Department of Library and Information Science

Bharathidasan University
Tiruchirappalli-620024

Name of the Programme: M.Lib.I.Sc

Course - 2.2 Information Processing and Retrieval Systems

Course Code:P21MLS7

Unit-V: Evaluation of Information Retrieval Systems: Purpose – Criteria; Recall and Precision and steps in evaluation – Major Evaluation Studies – MEDLARS and SMART Retrieval.

Course Teacher: Dr.C.RANGANATHAN
Professor
e-mail: cranganathan72@gmail.com

EVALUATION OF INFORMATION RETRIEVAL SYSTEMS

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.

PURPOSE OF EVALUATION

- 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.

Purposes of Evaluation

- 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;

Cont.,

- 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.

LEVELS OF EVALUATION

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

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.

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.

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.

Quality Criteria:

- Coverage of database;
- Completeness of output (Recall);
- Relevance of output (Precision);
- Novelty of output; and
- Completeness and accuracy of data.

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.

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.
-

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.

Cont..,

- 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

Cont..,

- C.W. Clever don [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.

Recall

- 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:
- Recall Ratio Number of relevant document

Total number of relevant document

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

- 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 $(35 / 50) \times 100 = 70\%$.
-

Precision

- 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:
- Precision Ratio Number of relevant items

Total number of relevant document

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

- 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$(8 / 50) \times 100 = 16\%$.

EVALUATION EXPERIMENTS

MEDLARS Test

- Most of the evaluation studies were carried out on small collection.
- 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.
- 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.

Objectives

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

Cont..,

- 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).

Cont..,

- 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).

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.

Cont..,

- 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.

Cont..,

- 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.

Cont..,

- Three automatic language analysis procedures include in the SMART system,
- **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).

THANK YOU