Information retrieval evaluation

9.3. Evaluating IR effectiveness

Overview

Evaluation is a critically important stage in IR system design. It should be built into the initial system development plan. Many IR applications are developed with little or no user input and no coherent plan for collecting evaluation data. How can one assess system effectiveness with no clear indicators of expected performance? The system design objectives should identify specific criteria against which to evaluate the system before the system is finished. If systems are evaluated only after implementation, it may be too late to make major changes.

This module builds on ideas about research discussed in other modules and provides some specific guidance for developing your approach to evaluating your own information organization system. It presents a description of evaluation research, considerations for assessing user satisfaction and IR performance, a list of factors affecting IR performance, and a discussion of measurement issues.

the determination of user satisfaction is based on users’ value judgments, while IR system performance is based on searching the system to test the quality of record structure, indexing, and so forth. For our purposes, both approaches address only system effectiveness (usefulness of IR results, benefits to users, etc.), not system efficiency (storage capacity, speed, costs, etc.).

Evaluation research

Most studies, including evaluation, follow the same process. Steps in blue italics are assigned for the IOP.

1. Define the research problem (e.g., the need to evaluate an IR system).
2. Establish research objectives (e.g., kinds of results to report).
3. Write research questions (in evaluation study, identify criteria for evaluation).
4. Develop research design.
   • Identify type(s) of data and data content to collect.
   • Choose data-collection method.
   • Obtain or develop data-collection instrument (e.g., questionnaire).
   • Determine respondent sample (source of data).
   • Determine schedule (who collects data, frequency and duration of data collection).
5. Collect data (in evaluation, conduct user satisfaction study and IR performance test).
6. Analyze the data and interpret the results.
7. Write research report.
Lancaster and Warner (1993) discuss three levels of evaluation:

- **Level I** is evaluation of **functional effectiveness**: criteria related directly to user-system interactions. Assessment is based on criteria such as ease of use, visual appeal, specificity and precision of indexing terms, and effectiveness of the system’s classification scheme.
- **Level II** is evaluation of **cost-effectiveness**: criteria related to internal system efficiency and costs, such as system speed and capacity.
- **Level III** is **cost-benefit evaluation**: the value of the system compared to costs of operating or using it. Level III asks whether a cheaper system could have been used.

The Level I criteria for functional effectiveness are most pertinent to the user emphasis in this course.

---

**Assessing user satisfaction**

User satisfaction judgments are based on users’ criteria for the quality of IR system performance, or the degree to which the system meets their needs and expectations. A criterion sets a standard. The questions one asks users should pertain to their experience with the system based on certain criteria.

**Criteria for effectiveness**

Lancaster and Warner (1993) are concerned with users’ perceptions of IR system performance. They suggest a wide range of user criteria:

A. Cost criteria
   1. Monetary cost to user
   2. Other, less tangible, costs
      a. Effort involved in learning to use system
      b. Effort involved in actual use
      c. Effort involved in retrieving documents
      d. Form of output provided by the system

B. Time criteria
   1. Time from request to retrieval of references
   2. Time from request to retrieval of documents
   3. Other, e.g., waiting to use system

C. Quality criteria
   1. Coverage of database
   2. Completeness of output (recall measure)
   3. Relevance of output (precision measure)
   4. Novelty of output (new to the user)
   5. Completeness and accuracy of data

Several of these criteria as important for you to consider as you develop your evaluation plan for the IOP.
System usability

Cost criteria fall into an area called usability. The designer might ask:

- How easily do users interact with the system?
- What features do they use or not use?
- Are instructions accessible? Easy to understand?

Cost criteria are important factors. Many online databases such as DIALOG charge fees that make the system unaffordable to many users. Another cost is the time invested in training to learn to use a cumbersome interface as well as the ongoing effort required to conduct searches. These are often considerations in corporate libraries, where the amount of money and time needed to retrieve data determine whether a system is actually used.

Time criteria are also critical. System response time is a prime consideration because users express definite preferences for faster systems. Much research has been done on user patience. In the web environment, the threshold of patience is measured in seconds.

Similarly, the amount of time it takes to obtain actual documents once citations have been retrieved from the system is important. Many libraries that offer search services on DIALOG and other commercial fee-based systems warn users that the library may not have all the documents represented in the database records. This is because online information services index entire domains of bibliographic literature, not the holdings of individual libraries. It is common for a DIALOG search to identify a set of documents that would be useful to the user but that are not available at the search site. Some online database vendors, however, now offer ways to indicate which documents identified in a search are available in a local library.

Relevance criteria

Quality criteria fall into an area generally called relevance. Some researchers view relevance as an umbrella criterion for evaluating IR system output and view other criteria as the reasons underlying users' judgments of "relevant" or "not relevant." The list by Lancaster and Warner (1993) above suggests the relevance criteria of novelty, completeness, and accuracy of data. These and other criteria, including aboutness, currency, credibility, understandability, availability, clarity, and style are suggested in the LIS literature (see Barry & Schamber, 1998).

Relevance criteria are useful for evaluating systems because they can direct the system designer to specific features to improve. Suppose a user judges a document represented in a database record to be not relevant or not worth pursuing. The researcher asks the user to explain that judgment. The user says the document is on the topic (aboutness), but is outdated (currency). The user then examines the full-text document and says it is worthwhile after all because the author works at a major research institution (credibility). These responses confirm the importance of representing the attributes of topic and date in the record and further suggest representing author affiliation.

It should be noted that the meaning of the concept of relevance, like that of information, is much debated in the LIS literature. See measurement issues below.
User abilities

User needs analyses conducted at the outset of the system design process suggest more considerations related to users' levels of knowledge and skills:

- Do users have the system and information-seeking skills to plan good search strategies?
- Do users have the vocabulary to formulate queries successfully?
- Do users have the domain knowledge to make reasonable relevance judgments?

Measuring IR performance

IR performance tests consist of conducting searches and recording how well the system responds. Ideally, the queries are based on requests from actual users.

LIS has dozens of performance measures, but two that are fundamental to the field are precision and recall. A measure is a quantifiable standard or unit for determining extent, quantity, or degree. Precision and recall are comparative measures that are calculated as ratios.

Calculating precision and recall

Suppose the IR system retrieves 12 documents (or document representations in records) out of a total of 1,000 documents, as follows:

<table>
<thead>
<tr>
<th>Retrieved</th>
<th>Relevant</th>
<th>Not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Not retrieved</td>
<td>2</td>
<td>986</td>
</tr>
</tbody>
</table>

Precision assesses the ability of the system to retrieve only the relevant documents:

\[
\text{Precision} = \frac{\text{Total relevant documents retrieved}}{\text{Total documents retrieved}} \times 100 = \frac{8}{12} \times 100 = 67\%
\]

Recall assesses the ability of the system to retrieve all the relevant documents it contains:

\[
\text{Recall} = \frac{\text{Total relevant documents retrieved}}{\text{Total relevant documents in system}} \times 100 = \frac{8}{10} \times 100 = 80\%
\]

Precision and recall results

How good are these scores? Not very! The truth is that scores range widely due to many factors (see below). Rarely does a system score 100% in either precision or recall, although scores in the 80% range may be acceptable. However, in cases where librarians are aiding users in the search process, the notion of acceptability may differ considerably from the librarian to the user. The librarian may be disappointed with any results below 80%, while the user may be happy with 20%. In other words, human judgments are subjective—a reality that is central to various measurement issues (see below).
Factors affecting IR performance

Precision and recall are only performance measures; they do not diagnose the reasons for certain levels of performance, which are more important than the scores per se.

The example above raises obvious questions. With regard to precision, why were four nonrelevant documents retrieved? That kind of performance wastes money in systems where the user pays fees based on number of documents retrieved, rather than on relevant documents retrieved. With regard to recall, why were two relevant documents not retrieved? Those documents could have contained information that was vital to helping the user.

Some factors that might have contributed to these results are:

- Poor indexing, in that inappropriate terms were applied to the four articles
- Poor semantics in search aids, such as thesauri; that is, thesaurus terms that were not specific enough, or that were simply erroneous
- Inaccuracy in the system’s search mechanism, perhaps based on faulty specifications
- Inaccuracy in the user's query, perhaps stemming from a low level of knowledge or ability

These are just a few of many possible factors to consider when analyzing performance test results. Taylor (2004, chapter 5), Borgman (1996), and Lancaster and Warner (1993) provide many more ideas. A good starting place is this summary by Lancaster and Warner (p. 204):

<table>
<thead>
<tr>
<th>System factors</th>
<th>Database</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time lag and frequency of update</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Indexing and vocabulary policies</td>
</tr>
<tr>
<td>Index language</td>
<td>Specificity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Level of coordination</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amount of explicit structure</td>
<td></td>
</tr>
<tr>
<td>Indexing</td>
<td>Exhaustivity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Specificity of term assignment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td></td>
</tr>
<tr>
<td>Search strategy</td>
<td>Exhaustivity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td></td>
</tr>
</tbody>
</table>

Clearly indexing is a heavy player in system performance. The module on how indexing languages affect retrieval relates the indexing criteria of specificity and exhaustibility to the precision and recall of system performance. Precision and recall are so closely related to indexing, in fact, that they are also called measures of indexing effectiveness.
Lancaster and Warner (1993, p. 204) also summarize factors related to indexers and searchers:

<table>
<thead>
<tr>
<th>Human factors</th>
<th>Indexing</th>
<th>Searching</th>
<th>Screening [filtering]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consistency</td>
<td>Consistency</td>
<td>Consistency</td>
</tr>
<tr>
<td></td>
<td>Subject expertise</td>
<td>Subject expertise</td>
<td>Subject expertise</td>
</tr>
<tr>
<td></td>
<td>Indexing experience</td>
<td>Indexing experience</td>
<td>Indexing experience</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cognitive style</td>
<td>Screening experience</td>
</tr>
</tbody>
</table>

Here you can see the significance of types and levels of knowledge of the searcher as well as the indexer. These are human factors that should be discovered in the initial user needs analysis.

**Needs analysis factors**

It is worthwhile considering characteristics of users from the user needs analysis conducted at the beginning of the system design process. The levels of various kinds of knowledge directly affect users’ ability to formulate queries effectively. Their information need situations also influence how they search. The importance and pressures of the situation may affect users’ motivation levels and emotional states. For example, a user who has little time to search but more time later to browse through retrieved results may settle for low precision and high recall, while another user under no time constraints may demand very high precision and exhaustive recall.

**Measurement issues**

No measure is perfect, and precision and recall are particularly troublesome. The first problem concerns the **relevance judgments** in these measures. What's wrong with relevance judgments? They are:

- Extremely subjective
- Change constantly
- Cannot be known in advance
- Ultimately depend on judgments of individual end users

This means that relevance judgments are difficult to predict because they vary widely depending on the judges and the circumstances of the search. However, they are easy to collect and count or measure with scales (e.g., degrees of relevance from 1 to 10). Further, researchers have found relevance judgments to be remarkably consistent and stable for similar kinds of users and settings. Interestingly, the same kind of consistency applies to judgments based on relevance criteria (accuracy, currency, etc.), so they can be handled in the same way as relevance judgments.

The second problem concerns the denominator of the recall ratio, or the **recall base**. How can one know how many relevant items are in the system, especially the relevant items that are not retrieved?

One solution is to examine every item in the database and judge its relevance. Obviously that's not practical in a database of any size. Another solution is to **estimate the recall base**. Two methods:

- Create a test collection of documents with predetermined relevance judgments, based on a particular query.
• Conduct overlap tests. Run the same query multiple times in the same database, or using different search mechanisms in the same database, or on different databases with parallel content. Compare the results: find documents that overlap, or that were retrieved in every search. Assume that these are the relevant documents for the recall base.

In estimating recall as with determining relevance, it also matters who the judges are: domain experts, expert searchers, or end users.

Summary

The steps of the research process are well-established. One begins by defining the research problem because research is expensive and there should be a sound reason for doing it. The research problem drives the objectives, design, data collection, analysis, and research report. Deciding on evaluation criteria before developing the data-collection instrument avoids finding the right answer to the wrong question. Ideally, the designer should:

• Build evaluation into the overall system design plan.
• Think carefully about the research questions or evaluation criteria.
• Know what data to collect and how they will be used.
• Choose appropriate data-collection methods.
• Try to use more than one data-collection method.
• Choose the sample carefully.
• Pretest the data-collection instrument (e.g., questionnaire).
• Set reasonable boundaries.
• Keep clear, precisely documented records.
• Analyze data as soon as possible.
• Draw conclusions carefully based on the evidence of the data.

This module cannot end without emphasizing the importance of relevance, precision, and recall. These may be imperfect measures, but they are central concepts in library and information science. These terms are used throughout the LIS literature for their meanings, beyond evaluation research and often with no attempt at quantification.

Cites & sites


