Assignment on Evaluation of MEDLARS
(Medical Literature Analysis and Retrieval System)

Submitted To

Dr. Rekha Kishore
Department of Library & Information Science
Pondicherry University

Submitted by
Derhashar Swargiary
M.L.I.S. 3rd semester
Department of Library & Information Science
Pondicherry University

INTRODUCTION:

F. W Lancaster professor emeritus at the University of Illinois Graduate School of Library and Information Science and former NLM employee is the Pioneer in Evaluation of Early MEDLARS Systems. Lancaster contributions to the NLM in the early days of automated information retrieval have had a lasting impact on our information systems and services. He earned a reputation for greatness in the evaluation of information storage and retrieval systems, based in part on his early experience with a comprehensive evaluation of NLM's MEDLARS (Medical Literature Analysis and Retrieval System). The National Library of Medicine is the world's largest library of the health sciences and a component of the National Institutes of Health (NIH). NLM collects, organizes and makes available biomedical science information to scientists, health professionals and the public.

The evaluation of the MEDLARS Demand Search Service in 1966 and 1967 was one of the earliest evaluations of a computer-based retrieval system and the first application of recall and precision measures in a large, operational database setting. The use of computers for bibliographic retrieval systems was in its infancy, and many of the extant systems were small or experimental. Planning for the evaluation began in December 1965, when Lancaster joined the NLM staff as Information Systems Evaluator. Following completion of the MEDLARS evaluation, he developed NLM training programs in his roles as Deputy Chief of the Bibliographic Services Division and Special Assistant to the Associate Director for Library Operations. In 1970-1971, Wilf conducted an evaluation of the MEDLARS AIM-TWX system, an innovative experimental service that was the precursor of MEDLINE/Pub Med.

Evaluation of operating efficiency of MEDLARS:

NLM conducted a wide-ranging program to evaluate the performance of MEDLARS. It was conducted in 1966-67.

Purpose:

1. To study the demand search requirements of MEDLARS users.
2. To determine how effectively and efficiently the present MEDLARS services are meeting its requirements.
3. To determine factors adversely effecting the performance.
4. To discover ways in which more effectively or more economically user requirements can be satisfied.

The prime requirements of demand search users were presumed to relate the following factors:

1. The coverage of MEDLARS(i.e. the proportion of useful literature on a particular topic, within the time limit imposed, i.e. indexed into the system)
2. Its recall power(i.e. its ability to retrieve relevant documents)
3. The precision power(i.e. the ability to hold back the non relevant items
4. The response time of the system
5. The format in which the search results are presented
6. The amount of effort the user must personally expend in order to achieve a satisfactory response from the system.

The evaluation program was conducted in order to meet user requirements and tolerances in relation to these various factors and to determine MEDLARS performance in regard to the user requirements.

The two most critical problems faced in the evaluation of MEDLARS were

1. Ensuring that the body of test request was as far as possible representative of the complete spectrum of kinds of requests processed.
2. Establishing methods for determining recall and precision performance.

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).

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 analyzed 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: **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)

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.

References:

1. Jones, Karen Spark and Willet, Peter (1997) Readings in Information Retrieval. USA : Elsevier science.
2. Retrieved from http://www.egyankosh.com on October 13th, 2013.
3. Lal, C. and Kumar, K. (2005).Understanding basics of Library and information Science. New Delhi : Ess Ess Publication.