UNIT 5 EVALUATION OF INDEXING SYSTEMS

Structure

5.0 OBJECTIVES

In the previous Units, you have learnt the fundamentals of information processing and organisation. You have also been acquainted with different types of indexing systems developed for organisation of bibliographic information in the context of development of information storage and retrieval systems. Different systems have originated to suit different types of requirements. For having an efficient system to suit particular needs it is necessary the system developed is evaluated to judge its efficiency in retrieval. The present Unit discusses various aspects of evaluation. Indexing is a sub-system of the overall ISAR system. Though the Unit is discussed in the context of IR as a whole system, the methodology is equally applicable in the context of indexing systems.

After reading this Unit, you will be able to:

- analyse the reasons for evaluation of IR systems;
- understand different levels and criteria for evaluating an IR system;
- learn the methodology for evaluating an IR system;
- develop strategies in evaluating an IR system; and
- know the different evaluation experiments.

5.1 INTRODUCTION

Evaluation of an information retrieval system essentially means measuring the performance of the system, success or failure, in terms of its retrieval efficiency (ease of approach, speed and accuracy), and its internal operating efficiency, cost effectiveness and cost benefit to the managers of the system. In other words, we evaluate a system in order to ascertain the level of its performance or its value. An indexing system is a sub-system of an information retrieval system and hence, its performance is directly linked with its overall performance of the entire information retrieval system. An information retrieval system can be evaluated by considering the following two issues:

i) how well efficiently the system is satisfying its objectives, that is, how well it is satisfying the demands placed upon it, and

ii) whether the system justifies its existence.

We evaluate a system in order to ascertain the level of it's performance or value in terms of its *effectiveness* and *efficiency*. By effectiveness we mean the level up to which the given system attains its stated objectives. In an information retrieval system, the effectiveness may be a measure of how far it can retrieve relevant information withholding non-relevant information. By efficiency we mean how economically the system is achieving its objective. In an information retrieval system efficiency can be measured by such factors, such as at what minimum cost and effort does the system function effectively. It may be necessary that the cost factors are to be calculated indirectly, such as response time (i.e. that is time taken by the system to retrieve the information), user effort (i.e. the amount of time and effort required by a user to interact with the system and analyse the output retrieved in order to get the required information), the cost involved, and so on. Evaluation is generally done for computer-based systems. It can also be possible in case of manual system (Card catalogues) and other types of bibliographical products and services, including printed indexes.

5.2 PURPOSE OF EVALUATION

As it has been mentioned earlier, the purpose of evaluation in general is to judge the efficiency of a system in relation to retrieval, identify the shortcomings, if any, rectify them and improve upon the system.

An IR system is evaluated to obtain two types of data:

a) Performance figures for a representative group of searches; and

b) Examples of system failures to allow analysis of causes of failures in searches.

Evaluation studies investigate the degree to which the stated goals or expectations have been achieved or the degree to which these can be achieved. The above-mentioned data, when analysed, and interpreted, should yield recommendations from which decisions can be made on how best the system performance may be improved. Evaluation studies have one or more of the following purposes:

- To show at what level of performance the system is now operating;
- To compare the performance of two or more systems against a standard or norm;
- To determine whether and how well goals or performance expectations are being fulfilled;
- To identify the possible sources of system failure or inefficiency with a view to raising the level of performance at some future date;

- To justify the system's existence by analysing the costs and benefits;
- To explore techniques for increasing performance effectiveness;
- To establish a foundation of further research on the reasons for the relative success of alternative techniques; and
- To improve the means employed for attaining the objectives or to redefine the goals in view of research findings.

5.3 LEVELS OF EVALUATION

An IR system may be evaluated at various levels. F.W. Lancaster has identified the following levels of evaluations:

5.3.1 System Effectiveness

It is the evaluation of the system performance in terms of the degree to which it meets the users' requirements. It considers the users' satisfaction level and is measured by determining the utility of information to the users in response to a user query. Precision has been widely used to measure the retrieval effectiveness. It takes into consideration cost, time and quality criteria.

1) Cost Criteria:
   i) Monetary cost of user (i.e. cost incurred per search, per subscription, per document);
   ii) Other less tangible cost considerations—such as, Users' effort involved :
      - in learning the working of the system;
      - in actual use;
      - in getting the documents through back-up document delivery systems; and
      - in retrieving information from the retrieved documents.
2) Time Criteria:
   i) Time taken from submission of query to the retrieval of bibliographical references;
   ii) Time elapsing from submission of query to the retrieval of documents and the actual information; and
   iii) Other time considerations—such as, waiting time to use the system such as online.
3) Quality Criteria :
   i) Coverage of database;
   ii) Completeness of output (Recall);
   iii) Relevance of output (Precision);
   iv) Novelty of output; and
   v) Completeness and accuracy of data.

5.3.2 Cost Effectiveness

There may be many systems where system effectiveness may be there but cost effectiveness may not be satisfactory. A cost effectiveness evaluation relates measures of effectiveness to measure of cost. It is the evaluation in terms of how to satisfy user requirements in the most efficient and economical way. It takes into considerations:

a) Unit cost per relevant citation retrieval;

b) Unit cost per new, that is, previously unknown, relevant citation retrieval; and

c) Unit cost per relevant document retrieval.

5.3.3 Cost-benefit Evaluation

It is the evaluation to assess the worthiness of the system. A cost-benefit study attempts to relate the cost of providing some service to the benefits of having the service available. A number of studies have been conducted so far to determine the costs of information retrieval systems or subsystems.

Self Check Exercises

1) How do you distinguish between the efficiency and effectiveness of an IR system?

2) What are the different levels of evaluation?

Note: i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

5.4 EVALUATION CRITERIA

During 1950s Perry and Kent while bring out the concept of evaluation for IR systems suggested the following:

i) **Resolution factor**: The proportion of total items retrieved over a total number of items in the collection.

ii) **Pertinency factor**: The proportion of relevant items retrieved over a total number of retrieved items. This factor is popularly named as the *precision ratio* in the subsequent evaluation studies.

iii) **Recall factor**: The proportion of relevant items retrieved over a total number of relevant items in the collection.

iv) **Elimination factor**: The proportion of non-retrieved items (both relevant and non-relevant) over the total items in the collection.

v) **Noise factor**: The proportion of retrieved items which are not relevant. This factor is considered as the complement of the *pertinency factor.*

vi) **Omission factor**: The proportion of non-relevant items retrieved over the total number of non-retrieved items in the collection.

Perry and Kent suggested the following formulae for the estimation of the above mentioned evaluation criteria:

L / N	=	Resolution factor	(N—L) / N	=	Elimination factor
R / L	=	Pertinency factor	(L—R) / L	=	Noise factor
R / C	=	Recall factor	(C—R) / C	=	Omission factor

Where, N = Total number of documents

L = Number of retrieved documents

C = Number of relevant documents

R = Number of documents that are both retrieved and relevant

C.W. Cleverdon [1962] identified six criteria for the evaluation of an information retrieval system. These are:

i) **Recall**: It refers to the ability of the system to retrieve all the relevant items;

ii) **Precision**: It refers to the ability of the system to retrieve only those items that are relevant;

iii) **Time lag**: It refers to the time gap between the submission of a request by the user and his receipt of the search results.

iv) **User Effort**: It refers to the intellectual as well as physical effort required from the user in obtaining answers to the search requests. The effort is measured by the amount of time user spends in conducting the search or negotiating his enquiry with the system. Sometimes, response time may be good, but user effort may be poor.

v) **Form of presentation** of the search output, which affects the user's ability to make use of the retrieved items, and

vi) **Coverage of the collection**: It refers to the extent to which the system includes relevant matter. It is a measure of the completeness of the collection.

Vickery identifies six criteria under two sets as follows:

Set 1

i) *Coverage*: the proportion of the total potentially useful literature that has been analysed,

ii) *Recall*: the proportion of such references that are retrieved in a search, and

iii) *Response time*: the average time needed to obtain a response from the system.

These three criteria are related to the availability of information, while the following three are related to the selectivity of output:

Set 2

i) *Precision*: the ability of the system to screen out irrelevant references,

ii) *Usability*: the value of the references retrieved, in terms of such factors as their reliability, comprehensibility, currency, etc., and

iii) *Presentation*: the form in which search results are presented to the user.

All these factors are related to the system parameters, and thus in order to identify the role played by each of the performance criteria mentioned above, each must be tagged with one or more system parameters. Salton and McGill identified the various parameters of an information retrieval system as related to each of five evaluation criteria. Table 5.1 presents the relationships between these evaluation criteria and the system parameters as identified by Salton and McGill.

Table 5.1 : Relationship between Evaluation Criteria and System Parameters

Evaluation Criteria	System Parameters
Recall and precision	1) Indexing exhaustivity Recall tends to increase the exhaustivity of indexing terms. 2) Term specificity Precision increases with the specificity of the index terms. 3) Indexing language Availability of measures for recognition of synonyms, term relations, etc., which improve recall. 4) Query formulation Ability to formulate an accurate search request. 5) Search strategy Ability of the user or intermediary to formulate an adequate search strategy.
Response time	1) Organisation of stored documents. 2) Type of query. 3) Location of information centre. 4) Frequency of receiving users' queries. 5) Size of the collection.
User effort	1) Accessibility of the system. 2) Availability of guidance by system personnel. 3) Volume of retrieved items. 4) Facilities for interaction with the system.
Form of presentation	1) Type of display device. 2) Nature of output – bibliographic reference, abstract, or full text.
Collection coverage	1) Type of input device and type of size of storage device. 2) Depth of subject analysis. 3) Nature of users' demands. 4) Nature of core subject area. 5) Physical forms of documents.

Some of the performance criteria mentioned above can be measured easily. For example, the parameters related to the collection coverage, and forms of presentation are related to policy matters, and thus are defined by the system managers beforehand. Response time and user effort can be measured without much difficulty. However, the two other criteria, recall and precision, cannot be measured so easily. In fact, measurements of these factors often cause a number of problems for the system investigators. Much research effort in the area of the evaluation of information retrieval system has concentrated on these two factors.

5.4.1 Recall and Precision

The most important parameters used for evaluating an indexing system are: *Recall* and *Precision.* The term *recall* refers to a measure of whether or not a particular item is retrieved or the extent to which the retrieval of wanted items occur. *Recall ratio* is nothing but the proportion of relevant items retrieved and thus, it is a measure of the completeness of a search in an index file. The *recall ratio* has been variously called as *hit rate, sensitivity*, and *conditional probability of a hit.* The general formula for calculation of *recall ratio* may be stated as:

$$\text{Recall Ratio} \quad \frac{\text{Number of relev}}{\text{Total number of rel}}$$

Example: If a database holds 50 relevant documents as answer to a query and retrieves only 35 of them against the enquiry as output, the *recall ratio* is (35 / 50) × 100 = 70%.

The term *precision* relates to the ability of an indexing system not to retrieve irrelevant items. *Precision ratio* is nothing but the proportion of retrieved items that are relevant. By it, we measure how precisely an indexing system functions. It is quite obvious that when the system retrieves items that are relevant to a given query it also retrieves some documents that are not relevant. These non-relevant items affect the success of the system because the user, for whom it results in the wastage of significant amount of time, must discard them. The *precision ratio* is sometimes referred to as a *relevance ratio.* The general formula for calculation of precision ratio may be stated as:

$$\text{Precision Ratio} \quad \frac{\text{Number of relev}}{\text{Total number}}$$

Example: If a system retrieves 50 documents as answer to a query as output and only 8 of them are relevant to that particular query, the *precision ratio* is (8 / 50) × 100 = 16%.

In the context of the evaluation, the following matrix as a common frame of reference would be useful:

	User relevance decision		
Retrieved	a Hits	b Noise	c Total retrieved
Not retrieved	c Misses	d Correctly rejected	c+d Total not retrieved
Total	a + c Total relevant	b + d Total not relevant	a + b + c + d Total collection

It appears from the above noted matrix that the system retrieves *'a'* relevant documents (i.e. *'hit'*) along with '*b'* non-relevant documents (i.e. 'noise'), and out of the remaining [*c+d*] documents, the system misses '*c'* documents that should have been retrieved, but it correctly rejects *'d'* documents that are not relevant to the given query. From this, the recall and precision ratios can be calculated as:

$$\text{Recall ratio} = [a / (a+c)] \times 100$$

$$\text{Precision ratio} = [a / (a+b)] \times 100$$

During the evaluation programme, values (a + b) can be established easily, but it is difficult to determine values (c + d). This can be done by asking the enquirer to verify all the non-retrieved documents (c + d) and judge which of them are relevant (c) and which are not (d).

An ideal indexing system attempts to achieve 100% recall and 100% precision, i.e. it attempts to retrieve all the relevant documents only. However, this is not possible in practice because as the level of recall increases, precision tends to decrease. Lancaster therefore states that *recall and precision tend to vary inversely in searching*. By this we mean that when we broaden a search to

achieve better recall, precision tends to go down. Conversely, when we restrict the scope of a search by searching more stringently in order to improve the precision, recall tends to deteriorate. In fact, by conducting a search or whole group of searches at varying strategy levels, from very broad to very specific, a series of performance points can be obtained in terms of recall and precision. If recall and precision ratios are derived for each of these search approaches and if these ratios were plotted against each other, the plot would result in *performance curve,* as shown below:

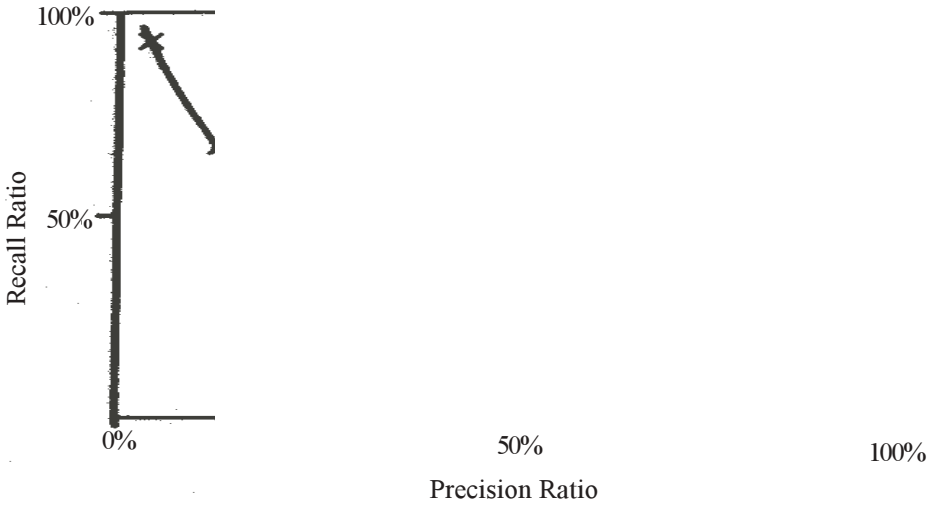


Table 5.1 : Performance curve

Fugmann has questioned the view of Lancaster in respect of recall—precision inverse relationship and by several examples he has shown that:

a) an increase in precision is by no means always accompanied by a corresponding decrease in recall, and

b) an increase in recall is by no means observed to have always in its wake a decrease in precision.

5.4.1.1 Factors Affecting Recall and Precision

Two performance criteria, i.e. recall and precision are influenced by the following factors:

a) Indexing policy: exhaustivity and specificity;

b) Requests that imperfectly represent information needs; and

c) Search strategy.

The effectiveness of an indexing system is governed by the indexing policy—*exhaustivity* and *specificity* (See the Section 4.2.7 under the Unit 4 of this course), which affect two most important parameters—*Recall* and *Precision* used for evaluating an indexing system.

A high level of *exhaustivity* of indexing tends to ensure high *recall.* For example, if we do not recognise the concept E in indexing a document dealing with five concepts A, B, C, D, E, we will never be able to retrieve it in response to a request for literature on E unless the concept E is related, hierarchically or non-hierarchically, to the concepts A to D. A high level of *exhaustivity* of indexing also tends to reduce *precision.* If we recognise all or substantial proportion of the indexable concepts for every document input, we will tend to index many concepts that are treated in only a very minor way in the documents concerned and as a result, the system will retrieve many documents containing little required

information. Further, If we use more concepts representing more terms in indexing, the greater is the potentiality of false coordination in searching. Thus, in a document containing four concepts ABCD, in which A and B are related, C and D are related, there may be the possibility of retrieving documents in which A and C are related, B and D are related, and so on. Conversely, a low level of *exhaustivity* of indexing leads to *low recall* and *high precision.* Suppose, we index each input document under a central or core topic only after the adoption of the policy of indexing at the minimum level of *exhaustivity.* Obviously, the system will not retrieve the document(s) in response to a request for a topic other than the core topic and thus, will eliminate the possibility of false coordination.

On the question of *specificity*, it is to be noted that the specificity of the vocabulary is the single most important factor influencing the *precision* with which a search can be conducted. For example, we want to retrieve documents on 'POPSI' from a retrieval system. If the system includes only its broader term like 'pre-coordinate indexing', or the still broader 'indexing', we shall retrieve a large number of documents, many of which will not discuss the required topic. In such a situation, we will achieve *high recall* at the cost of *low precision.* However, we can decrease the level of *recall* by increasing the level of *specificity.* For example, if we select 'pre-coordinate indexing' we can retrieve many documents where POPSI has been discussed along with other pre-coordinate indexing techniques. But, if the system includes a more specific term 'POPSI', the number of retrieved items will come down to our specific requirements and thus, can achieve *high precision.*

An analysis of recall and precision failures is undertaken after establishing the recall and precision ratios. This *failure analysis* implies diagnosis and is intended to lead to therapeutic action. It is the most important part of the entire evaluation programme, and involves a careful examination of the documents involved, how they were indexed, the request posed to the system, the search strategy used, and the complete assessment forms of the requesters. *Failure analysis* will attribute the recall and precision failures encountered to the principal subsystems like indexing, searching, index language, and user-system interface. The most important of the possible sources of failures as identified by Lancaster in the operation of MEDLARS are illustrated below:

Table 5.2 : Sources of Failures

Factors	Recall failures	Precision failures
Index language	Lack of specific terms (entry vocabulary), Inadequate hierarchical cross-reference structure, Roles, or other relational indicators causing over preciseness.	Lack of specific terms (index terms), Defects in hierarchy, False coordination Incorrect term relationships.
Indexing	Lack of specificity, Lack of exhaustivity, Omission of important concepts, Use of inappropriate terms.	Exhaustive indexing, Use of inappropriate terms.
Searching	Failure to cover all reasonable approaches to retrieval, Formulation too exhaustive, Formulation too specific.	Formulation not sufficiently exhaustive, Formulation not sufficiently specific, Use of inappropriate terms or term combinations.
User-System	Requests more specific than actual information needs.	Requests more general than actual information needs Computer processing,
Others	Computer processing, Clerical	Clerical, Value judgment, Inevitable retrievals.

5.4.1.2 Recall and Precision Devices

Devices used to manipulate indexing files to obtain optimum recall and precision have been categorised as follows:

Table 5.3 : Devices for Optimization of Recall and Precision

Recall–oriented	Precision–oriented
Synonym control	Co-ordination
Word form control	Links
Classification, including	Roles
a) Hierarchies	Weighting
b) Lattices	Relational indexing
c) Facet analysis	
d) Semantic factoring	
e) Clumps and clusters	

5.4.2 Other Performance Measures

Apart from the recall and precision ratios, various other measures of retrieval performance have been used and suggested. We have already discussed the performance measures as suggested by Perry and Kent in the Section 5.4 of this Unit. The following measures can also be derived from the matrix representing a common frame of reference for evaluation as mentioned therein:

i) **Snobbery ratio:** It refers to the number of relevant items that are not retrieved over the total number of relevant items in the collection. It is sometimes referred to as the conditional probability of a 'miss' and is measured by c / (a+c). It is complementary of recall ratio.

ii) **Noise factor:** The proportion of retrieved items which are not relevant. It is complementary of the precision ratio and is measured by b / (a+b).

iii) **Fallout ratio:** It refers to the number of non-relevant items that are retrieved over the total number of non-relevant items in the collection. It is sometimes referred to as 'discard' or the conditional probability of a 'false drops' and is measured by b / (b+d). It measures the proportion of the non-relevant items in the file, which are retrieved by a request. The fall out thus represents the fractional recall of irrelevant items. The fallout ratio taken together with the recall ratio may provide important clues to the underlying causes of retrieval performance.

iv) **Specificity ratio:** It refers to the conditional probability of a correct rejection and is measured by d / (b+d). It is complementary to the fall out ratio.

v) **Generality number:** It expresses the number of items relevant to a particular request over the total number of items in the collection. The higher the generality number the greater the density of relevant items to total collection, and the greater this density the easier the search tends to be.

vi) **Novelty ratio:** The proportion of relevant documents retrieved in a search that are new to the requester, that is, brought to his attention for the first time by the search. This ratio is particularly appropriate in the evaluation of literature searches conducted for current awareness purposes. The *novelty ratio* can be expressed in one of two ways:

$$\frac{\text{No. of new relevant documents retrieved}}{\text{No. of relevant document retrieved}} \quad \text{OR} \quad \frac{\text{No. of new relevant documents retrieved}}{\text{No. of documents retrieved}}$$

5.4.3 Relevance

Relevance is considered as one of the important criteria underlying existing

performance measures and is most highly debated in IR research. The term *relevance* is very difficult to define. This is because there are degrees of relevance. The individual view of *relevance* led to the concept of *pertinence* or *utility*. The term *pertinence* refers to a relationship between a document and an information need, whereas the term *relevance* refers to a relationship between a document and a request statement (i.e. expressed information need). The concept of relevance needs to be viewed in the broader context of a person needing information and expressing it in the form of a request (i.e. request statement). If a document is retrieved in response to a particular request, a panel of judges may assess its relevance, but the requester can only assess its pertinence. Relevance is consensus judgment, pertinence relates to an individual judgment. Another way of looking at the matter is that a document retrieved in response to a request may be useful to the enquirer, but its utility may change. For example: If the same document is retrieved in the second search, it will have lost its utility to the enquirer. Though its relevance will not have changed, but the enquirer view of it will change. Many retrieved relevant documents may be repetitive in the second search and they do not add anything significant to our knowledge.

The degree to which a requester is able to recognise the exact nature of his information need and the degree to which his need is accurately expressed determine how successful the IR system is able to satisfy the user. In relation to most of the evaluation studies *relevance* was applied to stated requests (i.e. expressed need). But, it has now been well established that the users' requests do not reflect their information needs completely. Therefore, the current view is that the *relevance* is to be judged in relation to both expressed and unexpressed needs rather than restricting only to stated requests. But whether this would be possible is a remains the question.

The term *precision* is used very widely in the literature in preference to relevance, and the term *relevance* is used when a subjective judgment is involved. *Relevance* is used to refer to the real-life situation, *precision* is to refer to the experimental situation.

Relevance of a document to a particular request cannot be measured precisely: the relationship is subjective and equivocal rather than objective and unequivocal. Different users may make different decisions on the *degree of relevance* between a document and a request. It is also quite possible that the same user may make different decisions on *relevance* in respect of a particular request-document pair at different times. *Recall and precision ratios* measure the degree of coincidence between the user relevance assessments and the *system relevance predictions*. In a perfect search these exactly coincide. Unfortunately, such perfect searches are relatively rare. Different users have different requirements for *recall* and *precision*, and a particular individual has different requirements at different times.

The idea of measuring *relevance* spread to the Cranfield project (See Section 5.6.1 of this Unit). However, the concept of *relevance* as viewed at Cranfield was different from that of other experts. In 1967, Cuadra and Katter published a paper entitled *Opening the black box of relevance* wherein they reported that it is possible to measure relevance by obtaining higher or lower relevance scores simply by telling relevance judges how documents are to be used. For example:

a) Use in stimulating ideas, creative approaches, etc.;

b) Use in relation to a specific task; and

c) Use in preparing an exhaustive bibliography.

Lancaster opined that the *relevance* is relative and capable of being judged on some type of scale in dividing a set of documents on the basis of extent relevance [1979]:

a) clearly relevant to a particular request statement,

b) relevant to the request statement but less relevant than the first document, and

c) not relevant.

We may get disagreement among a group of judges as to which documents are relevant to a particular request statement. There does not seem to be any consensus among the judges on the question of *relevance of relevance.*

Self Check Exercises

3) What are the different criteria used in evaluating an IR system?

4) How recall and precision ratios help in measuring the performance of an indexing system?

Note: i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

.....

5.5 EVALUATION METHODOLOGY

An evaluation study is conducted to determine the level of performance of the given system and also to identify those factors that are the reasons for weaknesses of the system. In other words, in evaluation an attempt is made to find out the different parameters and their interrelations with a view to assessing their contribution towards the overall performance of the system. Lancaster identifies five major steps involved in the evaluation of an information retrieval system, which are discussed below:

a) **Defining the scope of evaluation:** The first step of an evaluation study involves the designing of a precise set of questions in conformity with the evaluation objectives in order to learn capabilities and weakness of the system. The questions are usually concerned with one or more of the following:

   i) Overall performance level of the system;

   ii) Coverage and processing;

   iii) Indexing;

   iv) Index language;

   v) Searching; and

   vi) Input procedure and computer processing.

   The purpose and scope of the whole evaluation programme are set at this step. How the evaluation study will be conducted – in a laboratory-type set-up or in a real-life situations; and the probable constraints – in terms of cost, staff, time, etc., are also considered at this stage. In fact, a detailed plan is chalked out at this stage that forms the basis for the rest of the programme.

b) **Designing the evaluation programme**: This step involves the preparation of a detailed plan of action concerning the identification of parameters and procedure for collection of data needed to answer the question(s) set in the definition of scope. The designer might need to control some of the parameters of the system while conducting an evaluation programme. The designer should point out which parameters are to be held constant during the study and how this is to be done. Data may be collected either from the existing system for an answer to the question like "What is the present response time of the system (expressed in ranges, means, median and mode)?" or by making some changes in the normal functioning of the system for an answer to the question like "What would be the effect of response time if X is applied?" Here question is answered by applying X to a representative sample of transactions and comparing the response times with the existing system.

c) **Execution of the evaluation:** It is the stage in which data (performance results like recall and precision ratios, etc.) are continuously derived in a way prescribed in the design stage. In most cases, a repeated number of observations are required to avoid sampling error and bias. Derived data are manipulated and reduced to a form suitable for interpretation and analysis so that it can answer or contribute to answering the questions set in the definition of scope. The execution of the evaluation is obviously the most time-consuming step in an evaluation study.

d) **Analysis and interpretation of results:** The success of an evaluation programme rests upon the method of interpretation of results and its accuracy. This stage should begin before the execution stage is completed. Performance results or data collected on different parameters during the execution of the evaluation programme are analysed and interpreted in this stage. Although the methodology for manipulation of the data is determined at the design stage, the evaluator might need to make some changes so as to arrive at a better conclusion. Once the data have been manipulated suitably, the evaluator gets a set of results that is to be interpreted in the light of the set of objectives. No precise guidelines for analysis and interpretation are available since they vary considerably from one evaluation application to another. Appropriate statistical techniques are applied to the analysis and interpretation of the performance results. The evaluator might need to conduct failure analysis so as to justify the results and also to suggest improvements. Lancaster mentions that the joint use of performance figures and failure analysis should answer most of the questions identified in the objectives of the evaluation.

e) **Modifying the system in the light of the evaluation results:** Finally, the retrieval system is modified, if necessary, in the light of the results of the evaluation study.

Self Check Exercise

5) What are the major steps involved in the evaluation of the performance of an IR system?

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

5.6 EVALUATION EXPERIMENTS

5.6.1 The Cranfield Tests

5.6.1.1 Cranfield 1

The first extensive evaluation of retrieval systems was carried out by the ASLIB with the financial assistance from the National Science Foundation at the College of Aeronautics, Cranfield, UK, under the supervision of C.W. Cleverdon in 1957. It is popularly known as the Cranfield 1 Project. Its objective was to investigate the comparative efficiency of four indexing systems: UDC, Faceted Classification, Alphabetical Subject Headings List and Uniterm Indexing system.

a) **Input Elements**

A) Four indexing systems were taken into consideration:

1) An alphabetical subject catalogue based on a subject headings list and a set of rules for construction of heading;

2) A classified catalogue based on UDC with an alphabetical chain index to the class headings constructed;

3) A catalogue based on a faceted classification (Colon Classification) and an alphabetical chain index to the class headings constructed; and

4) A uniterm coordinate index controlled by an authority list of uniterms compiled during indexing.

B) Indexers: Three indexers with varied backgrounds in proficiency in indexing were taken:

1) One indexer with experience of indexing and subject knowledge of Aeronautics;

2) One with experience of indexing but with little subject knowledge; and

3) One indexer straight from library school (fresh student) with neither indexing experience nor subject knowledge.

Each indexer was asked to prepare index entries by following the above-mentioned four indexing systems.

C) *Time-period used:* Documents were indexed 5 times using variable indexing times (2, 4, 8, 12, 16 minutes). The system was run through three phases with a view to find out whether the level of performance increases with the increasing experience of the system personnel.

D) *Materials Indexed:* 100 technical reports and articles on the general field of aeronautics and the specialized field of high-speed aerodynamics, half of which were research reports and half periodical articles.

E) *Items Indexed:* With 3 indexers X 4 indexing systems X 5 time periods X 3 runs = 18,000 items were indexed.

F) *Number of Descriptors:* The study involved the use of following types of descriptors:

1) *Alphabetical Subject Heading List:*	2864 Main headings 592 Sub-headings 1560 'See' references
2) *Universal Decimal Classification:*	2350 Class numbers 4052 Headings in alphabetical order
3) *Faceted Classification Scheme:*	1686 Class numbers

4) *Uniterm Indexing System:* 3174 Terms including
607 Proper names
267 Numerals

b) **Methodology**

1) A number of people from different organizations were asked to select documents from the collection and in each case to frame a question to which that document would be an answer.

2) The project used manufactured queries, which were formulated before the beginning of the actual search. Altogether 400 queries were formulated and all were processed by the system in each of the three phases. Thus the system worked on a total of 1200 search queries.

3) The questions were put to the indexers.

c) **Results**

1) All four systems were operating with an effectiveness that could be expressed by a recall ratio between 60% and 90% with an overall average of 80%.

2) The average recall ratios for the different systems were as follows:

- Alphabetical Index: 81.5%
- Faceted classification: 74%
- UDC Scheme: 76%
- Uniterm Indexing: 82%

Subsequently, when the facet sequence was modified the recall factor of the faceted classification scheme increased to 83%.

3) Increased time spent in indexing improved the chance of recall. Recall ratios for different timing were as follows:

Time (in minutes)	Recall (%)
2	73
4	80
8	74
12	83
16	84

The apparent drop in efficiency at the 8 minutes level was difficult to explain. Even Cleverdon himself fails to explain the same.

4) There was no significant difference in retrieving documents indexed by the three different indexers. In other words, there was no significant difference in the performance of the three different indexers.

5) Success rates in retrieving documents in broad fields of Aeronautics were noted to be 4-5% better than those on the specialized field like high speed Aerodynamics.

6) The success rate in the third group of 6000 items was 3-4% greater than the second group suggesting that the documents were indexed better in the third group. That is to say, trained indexers without indexing experience were able to perform consistently good indexing job in spite of lack of subject knowledge.

d) **Failure Analysis**

Altogether 495 failures were listed to which 526 causes could be given. The causes were categorized as:

Question failure:	17%
Indexing failure:	60%
Searching failure:	17%
System failure:	6%

e) **Significance**

The results of the Cranfield 1 test contradicted the general belief regarding the nature of information retrieval systems in many ways. Firstly, the test proved that the performance of a system does not depend on the experience and subject background of the indexer. Secondly, it showed that systems where documents are organized by a faceted classification scheme perform poorly in comparison to the alphabetical index and uniterm system. Cranfield 1 test was important in two other respects. Firstly it identified the major factors that affect the performance of retrieval systems, and secondly it developed for the first time the methodologies that could be applied successfully in evaluating information retrieval systems. Moreover, it also proved that recall and precision are the two most important parameters for determining the performance of information retrieval systems, and that these two parameters are related inversely to each other.

f) **Criticisms**

Cranfield 1 was the most significant study in the decade of 1958-68. Even then it was not free from criticisms. One of the main criticisms against this study was the artificiality without much relation to the real life situation. The questions used for the test were 'manufactured' from the source documents. It was argued by the critics that to use an article as a source document for framing a question and then to report the results of search for the same document in response to the same question is very much artificial. In a real life situation queries are not based on previous knowledge of availability of the document in the store. Another criticism was that this test identified the level of performance of the indexing systems concerned, but it did not throw any light on reasons for failure of these systems. In fact these questions were taken into consideration in the second Cranfield test.

5.6.1.2 Cranfield—-WRU Test

Cranfield 1 was accompanied by other tests. An intermediate test between the Cranfield 1 and 2 conducted jointly with the Western Reserve University (WRU) needs special mention here. This was considered as a study of testing techniques rather than an evaluation of the index. The WRU system of indexing in metallurgical index was compared with a faceted system. The study took 950 documents and 114 questions based on source documents as input elements. Since this practice followed in Cranfield 1 was subjected to criticisms, this study was modified to include an exhaustive assessment of other documents for measuring recall and precision and deciding relevance. Two grades for relevance were considered–documents as relevant as the source and less relevant documents. The results of the study showed 75.8% recall and 33.7% precision for the Cranfield facet index. A critical analysis of failures placed major (67.1%) responsibility on searching and minor (18.4%) responsibility for indexing. The main causes of poor precision result of this study were grouped into two: (i) poor standard of the search programmes, and (ii) high level of exhaustive indexing. This study emphasized on the inverse relationship between recall and precision and influenced the programme of Cranfield 2 in sharpening the idea of index language devices.

5.6.1.3 Cranfield 2

The drawbacks of Cranfield 1 necessitated the conduct of further studies. The

second stage of Cranfield studies, known as Cranfield 2, began in 1963 and completed in 1966.The Cranfield 2 was a controlled experiment that attempted to investigate the components of index languages and their effect on the performance of retrieval systems. In Cranfield 2, the various index language devices were evaluated in terms of their effect on the recall and precision of a retrieval system. This study tried to assess the effect by varying each factor, while keeping the others constant. Some of the drawbacks of Cranfield 1 were eliminated in Cranfield 2 by bringing real-life situation in it and allowing feedback mechanisms between the indexers and users.

Test collection

1400 reports and research papers in the field of high-speed aerodynamics and aircraft structures were taken as input.

Query formulation

Each author of 200 selected research papers was asked to formulate questions for which he had cited the reference(s) in the paper. The authors were also asked to point out the documents that were not cited in their works but might have been relevant for the question they had formulated. The abstracts of the whole set of cited references were sent to the authors who were asked to assess the relevance of these with the questions they had formulated. The authors identified 1961 papers to be fully or partially relevant to the 279 questions obtained. Finally, 221 questions and 1400 documents were selected for the experiment. The success of the system was calculated by counting how many of the relevant papers thus assessed were retrieved by a given search.

Indexing

The documents were indexed in three different ways:

a) Each document was analysed and important concepts were selected and recorded in a natural language;

b) Concepts denoted by the single words were listed; and

c) Concepts with a weighting (ranging from 1 to 3) were combined to represent the subject contents of the documents.

Five different types of indexing languages with variations were used in this study:

- Single term index language with 8 variations like uncontrolled natural language, controlled synonyms, controlled word form, etc.;
- Simple concept index language with 15 variations by applying various controls;
- Controlled term index language with 6 similar variations;
- Title index with two variations; and
- Index language generated from abstracts with 2 variations.

Thus, 33 index languages were formed in this study by different levels of coordination of index terms.

Searching

To conduct searches 221 questions generated by the authors of the research papers were used. Relevancy of each document was ascertained for the questions by grading them from 1 to 4 in the following manner:

1) Complete answer to the question;

2) High degree of relevance;

3) Useful, providing general background of the work or dealing with a specific area; and

4) Minimum interest, providing information like historical viewpoint.

For the assessment, a single performance measure, called normalised recall, was introduced. This is a ratio of cumulated recall ratio and number of search stages involving document cut-off groups.

Results

The results of Cranfield 2 were rather unexpected because, other than corroborating inverse relationships of recall and precision, the results showed that :

1) A best performance result was obtained by the use of natural language single term index, such as Uniterm, based on words occurring in document texts;

2) With the natural language formation of groups or classes of terms beyond the stage of true synonyms or word forms resulted in fall of efficiency;

3) Use of precision devices like partitioning or intermixing was not as effective as the basic precision of coordination;

4) It was suggested that the terms taken from documents may be used successfully with minimum control in a post-coordinate index and it is helpful to eliminate synonyms. But any measures taken to control the vocabulary are likely to decrease its efficiency;

5) In the case where concepts were used for indexing, the system performance worsened with the introduction of superordinate, subordinate, and collateral classes along with the original concepts;

6) When broader and narrower terms were included along with the controlled languages of the thesaurus, the performance worsened; and

7) Index languages formed out of titles performed better than those formed out of abstracts.

Results and Conclusions

The results of the Cranfield 2 tests were unexpected because the test performing index languages were composed of uncontrolled single words occurring in documents. However, the variables used in the study were subject to criticisms. Each index language consisted of different units of words, phrases, or combinations of both. Both the documents and queries were formed in the same way. Thus the matching of questions to documents would evaluate the relative effectiveness of the languages of different specificity. Vickery comments that the measures used in the second Cranfield project do not adequately characterise those aspects of retrieval performance, those are of operational importance [1970].

Self Check Exercise

6) What were the objectives and input elements of the Cranfield 1 project?

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

5.6.2 MEDLARS Test

Most of the evaluation studies were carried out on small collection. But the evaluation conducted by F.W. Lancaster [1979] on the performance of the Medical Literature Analysis and Retrieval System (MEDLARS) of the National Library of Medicine, USA during August 1966–July 1967 was based on large collection of MEDLARS database which already contained 7,50,000 records relating to medical articles on magnetic tape (8). It was first major evaluation of an operating retrieval system. From the MEDLARS tape, monthly issues of *Index Medicus* were printed and terms used to index the subject of the articles (on average 6-7 terms per article) were drawn from a thesaurus of Medical Subject Headings (*MeSH)*, which then contained about 7000 main subject headings. The study involved the derivation of performance figures and conduct of detailed of failure analyses for a sample of 3000 real searches conducted in 1966-67. The main objectives of the MEDLARS test were:

1) to study user search requirements;
2) to determine how effectively and efficiently MEDLARS was meeting the user search requirements;
3) to identify factors adversely affecting the performance of the MEDLARS; and
4) to find out the ways to improve the performance of the MEDLARS.

a) **Methodology**

At the outset, a sample work statement consisting of a list of questions to be answered in the MEDLARS study was designed. It was decided that a target of 300 evaluated queries (i.e. fully analyzable test search requests) was needed to provide an adequate test. The range of queries should, as far as possible, be representative of the normal demand covering different subjects of medical literature like diseases, drugs, public health, and so on. Representativeness was achieved by stratified sampling of the medical institutions from which demands had come during 1965, and processing queries received from the sample institutions over a 12-month period. It was also decided to include all kinds of users (academic, research, pharmaceutical, clinical, Government) for the test and they should supply a certain volume of test questions for the test. The twenty-one user groups were so selected. Some 410 queries were received from the user group and processed, and finally 302 of these were fully evaluated and used in the MEDLARS test.

Queries were submitted to the MEDLARS and on receipt of the queries, MEDLARS staff prepared a search formulation (i.e. query designation) for each query using an appropriate combination of *MeSH* terms. A computer search was then carried out in the normal way. At this stage each user was also asked to submit a list of recent articles that he judged to be relevant to his query. The result of a search was a computer printout of references. Since the total items retrieved might be high (some searches retrieved more than 500 references), 25 to 30 items were selected randomly from the list and photocopies of these were provided to the searcher fro relevance assessment. Each searcher was asked to go through the full text of the articles and then to report about each article on the following scale of relevance:

H1 – of major value (relevance)

H2 – of minor value

W1 – of no value

W2 – of value not assessable (for example, in a foreign language).

The precision ratios were calculated by using the above scale of relevance: If L items were in the sample, the overall precision ratio was 100(H1 + H2)/L and the 'major value' precision ratio were 100H1/L. It was obviously not feasible to examine the whole MEDLARS database in relation to each search in order to establish recall ratio. Therefore, an indirect method was adopted for calculation of recall ratio: Each user was asked to identify relevant items for his query before receiving the search output and then search was carried out to find out whether those items were indexed in the database and retrieved along with other items that are both relevant and irrelevant. If t such relevant items were identified by the user and available on the database for a given query, and H were retrieved in the search, the overall recall ratio and 'major value' recall ratio was estimated as 100H/t and *100H1/t1* respectively.

The next stage of the evaluation was an elaborate analysis of retrieval failures, i.e., examining, for each search, collected data concerning failures include:

a) Query statement;

b) Search formulation;

c) Index entries for a sample of 'missed' items (i.e. relevant items that are not retrieved) and 'waste' items (i.e. noise—retrieval of irrelevant items); and

d) Full text (c).

b) Results

The average number of references retrieved for each search was 175, with an average or overall precision ratio of 50.4%; that is, of the average 175 references retrieved, about 87 were found to be not relevant. The overall recall ratio was 57.7% as calculated by an indirect method. Taking the average search, and assuming that about 88 of the references found were relevant, with an overall recall ratio of 57.7% implies that about 150 references should have been found, but 62 were missed. However, the recall and precision ratios for each of the 302 searches were analysed and individual ratios were then averaged in the MEDLARS test. The results were:

	Overall	*Major value*
Recall ratio	57.7%	65.2%
Precision ratio	50.4%	25.7%

An elaborate analysis of retrieval failures revealed that over the 302 searches, there were 797 recall failures and 3038 precision failure. Major categories of the failures were attributed to the principal system components as shown below:

Table 5.4 : Retrieval Failures

System attribute/component	Recall failure (%)	Precision failure (%)
Index language	81 (10.2)	1094 (36)
Indexing	298 (37.4)	393 (12.4)
Searching	279 (35)	983 (32.4)
User-system interaction	199 (25)	503 (16.6)
Others	11 (1.4)	78 (2.5)

c) Conclusions

The results of the MEDLARS test led to a series of recommendations for the improvement of the MEDLARS performance. Some notable changes made to the MEDLARS as a result of this test include design of a new search request

form (intended to ensure that a request statement is a good reflection of the information need behind it); and expansion of the entry vocabulary and improvement of its accessibility, and the adoption of an increased level of integration among personnel involved in indexing, searching and vocabulary control devices.

Self Check Exercise

7) What was the scale used in MEDLARS test for calculating precision ratio?

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

5.6.3 SMART Retrieval Experiment

The SMART retrieval system, based on the processing of abstracts in natural language forms, was launched in 1964 and Gerard Salton carried out the evaluation of various searching options offered by the SMART under laboratory condition.

The SMART retrieval system was a unique experimental environment for the development and evaluation of automated retrieval techniques. Documents were represented in the SMART system by a set of weighted terms (term vectors) and a set of documents is represented in a term assignment array. Each term was assigned a weight, which was positive, if an index term was actually assigned to a document and zero if it was not. An individual query was similarly represented as a vector of query terms. Automatic indexing using the term discrimination model carried out the creation of the document vector. In this model index terms were evaluated on the basis of their ability to increase the average dissimilarity of document description in a database. A good indexing term increased the average dissimilarity of documents in the collection; a bad one decreased it. Terms were then assigned discrimination values according to the degree to which they increase or decrease average document dissimilarity. Retrieval was based on the degree of similarity between the document vectors and the query vectors.

A number of methods were adopted for automatic content analysis of documents like Word suffix cut-off methods, Thesaurus look-up procedures, Phrase generation methods, Statistical term associations, Hierarchical term expansion, etc.

In order to develop a prototype for a fully automated information retrieval system a large number of search experiments were performed in the SMART experiment. The user environment was simulated by running iterative searches based on user feedback. The evaluation procedures of the SMART experiment was based on a pair-wise comparison of the effectiveness of two or more processing methods and a number of evaluation parameters were computed for each of the processing methods. It was followed by a comparison of the corresponding measures for two or more methods to produce a ranking of the methods in decreasing order of retrieval effectiveness. The following evaluation measures were generated by the SMART system:

i) A recall-precision graph reflecting the average precision value at ten discrete recall points — from a recall of 0.1 to a recall of 1.0 in intervals of 0.1;

ii) Two global measures, known as normalized recall and normalized precision, which together reflect the overall performance level of the system; and

iii) Two simplified global measures, known as rank recall and log precision, respectively.

a) **Methodology**

A collection of 1268 abstracts in the field of library science and documentation, mainly published in American documentation, 1963-1964, and also in some other journals was used for this experiment.

Eight different persons familiar with the subject either as a librarian or as a student of library science were asked to generate a total of 48 different search requests in the documentation field in grammatically correct and unambiguous English language. Following the receipt of the query formulation from each of the eight persons, a series of forty-eight requests were searched against a file of 1268 records, using the various search options that SMART system allows. Then, the text of the document abstracts were distributed, and each person was asked to assess the relevance of each document abstract with respect to each of his six queries. An essential element in the SMART experiment was the *relevance feedback*. If, in a preliminary output, the user can indicate which items are relevant and which are not, the system recalculate the weight of the items in the database. This is done by reducing weights associated with the characteristics of non-relevant items and increasing the weight of the characteristics associated with the relevant ones. Four sets of judgements were compared. The relevance judgement groupings were as follows:

Group	Judges	Function
A		Original group of query authors. Each person in the A group made relevance judgments for his six queries.
B		Non-author judges. Each person in the B group made relevance judgments for six queries corresponding to six different authors from the A group.
C		The document is relevant to a given query if either the A judge or B judge termed it relevant.
D		The document is relevant to a given query if both A and B judges termed it relevant.

After receiving the relevance judgements, called the 'A judgements', a second and independent set of relevance judgements called the 'B judgements', was obtained by asking each person in the test group to judge for relevance six additional queries originated by six different people. The same relevance criteria were used for the second relevance judgements as for the original ones; the only difference was that the 'A judgements' were rendered by the query authors, whereas the 'B judgements' were the non-author judgements. In order to achieve the objectivity in the experiment, the 'B judges' were not informed of the 'A judgements' obtained previously, nor was there any interaction between assessors either before or during the process of judgement.

Thus, for each of the 48 queries, a set of four different documents sets became available, each consisting of the items termed relevant by a different set of people as follows:

A set – relevance assessed by the query author;

B set – relevance assessed by outside subject expert;

C set – relevance asserted by either A or B assessor; and

D set – relevance asserted by both A and B assessor.

Three automatic language analysis procedures include in the SMART system, known as word form, word stem, and thesaurus, are furnished below:

1) Word form: Common words and final 's' endings were removed from the texts of document abstracts and queries, and weights were assigned to the remaining word forms; the resultant texts were then matched to obtain the document-query correlation coefficient.

2) Word stem: Texts were treated in the same way as word form, except that complete suffixes are removed from the text words to reduce the texts to weighted word stems; the query-document matching process remains the same.

3) Thesaurus: Each word stem produced by procedure (2) was checked with a thesaurus providing synonym recognition, and the resulting weighted concept identifiers assigned to queries and documents were compared (instead of word forms and word stems).

b) **Results**

It was found that, under normal circumstances, an evaluation of performance for a variety of processing methods required an examination of the ranking of the corresponding recall-precision curves, rather than a detailed comparison of the actual recall and precision values. From a ranking of the recall-precision graphs obtained from the several processing methods it was noted that :

a) Although the overall consistency of relevance agreements between the groups was not particularly high, the relative performance of the various retrieval methods was unaffected by changes in the relevance decisions; that is, all four sets of relevance judgements caused the same ranking of alternative search procedures. To be more specific, the word-form process was found to be less powerful than the two other procedures, and the thesaurus process was slightly better than the word stem match.

b) The best results in terms of recall and precision were obtained for the D judgements, which represented the agreements between both the A and B relevance judges; for low recall the precision was about 20% higher for D than for A, B, or C.

It was observed that the SMART evaluation output did not vary with the variations in the relevance judgments. The recall-precision output was basically invariant for the collection under study. It was concluded that if the relevance assessments by the query authors are typical of what can be expected from the general user populations, then 'the resulting average recall-precision figures appear to be stable indicators of system performance which do in fact reflect actual retrieval effectiveness'.

Self Check Exercise

8) What were the methods adopted for automatic content analysis of documents in the SMART retrieval experiment?

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

5.6.4 TREC Experiment

It has been observed that almost all the earlier evaluation experiments were based on a small data set and unrelated to the real-life problem. The major problem for IR researchers was to base the evaluation experiments on large test collection to match a real-life situation with an infrastructural support for conducting test on them. Under such circumstances, TREC (Text REtrieval Conference) series of experiments in information retrieval, funded by DARPA (Defence Advanced Research Project Agency, Department of Defence, USA) and operated by the NIST (National Institute for Science and Technology, USA), was launched in 1991 in order to enable IR researchers to scale up from small collection of data to larger experiments. The TREC series of experiments has drawn attention of the LIS professionals all over the world since its inception and has shown that significant research results can be obtained through international efforts and collaboration.

a) **Objectives**

The objectives of the TREC experiments are to :

1) encourage IR evaluation experiments on large test collections;
2) enable the exchange of ideas among industry, academia, and government in an open forum created for the purpose;
3) facilitate the speedy transfer of technology from research labs to commercial products by demonstrating substantial improvements in retrieval methodologies on real-life problems; and
4) develop new evaluation techniques applicable to current retrieval systems and to make them available for use by industry and academia.

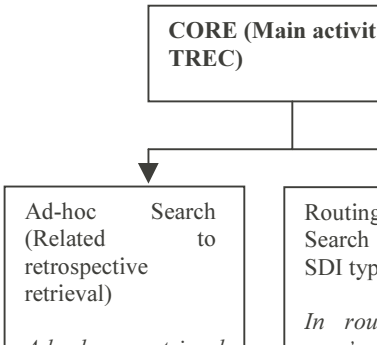
b) **Scope**

A wide range of information retrieval strategies was tested in different TREC experiments (i.e. from TREC 1 in 1992 to TREC 12 in 2003). Some notable examples are:

a) Boolean search;
b) Statistical and probabilistic indexing and term weighting strategies;
c) Passage or paragraph retrieval;
d) Combining the results of more than one search;
e) Retrieval based on prior relevance assessments;
f) Natural language-based and statistically based phrase indexing;
g) Query expansion and query reduction;
h) String and concept-based searching;
i) Dictionary-based searching;
j) Question-answering;
k) Content-based multimedia retrieval.

c) **Structure of TREC Experiment**

The two-fold information retrieval activities—i.e. (1) *Core* (i.e. Main activity), and (2) *Tracks* (i.e. Subsidiary activities) have been recognised by the TREC for the purpose of experiment. The following figure shows the structure of the TREC experiments:



TREC Experiments (1992-2003)	Area of Research / Investigation
TREC-1	Data structuring and system engineering
TREC-2	Automatic query formulation and NLP
TREC-3	Interactive query formulation in multiple databases
TREC-4	Problems of short user statements
TREC-5	Information retrieval of non-English languages
TREC-6	Cross-language and spoken document IR
TREC-7	Large query formulation
TREC-8	Web integrated IR
TREC-9	IR related to image and NLP interface
TREC-10	IR related to video
TREC-11	Fine tune searching within the ranked set of documents
TREC-12	IR specific to bioinformatics and genomics

Fig. 5.2 : Structure of TREC experiments

Core or main activity has been divided into (i) *Ad-hoc search* (i.e. retrospective retrieval), and (ii) *Routing or Filtering search* (i.e. SDI). The ad-hoc task investigates the performance of the system that search a static set of document—from a database or from the Web—in response to a question and produces a ranked list of items. In TREC jargon, an information need is termed as a *topic,* and the data structure submitted to a retrieval system is called a *query. Routing or filtering search* is needed either (a) by the researchers to keep track of the latest development in their field of interest, or (b) by the analysts to monitor news feed for items of interest on a particular subject. This type of search produces an unordered set of documents as opposed to a ranked list. Here, the retrieval system makes a decision whether or not a particular document is of relevance to the given user's query.

A *track* acts as an incubator for a new area of experiment. The first running of a *track* usually defines the problem and creates the necessary infrastructure, including test collections, evaluation methodology, etc. to support experiment on the specific task. TREC experiments have also become evolutionary in the sense that new tracks have been added as and when new problems areas are encountered by the researchers.

d) TREC Collection

The document collection of TREC (known as TIPSTER collection) reflect diversity of subject matter, word choice, literary styles, formats, and so on. The primary

TREC collection consists of about 2 gigabytes of data with over 8,00,000 documents. The document collection used in various tracks depends on the needs of the track and availability of data. The primary TREC document sets consist mostly of newspaper or newswire articles. Some government publications such as Federal Register, patent documents and computer science abstracts are also included. Each document is encoded in SGML and is assigned a document number.

e) Methodology

A program committee consisting of representatives from government, industry and academia was formed to supervise the TREC activities. For each TREC, NIST provided a test set of documents and questions. TREC participants were allowed to run their own retrieval system on the data and returned a list of the retrieved top ranked documents to NIST where the individual results were pooled, the retrieved documents were judged for correctness, and finally, the results were evaluated. The TREC cycle ends with a workshop— a forum for participants to share their experiences. The TREC workshop is organised every year and latest one was the 12th in the series, held at NIST in 2003.

Topic statements (i.e. query statements), which may be either fully automatic or purely manual or mixed, were created by the same person who performed the relevance assessment for that topic. Assessors came to NIST with their ideas for topics, searched the document collection and finally estimated the number of relevant documents per topic. Relevance judgments in TREC were based on binary decision—whether a document was relevant or not. Identifying all the relevant documents against a particular query from a collection of about a million records for estimating recall was logically absurd. Hence, relative recall instead of absolute recall for each query was calculated. In calculating relative recall, 100 top-ranking relevant documents for each system and for each query were combined together to produce a 'pool' of relevant documents. Then, each system's retrieved documents for a query or topic were compared with these relevant documents. The proportion of the 'pool' of relevant documents that a single system retrieved for a query or topic is relative recall. Output lists of 100 top-ranking documents from each participating research team were sent to NIST where the 100 top-ranking documents for each topic from all the participating research teams were merged into a single set and then handed over to the assessor for relevance evaluation. Most of the relevant documents were found by pooling all the results from all the participating teams. The results for each system were then analysed to generate various performance measures.

The ad-hoc retrieval tasks in TREC were evaluated using package called tree-eval, which reported about 85 different numbers for a run, including recall and precision measures at various cut-off points and a single value summary measure for recall and precision. The recall—precision curve (a plot of precision as a function of recall) and mean average precision (mean of the precision obtained after each relevant document is retrieved) were the most commonly used measure to describe TREC results.

f) Results

The TREC series of experiments has produced very significant and interesting results. The findings of each experiment (along with specific reports) appear regularly on TREC Website (http://trec.nist.gov). Some important findings of the TREC experiments are furnished below:

i) Use of 'pooling' to produce the sample results for relevance judgments was

found to be more than adequate for test collection;

ii) Automatic construction of queries from natural language queries seems to work well, or better than, manual construction of queries, which is encouraging for groups supporting the use of simple natural language interfaces for retrieval systems;

iii) Significant improvement of retrieval performance over TREC 1 was observed in TREC 2 in respect of the routing task with the increase of documents per topic (200 per topic to 1000) and database size (1 gigabyte to 3 gigabytes);

iv) Level of performance was same in spite of differences experimental designs. For examples: some groups generated queries automatically from the topic statements while others generated the queries manually; relevance feedback was not incorporated in many systems; the computer platform used ranged from PCs to a supercomputer;

v) Differences in precision-recall curve was minimal; and

vi) Although the precision-recall results were similar, there was a large scatter in the actual documents retrieved.

g) **Criticism**

The main criticism against TREC centres round on the methodological issues. TREC experiments were carried out within the traditional laboratory paradigm, which is very difficult to relate to users browsing information on the Web as opposed to traditional library setting.

Self Check Exercise

9) Mention the different information retrieval strategies that were tested in the TREC series of experiments.

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

5.7 SUMMARY

This Unit presents a holistic view of the evaluation of IR systems. The indexing system is a major component of IR system. In the context of information processing and organisation, evaluation of indexing system is thus most crucial for improving the efficiency of the system. This Unit thus lays major emphasis on the evaluation of indexing system. It starts with the discussion on different levels of evaluation like system effectiveness, cost effectiveness and cost-benefit evaluation, and different evaluation criteria consisting recall, precision, fallout, specificity, novelty, relevance, etc. Major steps involved in the evaluation of the performance of an information retrieval system have been discussed. Important evaluation experiments—such as, Cranfield 1, Cranfield-WRU test, Cranfield 2, MEDLARS test, SMART Project and TREC have also been discussed in terms of their objectives, scope, methodology, results, significance and criticisms.

5.8 ANSWERS TO SELF CHECK EXERCISES

1) By *effectiveness* we mean the level up to which the given IR system attains its stated objectives, whereas *efficiency* refers to how economically the given IR system is achieving its stated objectives. In order to know the *effectiveness* of the system we measure the extent to which the given IR system can retrieve relevant documents while withholding non-relevant documents. But, we measure only the cost factors like response time, user effort, cost involved per search, etc. for determination of the efficiency of the system.

2) An information retrieval system can be evaluated at the following levels:

   a) System effectiveness;

   b) Cost effectiveness; and

   c) Cost-benefit evaluation

3) Different criteria used in evaluating an IR system are :

   a) Recall;

   b) Precision;

   c) Response time;

   d) User effort;

   e) Form of presentation; and

   f) Collection coverage.

4) The term *Recall* relates to the ability of the system to retrieve relevant documents. Recall ratio is nothing but the proportion of relevant documents retrieved. This ratio helps in measuring the extent to which the retrieval of relevant documents occurs in a given system. The formula for the calculation of *Recall ratio* is:

$$\frac{\text{Relevant retrieval}}{\text{Total relevant}} \quad 100$$

On the other hand, *precision ratio* is the ratio between the relevant retrieved and total retrieved documents. It measures how precisely an indexing system functions. The formula for the calculation of the *precision ratio* is:

$$\frac{\text{Relevant retrieval}}{\text{Total retrieval}} \quad 100$$

5) The following major steps are involved in the evaluation of the performance of an IR system:

   a) Defining the scope of evaluation;

   b) Designing the evaluation programme;

   c) Execution of the evaluation programme;

   d) Analysis and interpretation of results; and

   e) Modifying the system in the light of the evaluation results.

6) The objective of the Cranfield 1 project was to investigate the comparative efficiency of four indexing systems: UDC, Faceted classification, Alphabetical subject heading list and Uniterm indexing system. Input elements for this project consist of :

   a) Four indexing systems as stated above;

   b) Three indexers with varied backgrounds;

c) Variable indexing time periods;

d) 100 micro documents on the general field of aeronautics and the specialized field of high-speed aerodynamics;

e) 18,000 indexed items (100 micro documents × 3 indexers × 4 indexing systems × 5 time periods × 3 runs); and

f) Descriptors.

7) In the MEDLARS test, the user was asked to characterize each retrieved item in a randomly selected sample output of 25 to 30 items by using the following scale: H1 – of major value; H2 – of minor value; W1 – of no value; W2 – of value unknown (because in foreign language). From this assessment, precision ratios were calculated: If L items were in the sample output, the overall precision ratio was 100(H1 + H2)/L; the 'major value' precision ratio was 100H1/L.

8) A number of methods was adopted for automatic content analysis of documents in the SMART retrieval experiment. Some notable examples include:

a) Word suffix cut-off methods,

b) Thesaurus look-up procedures,

c) Phrase generation methods,

d) Statistical term associations,

e) Hierarchical term expansion and so on.

9) A wide range of information retrieval strategies was tested in different TREC experiments—that is, from TREC 1 in 1992 to TREC 12 in 2003. Some notable examples are:

a) Boolean search;

b) Statistical and probabilistic indexing and term weighting strategies;

c) Passage or paragraph retrieval;

d) Combining the results of more than one search;

e) Retrieval based on prior relevance assessments;

f) Natural language-based and statistically based phrase indexing;

g) Query expansion and query reduction;

h) String and concept-based searching;

i) Dictionary-based searching;

j) Question-answering; and

k) Content-based multimedia retrieval.

5.9 KEYWORDS

Clumps : A group of words that tend to 'adhere'. Various criteria can be used to define the boundaries of a clump. In general, each word is associated with the others as a group above some given threshold. The formation of word classes is likely to be based on more than simple concurrence. Instead, two words are considered to be related if they co-occur more frequently than one expects them to co-occur probabilistically.

Clustering : A group formation mechanism that combines a number of related or associated terms. Four kinds of group may be found in clustering: strings, stars,

cliques and clumps. String occurs when term A is strongly associated with term B, term B with term C, and so on. In practice, strings tend to form loops fairly quickly: term A->B->C->D->E->A. Stars are found when one term is equally strongly related each to other. Cliques occur when a set of terms are strongly related each to the other. Clumps are weaker form of clique, in which a term is related to one or more of the others in the clump, but not necessarily to all.

Cost effectiveness : It is the evaluation in terms of how to satisfy user requirements in the most efficient and economical way. A cost effectiveness evaluation relates measures of effectiveness to measure of cost. It takes into considerations unit cost per relevant citation retrieval, per new relevant citation retrieval and per relevant document retrieval.

Cost-benefit : It is the evaluation of the worth of the system by relating the cost of providing the service to the benefits of having the service available.

Elimination factor : The proportion of non-retrieved items (both relevant and non-relevant) over the total items in the collection.

Evaluation : It refers to the act of measuring the performance of the system in terms of its retrieval efficiency (ease of approach, speed and accuracy) to the users, and its internal operating efficiency, cost effectiveness and cost benefit to the managers of the system in order to ascertain the level of its performance or its value.

Exhaustivity : It is the measure of the extent to which all the distinct topics discussed in the documents are considered.

Failure analysis : A diagnostic procedure, which involves the analysis of recall and precision failures encountered to the principal subsystems of an IR system like indexing, searching, index language, and user-system interface. It consists of a careful examination of the documents involved, how they were indexed, the request posed to the system, the search strategy used, and the complete assessment forms of the requesters.

Fallout ratio : It is the proportion of the irrelevant items in the file, which are retrieved by a request. The fall out thus represents the fractional recall of irrelevant items. It has been referred to as the conditional probability of 'false drops'. It is also called 'discard'.

Generality number : It expresses the number of items relevant to a particular request over the total number of items in the collection. The higher the generality number the greater the density of relevant items to total

collection, and the greater this density the easier the search tends to be.

Noise factor : The proportion of retrieved items those are not relevant. This factor is considered as the complement of the precision ratio.

Novelty ratio : The proportion of relevant documents retrieved in a search that are new to the requester, that is, brought to his attention for the first time by the search.

Omission factor : The proportion of non-relevant items retrieved over the total number of non-retrieved items in the collection.

Performance Curve : A performance curve is nothing but a line representing graphically a series of variable performance points in terms of recall and precision ratios affected by search strategy at varying levels, from very broad to very specific. This curve expresses the inverse relationship between recall and precision.

Resolution factor : The proportion of total items retrieved over a total number of items in the collection.

Semantic factoring : Involves the analysis of a word into largest number of ideas implied. During the 1950s a team at Case Western Reserve University worked on a system of analysis known as semantic factoring. The objective was to break down every concept into a set of fundamental concepts called semantic factors. Because of their fundamental nature, there would only be a limited number of these factors. For example, word such as 'urinalysis' factors into 'urine' and 'analysis', whereas 'magneto-hydrodynamics' factors into 'fluid flow', 'magnetism' and 'electrical conductivity'.

Specificity : It is measure of the degree of preciseness of the subject to express the thought content of the document.

System effectiveness : It is the evaluation of the system performance in terms of degree to which it meets the users' requirements. It considers the users' satisfaction and is measured by determining the utility to the users of the documents retrieved in answer to a user query. It takes into consideration cost, time and quality criteria.

Usability : The value of references retrieved in terms of such factors as their reliability, comprehensibility, currency, etc.

User effort : It refers to the intellectual as well as physical effort required from the user in obtaining answers to the search requests. The effort is measured by the amount of time user spends in conducting the search or negotiating his enquiry with the system.

5.10 REFERENCES AND FURTHER READING

Chakraborty, A.R. and Chakraborty, B. (1984). *Indexing: principles, processes and products*. Calcutta: World Press.

Chowdhury, G.G. (2004). *Introduction to modern information retrieval*. 2nd Ed. London: Facet Publishing.

Cleverdon, C.W. (1962). *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. Cranfield: College of Aeronautics.

Foskett, A.C. (1996). *Subject approach to information*. 5th Ed. London: The Library Association.

Ghosh, S.B. and Satpathi, J.N. (eds.) (1998). *Subject indexing systems: concepts, methods and techniques*. Calcutta: IASLIC.

Gopinath, M.A. (1999). *Evaluation of information storage and retrieval (ISAR) systems*. In: MLIS-03, Block 3, Unit 12 course materials. New Delhi: Indira Gandhi National Open University.

Jones, Karen Sparck. (ed.) (1981). *Information retrieval experiment*. London: Butterworth.

Lancaster, F W. (1979). *Information retrieval systems: characteristics, testing, and evaluation*. 2nd Ed. New York: John Wiley.

Salton, G. and McGill, M.I. (1983). *Introduction to information retrieval*. New York: McGraw-Hill.

Sarkhel, J.K. (2001). *Information analysis in theory and practice*. Kolkata: Classique Books.

TREC. Available at http://trec.nist.gov/pubs.html.

Vickery, B.C. (1970). *Techniques of information retrieval*. London: Butterworth.