A Gateway to All Post Graduate Courses

An MHRD Project under its National Mission on Education throught ICT (NME-ICT)

Subject: Library and Information Science

Production of Courseware

-Content for Post Graduate Courses

Paper No: 07 Information Storage and Retrieval

Module : 07 Evaluations and Measurement of Information Retrieval System

Development Team

Principal Investigator & Subject Coordinator	Dr. Jagdish Arora, Director INFLIBNET Centre, Gandhinagar
Paper Coordinator	Prof Devika P Madalli, Professor, Documentation Research and Training Centre (DRTC), Bangalore
Content Writer	Dr Nanaji Shewale, Librarian, Gokhale Institute of Politics and Economics (GIPE)
Content Reviewer	Prof Devika P Madalli, Professor, Documentation Research and Training Centre (DRTC), Bangalore

Evaluation and measurement of Information Retrieval System

I. Objectives

The objectives of this module are to:

- Introduce the need for evaluating information retrieval systems.
- Introduce different points of view of evaluation study of IR systems.
- Familiarize the reader about different factors which can affect the performance of IR systems.
- Enlist different criteria to evaluate information retrieval systems.
- Introduce the measures of precision and recall.
- Introduce various retrieval tests and experimental tools which will help in evaluating the effectiveness of IR systems.

II. Learning Outcomes

After reading this Module:

- The reader will gain the knowledge of evaluation study and its benefits in IR.
- The readers will enrich their knowledge about various evaluation criteria and the importance of user oriented evaluation criteria for evaluating the IR systems.
- The reader will gain the knowledge of relationship between the recall and precision, fall out and generality.
- The reader will also learn about different evaluation tests for evaluation of information retrieval system.

III. Structure

1. Introduction

Evaluation is a process of feedback against an investment in time, energy, money, knowledge and intelligence. Evaluation usually results in indicators that gauge the usefulness of systems and services. Information storage and retrieval systems are evaluated from viewpoints such as users, economy, coverage, hardware, software, man-power, environmental conditions, etc.

2. Need for Evaluation

Evaluation studies investigate the degree to which the stated goals or expectations have been achieved or the degree to which these can be achieved. Keen (1971) gives three major purposes of evaluating an information retrieval system as follows:

- The need for measures with which to make merit comparison within a single test situation. In other words, evaluation studies are conducted to compare the merits (or demerits) of two or more systems
- the need for measures with which to make comparisons between results obtained in different test situations, and
- The need for assessing the merit of real-life system.

Swanson (1971) states that evaluation studies have one or more of the following purposes:

- To assess a set of goals, a program plan, or a design prior to implementation.
- To determine whether and how well goals or performance expectations are being fulfilled.
- To determine specific reasons for successes and failures.
- To uncover principles underlying a successful program.
- To explore techniques for increasing program 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 objectives or to redefine sub goals or goals in view of research findings.

3. Different Evaluation Criteria

An evaluation study can be conducted from two different points of view. When it is conducted from managerial point of view, the evaluation study is called management-oriented; conducted from users' point of view it is called a user-oriented evaluation study. Many information scientists advocate that an evaluation of information retrieval system should always be user-oriented, i.e. evaluators should pay more attention to those factors that can provide improved service to the users.

Cleverdon (1962) says that a user oriented evaluation should try to answer the following questions which are quite relevant in modern context too:

- To what extent does the system meet both the expressed and latent needs of its users' community?
- What are the reasons for the failure of the system to meet the users' needs?
- What is the cost-effectiveness of the searches made by the users themselves as against those made by the intermediaries?
- What basic changes are required to improve the output?
- Can the costs be reduced while maintaining the same level of performance?
- What would be the possible effect if some new services were introduced or an existing service were withdrawn?

As with any other system, we expect the best possible performance at the least cost from an information retrieval system. We can thus identify two major factors, performance and cost. Now, if we try to determine how we measure the performance of an information retrieval system we have to go back to the question of its basic objective. We know that the system is intended to retrieve all relevant documents from a collection. The system, therefore, should retrieve relevant and only relevant items. One also needs to assess how economically a system performs. Calculations of costs of an information retrieval system are not quite easy as it involves quite a number of indirect methods of calculation of costs. Lancaster (1979) lists the following major factors to be taken into consideration for the cost calculation:

- cost incurred per search
- users' efforts involved
    - in learning how the system works
    - in actual use
    - in getting the documents through back-up document delivery system
    - in retrieving information from the retrieved documents, and
- users' time
    - from submission of query to the retrieval of references
    - From submission of query to the retrieval of documents and the actual information.

A number of studies have been conducted so far to determine the cost of information retrieval system and subsystems.

In 1966, Cleverdon identified six criteria for the evaluation of an information retrieval system. These are:

- **recall**, i.e., the ability of the system to present all the relevant items
- **precision**, i.e., the ability of the system to present only those item that are relevant
- **time lag**, i.e. the average interval between the time the search request is made and the time an answer is provided
- **effort**, intellectual as well as physical, required from the user in obtaining answers to the search requests
- **form of presentation** of the search output, which affects the user’s ability to make use of the retrieved items, and

- **Coverage of the collection**, i.e. the extent to which the system includes relevant matter.

Vickery (1970) identifies six criteria, grouped into two sets as follows:

Set 1

- Coverage = the proportion of the total potentially useful literature that has been analysed
- recall – the proportion of such references that are retrieved in a search, and
- 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

- precision – the ability of the system to screen out irrelevant references
- usability – the value of the references retrieved, in terms of such factors as their reliability, comprehensibility, currency, etc., and
- Presentation – the form in which search results are presented to the user.

In 1971, Lancaster proposed five evaluation criteria:

- coverage of the system
- ability of the system to retrieve wanted items (i.e. recall);
- ability of the system to avoid retrieval of unwanted items (i.e. precision)
- the response time of the system, and
- The amount of effort required by 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 (1983) identified the various parameters of an information retrieval system as related to each of five evaluation criteria:

No	Evaluation Criteria	System Parameters
1	Recall and precision	• Indexing exhaustively – Recall tends to increase the exhaustively of indexing terms. • Term specificity – Precision increases with the specificity of the index terms • Indexing language – Availability of measures of recognition of synonyms, terms relations, etc., which improve recall. • Query formulation – Ability to formulate an accurate search request • Search strategy – Ability of the user or intermediary to formulate an adequate search strategy.
2	Response time	• Organization of stored documents.

		• Type of query. • Location of information centre. • Frequency of receiving user's queries. • Size of the collection.
3	User effort	• Accessibility of the system. • Availability of guidance by system personnel. • Volume of retrieved items. • Facilities for interaction with the system.
4	Form of presentation	• Type of display device. • Nature of output – bibliographic reference, abstract, or full text.
5	Collection coverage	• Type of input device and type and size of storage device. • Depth of subject analysis. • Nature of users' demand • Physical forms of documents.

4. Evaluation of Outcome

Some of the performance criteria mentioned above can be measured easily. For example, the parameters related to the collation coverage, and form of presentation is related to policy matters, and thus is defined by the system managers beforehand. However, the two other criteria, recall and precision, cannot be measured so easily.

4.1 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 occurs. Whenever a user puts his / her query, it is the responsibility of the system to retrieve all those items that are relevant to the given query. However, in reality it may not be possible to retrieve all the relevant items from a collection, especially when the collection is large. Thus, a system may be able to retrieve a proportion of the total relevant documents in response to a given query. The performance of a system is often measured by recall ratio, which denotes the percentage of relevant items retrieved in a given situation.

The general formula for calculation of recall and precision may be stated as:

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items in the collection}} \times 100$$

$$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}} \times 100$$

Recall thus relates to the ability of the system to retrieve relevant documents, and precision

relates to its ability not to retrieve non-relevant documents. The ideal system attempts to achieve 100% recall and 100 % precision, i.e. it attempts to retrieve all the relevant documents and relevant documents only. However, this is not possible in practice because as the level of recall increases, precision tends to decrease. They are inversely proportional. Following example shows the relationship between recall and precision for a given search.

Let us suppose that in a given situation a system retrieves 'a+b' number of documents, out of which 'a' documents are relevant, and 'b' documents are non-relevant. Say, for example, 'c+d' documents are left in the collection after the search has been conducted. This number will be quite large, because it represents the whole collection minus the retrieved documents. Out of the 'c+d' number, let's say, 'c' documents are relevant to the query but could not be retrieved, and 'd' documents are not relevant and thus have been correctly rejected. For a large collection the value of 'd' will be quite large in comparison to c because it represents all the non-relevant documents minus those that have been retrieved wrongly (here b). Lancaster suggests that these statistics can be represented as in the following table.

	Relevant	Not-Relevant	Total
Retrieved	a (hits)	b (noise)	a+b
Not Retrieved	c (misses)	d (rejected)	c+d
Total	a+c	b+d	a+b+c+d

Table 1: Recall – Precision Matrix

So as per the above table, 'a' denotes the 'hit' and 'b' denotes the 'noise'. Now, 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. The recall and precision ratio in this case can be calculated as

$R = [a / (a + c)] X 100$
$P = [a / (a + b)] X 100$

Recently, the theory of the 'inverse relationship between precision and recall' has been questioned by Fugmann (1993). By several examples, he has shown that:

- An increasing in precision is by no means always accompanied by a corresponding decrease in recall, and
- An increase in recall is by no means observed to have always in its wake a decrease in precision.

4.2 Fallout and generality

While conducting a search, it is quite likely that some non-relevant items could be retrieved in a given search. This is often termed as the fallout ratio. At the same time, the proportion of relevant documents in the collection for a given query is called the generality ratio. Recall, precision, fallout, and generality ratios have been represented by Salton (1971) as shown in Tables 8.1 and Table 8.2. Thus from Table 8.1, the cut-off can be determined by the following formula:

Cut-off = [(a + b } / (a + b + c + d)]

Van Rijsbergen proposes that recall and precision can be combined in a single measure, called effectiveness, or E, which is a weighted combination of precision and recall where the lower the E value, the greater is the effectiveness. If recall and precision are represented by R and P respectively, the value of E can be measured through the following formula.

$E = 100 X [1 - (1 + \beta^2)^{PR} / (\beta^2 P + R)]$

Where β is used to reflect the relative importance of recall and precision to the user ($0 < \beta < \infty$); $\beta = 0.5$ corresponds to attaching half as much importance to recall as precision.

Symbol	Evaluation Measure	Formula	Explanation
R	Recall	a/(a+c)	Proportion of relevant items retrieved
P	Precision	a/(a+b)	Proportion of retrieved items that are relevant
F	Fallout	b/(b+d)	Proportion of non-relevant items retrieved
G	Ganerality	(a+c)/(a+b+c+d)	Proportion of relevant items per query

Table 2: Retrieval Measures

4.3 Limitations of recall and precision

Number of documents to be retrieved: Different users may want different levels of recall, like a person going to prepare a state-of-the-art-report (SOTAR) on a topic would like to have all the items so the he / she will go for a high recall. Whereas, the user wanting to know 'something' about a given topic will prefer to have 'a few items', and thus will not require a high recall. Here, the major problem is that users very often are unable to specify exactly how many items they want to be retrieved.

Recall assumes that all relevant items have the same value, which is not always true. The retrieved items may have different degrees of relevance and this may vary from user to user and even from time to time to the same user. Both recall and precision depend largely on the relevance judgments of the user. The judgement is quite subjective and there may be different degrees of the retrieved output.

5. Types of Evaluation Experiments

Lancaster (1979) identifies five major steps involved in the evaluation of an information retrieval system, which are

- designing the scope of evaluation
- designing the evaluation program
- execution of the evaluation

- analysis and interpretation of results, and
- Modifying the system in the light of the evaluation results.

Step 1

In the step 1, an evaluation study is conducted to determine the level of performance of the given system. The evaluation would also point out the weaknesses of the system and the reasons for the same. This stage is where planning of evaluation is done, setting its purpose and scope, choosing appropriate methods, costing and staff are all decided.

Step 2

In step 2 following the basic objectives set and the proposed plans, the parameters for data collection are worked out. The evaluator must identify the points on which data are to be collected and a methodology is proposed. Data collection plans are outlined. Also the plans for exploiting data and required manipulations are decided upon.

Step 3

Step 3 deals with execution of the evaluation. This is time consuming. The data collectors have to collect data according to the plan and methods prescribed in the previous stage. Any alterations required to the proposed plan due to constraints at data collection stage must be communicated to the evaluator by the data collection personnel. Mutually they can agree upon required adjustments and execute them.

Step 4

Step4 is about analysis and interpretation of the data. The success and impact of evaluation mainly depends on the accuracy of the interpretations. Data is manipulated and analysed according to aimed objective and results are obtained. These results are interpreted in the elight of the objectives of the evaluation.

Step 5

Finally, feedback based on the results and interpretation of evaluation is given to the retrieval system so that it may be modified accordingly.

5.1 Cranfield Tests

The first extensive evaluation of information retrieval systems was undertaken at Cranfield, UK, under the direction of C W Cleverdon, and is known as the Cranfield 1 project. The first Cranfield Study began in 1957 and was reported by Cleverdon in 1962.The project was designed to compare the effectiveness of four indexing systems, viz.

- An alphabetical subject catalogue based on a subject heading list.
- A UDC classified catalogue with alphabetical chain index to the class headings constructed.
- A catalogue based on a faceted classification and an alphabetical index to the class headings.
- A catalogue compiled by the uniterm coordinate index.

System Parameters

The study was involved 18,000 indexed items and 1200 search topics. The documents, half of which were research reports and half periodical articles, were chosen equally from the general field of aeronautics and the specialized field of high-speed aerodynamics.

Three indexes were chosen – one with subject knowledge, one with indexing experience and one straight from library school having neither subject background nor indexing experience.

Each indexer was asked to index each source document five times, spending 2, 4, 8, 12 and 16 minutes per document. One hundred source documents thus gave rise to a set of 6000 indexed items (100 documents X 3 indexers X 4 systems X 5 times). Each of these 6000 items was tested in three places, and therefore the system worked on altogether 18,000 (6000 X 3 phases) indexed items. The test was conducted in three phases with a view to find out whether the level of performance increased with increasing experience of the system personnel.

Significance

The results of the Cranfield 1 test contradicted the general belief regarding the nature of information retrieval system in many ways. The test proved that the performance of a system does not depend on the experience and subject background of the indexer. It showed that systems where documents are organized by faceted classification scheme perform poorly in comparison to the alphabetical index and uniterm system. It identified the major factors that affect the performance of retrieval systems, and 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 is the two most important parameters for determining the performance of information retrieval systems and that these two parameters are related inversely to each other.

The following findings of Cranfield are significant

- Indexing times over 4 minutes gave no real improvement in performance
- A high quality of indexing could be obtained from non-technical indexers
- The system operated at a recall rate of 70-90% and precision rate of 8 – 20 %
- A 1 % improvement in precision could be achieved at the cost of 3% loss in recall
- Recall and precision were inversely related to each other
- All four indexing methods gave a broadly similar performance

Cranfield test 2

The second Cranfield test was a controlled experiment that attempted to assess the effects of the components of index languages on the performance of retrieval systems. This study tried to assess the effect by varying each factor, while keeping the others constant. Altogether 29 index languages formed by the combination of concepts were tested on 1400 documents. The test was conducted on a collection of 1400 reports and articles collected from the field of high speed aerodynamics and aircraft structures.

5.2 MEDLARS

Performance of the Medical Literature Analysis and Retrieval System (MEDLARS) of the US National Library of Medicine was assessed during August 1966 to July 1967. The test was conducted on the operational database of MEDLARS, a database of biomedical articles, index entries being drawn from the MeSH. The objective of the MEDLARS test was to evaluate the existing MEDLARS system and to find out the ways for improvisation. The document collection available on the MEDLARS service at the time of the test consisted of about 7,00,00 items.

21 user groups were selected from the users community that would -
- supply some test questions;
- cover all kinds of subjects in the requests; and
- cover all categories of users.

The user group so selected provided 302 search requests. Each query was formulated in terms of MeSH by the system operator and searches were conducted. After completion of a search the sample output was sent to the users for relevant assessment. Photocopies of the articles, rather than mere reference, were supplied for the relevance assessment. The user was asked to mark each retrieved item using the following scales:

H1 – of major value;
H2 – of minor value;
W1 – of no value;
W2 – value unknown.

Precision of the searches were calculated from these figures with the following formula:
Precision ration = ((H1 + H2)/L) X 100
(Where L is the number of sample items retrieved)

5.3 SMART Retrieval Experiment

The SMART System was designed in 1964, largely as an experimental tool for the evaluation of the effectiveness of many different types of analysis and search procedures. Salton(1971) characterizes the system through the following steps of its function.

Overview of the functioning of SMART Retrieval System

It is used to

- take documents and search queries posed in English;
- perform a fully automatic content analysis of texts;
- match analysed search statements and contents of documents;
- Retrieve the stored items which are most similar to the queries.

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; and so on.

The following evaluation measures were generated by the SMART System:

- 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.
- Two global measures, known as normalized recall and normalized precision, which together reflect the overall performance level of the system.
- Two simplified global measures, known as rank recall and log precision, respectively.

5.4 The Stairs Project

In 1985, Blair and Maron (1985) published a report on a large scale experiment aimed at evaluating the retrieval effectiveness of a full-text search and retrieval system. This is known as STAIRS (Storage and Information Retrieval System) Study.

The database examined in the STAIRS study consisted of nearly 40,000 documents, representing roughly 350,000 pages of hard copy text used in the defence of a large corporate lawsuit. One important feature of STAIRS was that the lawyers who were to use the system for litigation support stipulated that they must be able to retrieve 75% of all the documents relevant to a given request. The major objective of the STAIRS evaluation was to assess how well the system could retrieve all the documents (and only those) relevant to a given request and measures of recall and precision were used for this purpose.

In STAIRS Project, precision was calculated by dividing the total number of ‘vital’, ‘satisfactory’, and ‘marginally relevant’ documents by the total number of documents retrieved. For the calculation of recall, a sampling technique was adopted. Random samples were taken and these were evaluated by the lawyers. The total number of relevant documents that existed in these subsets was estimated.

One of the reasons for the failure of STAIRS, as stated was, “It was impossibly difficult for users to predict the exact words, their combinations and phrases used in all or most of the relevant documents and only in those documents".

Now, we will see the last information retrieval system evaluation experiment in this module, i.e.

5.5 TREC: The Text Retrieval Conference

In 1991, the US Defence Advanced Research Projects Agency (DARPA), funded the TREC experiments, to be run by the National Institute of Science and Technology (NIST) in order to enable information retrieval researchers to scale up from small collections of data to larger experiments.

The objectives of the TREC experiments have been to

- Encourage retrieval research based on large test collections
- Increase communication among industry, academia and government
- Speed the transfer of technology from research labs to commercial products
- Increase the availability of appropriate evaluation techniques for use by industry and academia.

TREC is an annual workshop series aimed at building the infrastructure required for extracting relevant information from large volumes of electronic documents. The first TREC Conference was held in 1992 and research into improved text-retrieval methodologies has continued ever since. The document collection of TREC is known as the TIPSTER collection. TIPSTER reflects the diversity of subject matter, word choice, literary style, formats, and so on. Its primary collection is in gigabytes with over a million documents.

6. Summary

With this, we have completed the discussion on various aspects of information retrieval evaluation experiments and are now ready to conclude this module by saying that different users may want different levels of recall, like a person going to prepare a state-of-the-art-report (SOTAR) on a topic would like to have all the items so he / she will go for a high recall. On the other hand, the user wanting to know 'something' about a given topic will prefer to have 'a few items', and thus will not require a high recall. Here, the major problem is that users very often are unable to specify exactly how many items they want to be retrieved.

Recall assumes that all relevant items have the same value, which is not always true. The retrieved items may have different degrees of relevance and this may vary from user to user and even from time to time to the same user. Both recall and precision depend largely on the relevance judgments of the user. The judgement is quite subjective and there may be different degrees of the retrieved output.

Finally to conclude, one can say that the performance evaluation is crucial at many stages in information retrieval system development. At the end of development process, it is significant to show that the final retrieval system achieves an acceptable level of performance and that it represents a significant improvement over existing retrieval systems. To evaluate a retrieval system, there is need to estimate the future performance of the system.

7. References

1. Blair, D C and Maron, M E., An evaluation of retrieval effectiveness for a full-text document retrieval system. Communications of the ACM, 28(3), 1985. 289-99 pp.

2. Cleverdon, Cyril W. 1967. The Cranfield tests on index language devices. Aslib Proceedings 19, no. 6:173-194

3. Cleverdon, C W., Report on the first step of an investigation into comparative efficiency of indexing systems, Cranfield, College of Aeronautics, 1960

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

5. Fungmann, R., Subject analysis and indexing: theoretical foundation and practical advice by Robert Fungman, Frankfurt, IdeksVerlag, 1993

6. Keen, E. M., 'Evaluation parameters'. In: In: Salta G. (ed.), The SMART Retrieval System: experiments in automatic document processing. Englewood Cliffs, New Jersey: Prentice Hall, 1971, 74-111 pp.

7. Lancaster, F W. Information Retrieval Systems: Characteristics, testing and evaluation, 2nd edn, New York, John Wiley, 1979.

8. Salton, Gerald. 1971. Relevance feedback and the optimization of retrieval effectiveness. Ghap 15 in The SMART Retrieval System – Experiments in Automatic Document processing. Ed. G Salton. Pp. 324-336. Englewood Cliffs, New Jersey: Prentice Hall

9. Salton, Gerald, 'The SMART Project: Status report and plan'. In: Salta G. (ed.), The SMART Retrieval System: experiments in automatic document processing. Englewood Cliffs, New Jersey: Prentice Hall, 1971, 03-11 pp.

10. Salton, Gerald. 1971. Relevance feedback and the optimization of retrieval effectiveness. Ghap 15 in The SMART Retrieval System – Experiments in Automatic Document processing. Ed. G Salton. Pp. 324-336. Englewood Cliffs, New Jersey: Prentice Hall

11. Salton, G. 'Another look at automatic text-retrieval systems', communications of the ACM, 29(7), 1986, 648-656 pp.

12. Salton, G. and McGill, M. J., Introduction to modern information retrieval, Auckland, McGraw-Hill, 1983.

13. Swanson, D R. 'Some unexplained aspects of the Cranfield tests of indexing language performance', Library Quarterly, 41, 1971, 223-228 pp.

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

15. van Rijsbergen, C. J., Information retrieval, 2nd ed., London, Butterworth, 1979

16. Vickery, B. C. Techniques of information retrieval. London, Butterworth, 1970.