Lab 6: Evaluation of Information Retrieval Systems: Purpose and Criteria

Aim: Study Experiment for the Understanding of Evaluation of Information Retrieval Systems

Theory:

Evaluation is a systematic determination of a subject's merit, worth and significance, using criteria governed by a set of standards. It can assist an organization, program, project or any other intervention or initiative to assess any aim, realizable concept/proposal, or any alternative, to help in decision- making; or to ascertain the degree of achievement or value in regard to the aim and objectives and results of any such action that has been completed. The primary purpose of evaluation, in addition to gaining insight into prior or existing initiatives, is to enable reflection and assist in the identification of future change.

Evaluation is the structured interpretation and giving of meaning to predict or actual impacts of proposals or results. It looks at original objectives, and at what are either predicted or what was accomplished and how it was accomplished. So evaluation can be formative that is taking place during the development of a concept or proposal, project or organization, with the intention of improving the value or effectiveness of the proposal, project, or organization. It can also be summative, drawing lessons from a completed action or project or an organization at a later point in time or circumstance.

Evaluation is inherently a theoretically informed approach and consequently any particular definition of evaluation would have be tailored to its context - the theory, approach, needs, purpose, and methodology of the evaluation process itself.

1. A systematic, rigorous, and meticulous application of scientific methods to assess the design, implementation, improvement, or outcomes of a program. It is a resource-intensive process, frequently requiring resources, such as, evaluator expertise, labor, time, and a sizeable budget.
2. The critical assessment, in as objective a manner as possible, of the degree to which a service or its component parts fulfills stated goals'. The focus of this definition is on attaining objective knowledge, and scientifically or quantitatively measuring predetermined and external concepts

Evaluation of information retrieval system measure which of the two existing system perform better and try to assess how the level of performance of a given can be improved.

- Effectiveness and
- Efficiency

- Effectiveness it means the level up to which the given system attained its objectives. Thus in information retrieval system effectiveness may be measure of how far it can retrieve relevant information while with-holding non-relevant information.
- Efficiency means how economically the system is achieving its objectives. In an information retrieval system efficiency can be measured be factor such as cost. The cost factors are to be calculated indirectly. They include factor such as response time, time taken by the system to provide an answer. User effort, the amount of time and effort needed by a user to interact with the system and analyzed the output retrieved in order to get the correct information.

Lancaster state that evaluation of information retrieval system can be justified by the following three issues:

1. How well the system is satisfying its objectives
2. How efficiently it is satisfying its objectives and
3. Whether the system justified its existence.

PURPOSE OF EVALUATION

The main purpose of the evaluation is to focus on the process of implementation rather than on its impact, since this would be minimal after such a short time, accessing in particular the participatory approaches used to identify project beneficiaries and the communities role in implementing and monitoring the project.

To measure information retrieval effectiveness in the standard way, we need a test collection consisting of three things:

1. A document collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgments, standardly a binary assessment of either relevant or no relevant for each query-document pair.

Evaluation studies also investigate the degree to which the state goals have been achieved to which these can be achieved.

Swanson state seven purposes for evaluation:

1. To assess a set of goals, a programme plan, or a design prior to implementation.
2. To determine whether and how well goals or performance expectation are being fulfilled.
3. To determine specific reasons for success and failure.
4. To uncover principles underlying a successful programme.
5. To explore technique for increasing programme effectiveness.
6. To established a foundation of further research on the reason for the relative success of alternative technique and
7. To improve the means employed for attaining objectives or to redefine sub goals or goals in view of research findings.

Keen give three major purpose of evaluation for an information retrieval system:

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

EVALUATION CRITERIA

Evaluation of Information Retrieval is conduct into two different viewpoints.

1. Managerial view: when evaluation is conducted from managerial point of view it is called managerial oriented evaluation.
2. User view: when evaluation is conducted from the user point of view it is called user-oriented evaluation study.

Criteria for evaluation of information retrieval system:

- **Lancaster in 1971 proposed five evaluation criteria:**

1. Coverage of the system
2. Ability of the system to retrieve wanted items (i.e. recall)
3. Ability of the system to avoid retrieval of unwanted items (i.e. precision)
4. The response time of the system, and
5. The amount of effort required by the user.

- **Vickery advocate six criteria for evaluation of information retrieval system. He grouped into two sets as follows:**

Set 1

1. Coverage- the proportion of the total potentially useful literature that has been analyzed.
2. Recall- the proportion of such references that are retrieved in a search, and
3. Response time- the average time needed to obtain a response from the system.

Set 2

4. Precision- the ability of the system to screen out irrelevant references
5. Usability- the value of the references retrieved, in terms of such factors as their reliability, comprehensibility, currency and
6. Presentation- the form in which search results are presented to the user.

- **Cleverdon in 1966 identified six criteria for the evaluation of an information retrieval system. These are:**

1. Recall- the ability of the system to present all the relevant items.
2. Precision- the ability of the system to present only those items that is relevant.
3. Time lag- the average interval between the time the search request is made and the time an answer is provided.
4. Effort- intellectual as well as physical required from the user in obtaining answer to the search request.
5. Form of presentation- search output, which effects the user ability to make use of the relevant items and
6. Coverage of the collection- the extent to which the system includes relevant matter.

User-Centered Evaluation

User base evaluation is the most common evaluation system advocated by many information scientists. A criterion for evaluation of information retrieval system includes:

1. Recall
2. Precision
3. Fallout
4. Generality

1. Recall

The classic evaluative criteria of information retrieval system performance have been recall and precision, measures that were developed to evaluate the effectiveness of various types of indexing. Precision is defined as the proportion of documents retrieved that is relevant, while recall is defined as the proportion of the total relevant documents that is retrieved. These measures are expressed as a mathematical ratio, with precision generally inversely related to recall. That is, as recall increases, precision decreases, and vice versa.

The term recall refers to a measure of whether 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 if the system to retrieve all those items that is relevant to the given query. When the collection is large it is not possible to retrieve all the relevant items. Thus, a system is able to retrieve a proportion of the total relevant document in response to a given query. The performance of a system is often measured by recall ratio, which denotes the percentages of relevant items retrieved in a given situation.

The general formula for calculation of recall may be state as:

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

Example, if there are 100 documents in a collection that are relevant to a given query and 60 of these items are retrieved in a given search, then the recall is state to be 60% in other words the system has been able to retrieve 60% of the relevant items.

2. Precision

By precision we mean how precisely a particular system function. Precision is defined as the proportion of documents retrieved that is relevant. In precision the non-relevant items is discarded by the user.

The general formula for calculation of precision may be state as: Number of

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

Example, if in a given search the system retrieves 80 items, out of which 60 are relevant and 20 are non-relevant, the precision is 75%.

Thus recall related to the ability of the system to retrieve relevant documents, and precision related to its ability not to retrieve non-relevant documents. The ideal system attempts to achieve 100% recall and 100% precision is not possible in practice, because as the level of recall increase precision tends to decrease.

Example

Following example show the relationship between recall and precision of a given search. In a given situation a system retrieved a+b number of documents, out of which documents are relevant, and b documents are non-relevant. For example, c+d document are left in the collection after the search has been conducted. The number will be quite large, because it represents the whole collection minus the retrieved documents. Out of the c+d number, c document are relevant to the query but could not be retrieved, and document 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 the entire non- relevant document minus those that have been retrieved wrongly (b).

Lancaster suggests that these statistics can be represented in a 2 x 2 matrix, as shown below.

Recall-precision matrix

	Relevant	**Not-relevant**	**Total**
Retrieved	a (hints)	b (noise)	a+b
Not-retrieved	c (misses)	d rejected	c+d
Total	a+c	b+d	a+b+c+d

The system retrieves a relevant document along with b non-relevant documents. Thus following Lancaster it can be stated that a denoted hits and b denotes the noise. Now out of the remaining c+d document, the system misses c document that should have been retrieved, but it correctly rejected d document that are not relevant to the given query. The recall and precision ratio in this case can be calculated as

$R = [a/(a+c)] \times 100$ $P = [a/$

$(a+b)] \times 100$

The value of recall can be increase by increasing the value of a, that is by retrieving a greater number of relevant items. This can be achieved by increasing the number of retrieved document, but as the number of items retrieved increases, so also increase the likelihood of retrieval of non-relevant items that is b, which decreases the value of precision. Lancaster therefore states that recall and precision tend to vary inversely.

Limitations of recall and precision

- Different users may want different levels of recall. A person going to prepare a state-of-the-art report on a topic would like to have all the items available on the topics and therefore will go for high recall. Whereas, a user wanting to know about a given topic will prefer to have a few items and thus will not require a high recall.

- Another drawback of recall is that it assumes that all relevant items have the same value, which is not true. The retrieved items may have different degree of relevance and this may vary from user to user, and even form time to time to the same user. Both recall and precision depend largely on the relevance judgment of the user.

- Despite their apparent simplicity, these are slippery concepts, depending for their definition on relevance judgments which are subjective at best. Because these criteria are document-based, they measure only the performance of the system in retrieving items predetermined to be “relevant" to the information need.

- They do not consider how the information will be used, or whether, in the judgment of the user, the documents fulfill the information need. These limitations of precision and recall have been acknowledged and the need for additional measures and different criteria for effectiveness has been identified.

3. Fallout

Fallout ratio is the proportion of non-relevant items that has been retrieved in a given search.

4. Generality

Generality ratio is the relevant items that have been retrieved in a given search.

Salton state that larger the collection, the larger will be the number of non- relevant item in given query. Hence, an increase in the level of recall will cause a decrease in precision.

Retrieval Measure

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

Other Evaluation criteria

- Effectiveness
- Usability
- Satisfaction
- Cost

Effectiveness

It is obvious that the primary concern of an IR system is to retrieve information objects which meet the needs of the user. The two most commonly used measures of system performance are the recall ration and the precision ratio.

	Relevant	Not relevant
Retrieved	A	B
Not retrieved	C	D
Totals	A+C	B+D

The table suggests that in a search for documents or information objects there are four possible outcomes:

1. Some relevant documents are successfully received – call hits(A)
2. Some items that are not relevant are retrieved-noise (B)
3. The search fails to retrieve some relevant items- these are misses (C)
4. Some irrelevant items are not retrieved-these have been successfully dodged (D)

- Recall is a measure of a system's ability to retrieve relevant information.
- Recall=total relevant retrieved\total relevant in system

  = (a\a+c)*100

- Precision is a measure of the system's ability to suppress irrelevant or unwanted material.
- Precision = total relevant retrieved\ total retrieved

  = (a\a+b)*100

Indexing systems and search software should be designed to maximize both recall and precision: that is, to minimize noise and misses. Another measure, which has been quite widely used, is the fallout ration. Fallout is a measure of the ability of a system to suppress, or not to retrieve, non-relevant material.

- Fallout=total irrelevant retrieved\total relevant

  =(b\b+d)*100

This approach to the measurement of the performance of an IR system depends upon an acceptable means of determining the relevance of information objects to particular requests. It may be difficult to measure the total number of relevant document in an information retrieval system. Strictly using this measure involves examining every document in the system for its potential relevance to a specific search query.

- For web search engines such as Google this is clearly impossible

Usability

Usability: is a measure that embraces the interface through which the user interacts with the system, and also takes into account the user and their expectations, skills and experiences.

Satisfaction

There is no agreed definition of user satisfaction within the information science and information system communities. The original proponents of user satisfaction as a criterion for IR system evaluation assumed that it was correlated with search success. Applegate outlines three different models of searcher satisfaction namely:

- The material satisfaction model
- The emotional satisfaction- simple path model
- The emotional satisfaction- multiple path model

Cost

Users may experience costs in terms of any payment that they need to make for system or document access but the most significant cost is associated with the time that they expend in searching a system. Search algorithm, the options for the display of hits, the seamlessness of the stages in individual systems and interoperability between systems.