**Course: Advanced Bioinformatics**

**Module title: Information Retrieval**

**Module no. : 167**

For any sort of data, it can be explained at following four levels

* Representation
* Storage
* Organization
* Information access

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing.

Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

**Overview**

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

An object is an entity that is represented by information in a database. User queries are matched against the database information. Depending on the application the data objects may be, for example, text documents, images, audio, mind maps or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

**Types of Information Needs**

Retrospective: Searching the past; Different queries posed against a static collection, Time invariant

Prospective: Searching the future, Static query posed against dynamic collection, Time dependent

Pooled resources: shared with many users (remotely accessed)

Virtualization: high utilization of hardware resources (no idling)

Elasticity: dynamic scaling without capital expenditure and time delay

Automation: build, deploy, configure, provision, and move without manual intervention

**Sources of Information**

* Text (Documents and portions thereof)
* XML and structured documents
* Images/Video
* Audio: sound effects, songs Source code