.
- General assumption that the output of Reduce is smaller than the input to Map;large data source used to generate smaller final values.

11. What is Cross-Lingual Retrieval?

Cross – Lingual Retrieval refers to the retrieval of documents that are in a language different from the one in which the query is expressed. This allows users to search document collections in multiple language and retrieve relevant information in a form that is useful to them, even when they have little or no linguistic competence in the target languages.

12. Define Snippets.

Snippets are short fragments of text extracted from the document content or its metadata. They may be static or query based. In static snippet, it always shows the first 50 words of the document, or the content of its description metadata, or a description taken from a directory site such as dmoz.org.

13. List advantages of invisible web content.

- Specialized content focus – large amounts of information focused on an exact subject.
- Contains information than might not be available on the visible web.
- Allows a user to find a precise answer to a specific question
- Allow a user to find WebPages from a specific date or time.

14. What is collaborative filtering?

Collaborative filtering is a method of making automatic predictions about the interests of a single user by collecting preferences or taste information from many users. It uses given rating data by many users for many items as the basic for predicting missing ratings and/or for creating a top-N recommendation list for a given user, called the active user.

15. What do you mean by item-based collaborative filtering?

Item-based CF is a model-based approach which produces recommendations based on the relationship between items inferred from the rating matrix. The assumption behind this approach is that users will prefer items that are similar to other items they like.

16. What are problem of user based CF?

The two main problems of user-based CF are that the whole user database has to be kept in memory and that expensive similarity computation between the active user and all other users in the database has to be performed.

17. Define user based collaborative Filtering.

User-based collaborative filtering algorithms work off the premise that if a user(A) has a similar profile to another user (B), then A is more likely to prefer things that B prefers when compared with a user chosen at random.

Part- B

1. Explain Link Analysis

Many links are navigational.
Many pages with high in-degree are portals not content providers.
Not all links are endorsements.
Company websites don't point to their competitors.
Citations to relevant literature is enforced by peer-review.

2. Define Hub and authorities

- *Hubs* are index pages that provide lots of useful links to relevant content pages (topic authorities).

HITS Algorithm
- Computes hubs and authorities for a particular topic specified by a normal query.
- First determines a set of relevant pages for the query called the *base* set *S*.

Analyze the link structure of the web subgraph defined by *S* to find authority and hub pages in this set

Authorities and In-Degree
- Even within the base set *S* for a given query, the nodes with highest in-degree are not necessarily authorities (may just be generally popular pages like Yahoo or Amazon).
- True authority pages are pointed to by a number of hubs (i.e. pages that point to lots of authorities).

3. Explain Collaborative filtering and content

Collaborative filtering (**CF**) is a technique used by recommender systems.[1] Collaborative filtering has two senses, a narrow one and a more general one.

Types

Memory-based

Model-based

Hybrid

Context-aware collaborative filtering

4. Discuss about Cross Lingual Retrieval

Crawling: Documents from web are fetched and stored. Indexing: An index of the fetched documents is created. **Cross-Lingual Information Retrieval** (CLIR) refers to the **retrieval** of documents that are in a language different from the one in which the query is expressed.