A Cost Model for Evaluating Information Retrieval Systems

This paper develops a mathematical model for use in studying how to minimize the cost of operating a mechanized retrieval system. Through the use of cost analysis, the model provides a method for comparative evaluation between information retrieval systems. The cost model divides the costs of a retrieval system into two components: system costs and user costs. In addition, it suggests that a trade off exists between the performance level of the system and the combination of user and system time that is expended in working with the system. With this approach it is possible to determine the allocation of user and system. This minimizes the total cost of operating the system. This allocation is done for a given performance level and for a given cost per unit of user and system time.

MICHAEL D. COOPER
School of Librarianship
University of California
Berkeley, California 94720

Introduction

One method whereby information retrieval systems can be evaluated is through the use of costs. In order for accurate comparative analysis to be conducted, the cost methods used must be consistent and comprehensive. The model developed in this paper has two facets. It develops equations for a total cost of operating the system and thus allows comparative evaluation between other systems. Further, once the cost equations for the system have been presented, it shows that an optimal division can be made between those functions that the user should perform and those that the system should perform in optimizing the search process.

Most information retrieval systems that have been designed to date use a technique of comparing a query representation with stored document representations in order to retrieve documents that satisfy the request. Those documents whose representations are ‘closest’ to the query are retrieved. It is hypothesized that this type of comparison process is not sufficient to insure the ‘best’ operation of the retrieval system. Not only should the system perform matching but it should take into account:

1. The cost of the search to the system. This implies that the system provides a service to the user at a specific price.
2. The cost of the search to the user. This suggests that the user places a value on the time and effort he spends using the system. This quantity is in addition to the amount that must be paid to use the system.
3. The benefit that the patron can gain by using the system.
4. The most economic division of effort between the user and the system in accomplishing the user's search objective.

In this model, the user formulates a query and the system decides, on the basis of control parameters and previous experience, the retrieval rule most likely to yield documents relevant to the user's request.

Retrieval Activities

Acquisition of information from an automated retrieval system involves an interaction between the user and the computer. As with any man-machine interaction, the more demanding and more sophisticated the user is in his requests, the greater will have to be the system effort to achieve the desired goal. In this section the trade off between user and system effort is explored and a schema is developed for analyzing it.

It is possible to distinguish three stages in the interaction of a user with the retrieval system. The process begins with pre-search activities. For the user this involves determining what is to be asked of the retrieval system and mapping the request into the system’s formal query language. Since it is unlikely that the user will enter the correct query the first time he tries (perhaps due to syn-
tax or spelling errors) there will be some user-system dialog involved in putting the request into a form acceptable by the system.

Query negotiation may be a simple process of correcting syntax, as noted above, or can be more elaborate. For example, the system may be in a position to aid the user in query formulation by the use of a thesaurus and/or a word frequency list. The thesaurus is employed to tell the user the generality or specificity of the words that are present in the query. This allows the user to broaden or narrow the scope of the query depending on the search objective.

Through the use of word frequency distributions, the system can tell the user the extent to which a particular term has been used as, for example, an index term in the indexing of the document collection. When this information is supplied, the user can judge the quantity of material that will be retrieved for a given request.

The second stage in the retrieval process is the search activity. It is in this stage that the comparison of the formal representation of the user's request is made to stored document representations. Consequently the system's effort in this stage of the process is greatest, and the user is resigned to waiting for the results to be displayed.

The final stage is concerned with post-search activities. The retrieval system has predicted which documents satisfy the request and now must display the output for the user. With the documents displayed in front of him, the user then evaluates the retrieved documents in terms of their relevance to his information need. The system uses this information to calculate a performance measure for the search. In addition, a feedback mechanism operates to revise the search procedure in light of the user's satisfaction. Table 1 summarizes the user and system activities.

### Document Representation

A number of developments suggest that useful information is being ignored when the only keys that can be used to retrieve documents from a file are author, title and index terms. Consider, for example, the MARC format for the interchange of bibliographic information on magnetic tape (15). It is suggested that a large number of the fields in this format are useful means of accessing bibliographic information. Kessler's work on bibliographic coupling suggests the usefulness of storing document citations in the file to allow citation indexing and searching (15). Another possibility is that of including non-content information about the document (i.e. context information) in the file (10).* Context information, as distinguished from content information is that material that describes characteristics of the author (his academic background, degrees, current research interests, employer, etc.), the journal in which the paper was published, the editor, the references, etc.

Thus one can see that there are a number of alternate document surrogates that can be stored in the file and used for retrieval. In this paper these alternate forms are called *representations* of a document. Examples of document representations include the title, author(s), abstract(s), full text, index terms, citation, cluster center descriptions, etc. of a document.

### User-System Interaction

The user has a number of alternative strategies that he can employ in his information seeking behavior. Instead of employing an information retrieval system, the user may browse through his personal library, consult a co-worker, phone a friend, or consult a reference librarian.

The user's time is an economic quantity. Given the cost of this time and the fact that there are a number of information seeking alternatives, Simon's concept of *satisficing* appears particularly applicable (9, 14).** The user pursues a selected information seeking strategy until the cost incurred exceeds the level of satisfaction received. At this point another strategy could be adopted or the process could stop with the user *satisfied*.

It is suggested that a number of variables are tested by the user in deciding whether his cost of a particular strategy has exceeded the benefit. These variables include the time the user spends at the console of the informa-

* Preliminary evidence using a small corpus has not indicated the usefulness of this data for searching, but these results are based on a too small a sample to be considered indicative or final (18).

** Baker and Nance (1) have also suggested the applicability of this idea to information seeking behavior.

### Table 1. User and System Activities

<table>
<thead>
<tr>
<th>User Activity</th>
<th>System Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determine information</td>
<td>Syntax check of query</td>
</tr>
<tr>
<td>need</td>
<td>Thesaurus lookup</td>
</tr>
<tr>
<td>Enter query</td>
<td>Query term frequency analysis</td>
</tr>
<tr>
<td>Query negotiation</td>
<td>Map query into formal language</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pre-Search</th>
<th>Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wait</td>
<td>Select comparison method (retrieval</td>
</tr>
<tr>
<td></td>
<td>rule)</td>
</tr>
<tr>
<td></td>
<td>Search file</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Post-Search</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Read output</td>
<td>Display output</td>
</tr>
<tr>
<td>Mark relevant material</td>
<td>Calculate performance measure</td>
</tr>
<tr>
<td>Use relevant material</td>
<td>Revise strategy and/or query with</td>
</tr>
<tr>
<td></td>
<td>feedback</td>
</tr>
</tbody>
</table>

Journal of the American Society for Information Science—September—October 1972  307
tion retrieval system, the time required to map the request into the retrieval system's query language, and the waiting time until the results of the search are displayed.

Another group of variables that determine user cost, more specifically relate to man-machine interaction (3, p. 57). The design of the console, the flexibility of the programs in allowing the user to go as slow or as fast as he wants in a dialog with the machine, all contribute to his willingness to use the machine and the value that he places in the retrieval system.

Finally, the user is influenced by the results that he obtains from the system. It is in this area that measures of retrieval effectiveness are valuable. They provide the user with a measure of the degree to which he is satisfied with the system.*

The cost that the user assigns to the employment of the system is a combination of the factors described in this section. If the system does not satisfy the user requirements, it is not used. Thus the user cost-benefit function is a constraint that is considered by the system. This is accomplished in a number of ways. For a given query the retrieval system predicts for the user the cost of the search and the time required to perform it. The user can then broaden or narrow his request given his budget constraint.

- Search Methodology

A complex process is undertaken when a retrieval system attempts to find material relevant to a user need. The model of user interaction in the previous section suggests that information seeking patterns vary according to user cost and benefit. This section elaborates the problem further by suggesting the need to pick an optimal combination of search comparison method and document representation for the search.

There are a number of search comparison methods that can be used in an information retrieval system. These included the simple matching technique, in which the query is compared with the document representations and the degree of similarity between the two is calculated. Extensions of the methodology include the elaboration of the terms in the query with related terms to effect associative searching. Alternatively it is possible that instead of looking at every document representation in the file, clustering could be employed, and only cluster center representations be compared to the query representation.**

Thus there are a number of different strategies that can be employed. Traditionally one or perhaps two of these have been implemented in a given operational system. In the proposed system, however, all the comparison methods will be available to the user. This is done on the theory that a specific strategy will have certain properties that make its use advantageous in specific circumstances. For example, comparison method \( X \) may be found to be extremely exhaustive in its search for documents meeting the query objective. Method \( Y \), on the other hand, may be particularly useful when searching for one specific subject. Then the user will have the ability to decide which strategy to employ. Alternatively he may rely on the system to pick the strategy, or may be forced to pick one because of a requirement for a specific performance level or because of a given budget constraint.

In addition to the need to pick a particular search comparison method or retrieval rule for a specific purpose, the user has the ability to select a document representation that will be used to compare the query against.

A document has associated with it a number of representations. These surrogates include index terms, abstracts, subject headings, etc. When a search is made, the retrieval system picks a particular representation to compare the query against. For example, if it is desired to find an article written by author \( W \), a search of the author representation is conducted.

The search for a particular author is the easiest case for the system to handle. This is so because the system knows which representation to compare against. But take the case of a request which is in the form of a set of words characterizing a chosen subject. Here the problem is more complex because there are a number of representations that could be used for the comparison, such as the document index terms or the document abstract.

The user then has the flexibility to decide which representation will be used in his search or alternatively to let the system decide. If a broad survey of literature is desired, there may be more benefit in using subject headings than author assigned index terms for the search comparison. On the other hand, if a very specific question is posed, the query may only be able to be answered through searching the abstracts or full text of a document. Here again there is a trade off between the degree of generality or specificity required in the search and the cost and benefit of conducting it. By allowing this flexibility to exist, the system stands a better chance of satisfying a variety of users.

- Retrieval Model

Optimal performance of a retrieval system requires that both system and user resources be considered in determining an operating level. This section considers the issues involved in selecting such a level. It also sketches the manner in which the system's strategy can be modi-
fied in the light of changes in user assessment of system benefits. The total cost of a retrieval system's operation for a query, \( C_T \), is the sum of the system cost and the user cost. That is

\[ C_T = c_u + c_s, \]

(1)

where \( c_u \) is the cost per unit of user time and \( t_u \) is the amount of user time required for a given search. Similarly \( c_s \) is the cost per unit of system time and \( t_s \) is the amount of system time required for a given search.

It is presumed that retrieval system performance can be characterized by a measure called \( P \). It is believed that \( P \) is a complex function that may include variables such as those used in calculating measures of retrieval effectiveness. For this model the measure is considerably simplified so that it has the form

\[ P = f(t_u, t_s). \]

(2)

That is, the performance is a function of the amount of user time and system time expended on the search. A number of more specific formulations are possible. For example, the relation could be

\[ P = t_u t_s. \]

(3)

This is the form of an isoquant curve from economic theory. (See Figure 1). Each point on an isoquant curve represents the maximum output that can be produced with a given combination of inputs. Each of the curves of Figure 1 show combinations of user and system time that yield the same level of performance, \( P_t \).

Very little information is available about the precise shape of the performance curves. It could very well be that the curves are L shaped or even straight lines. However, assume for the following discussion that the performance can be characterized by an equation such as (3). Then it is possible to solve equations (1) and (3) to find the optimal level of \( t_u \) and \( t_s \) that minimizes total cost for a given performance level.

First rewrite equation (3) as

\[ t_u t_s - P = 0. \]

Then using the Lagrange multiplier \( \lambda \), form the equation

\[ F_\lambda = t_u c_u + t_s c_s + \lambda (t_u t_s - P). \]

(5)

The application of partial differentiation yields

\[ \frac{\partial F_\lambda}{\partial t_u} = c_u + \lambda t_s = 0 \]

(6)

\[ \frac{\partial F_\lambda}{\partial t_s} = c_s + \lambda t_u = 0 \]

(7)

\[ \frac{\partial F_\lambda}{\partial \lambda} = t_u t_s - P = 0. \]

(8)

Then

\[ t_u = -c_u/\lambda \]

(9)

and

\[ t_s = -c_s/\lambda. \]

(10)

Substituting equations (9) and (10) into (8) yields

\[ (-c_u/\lambda)(-c_s/\lambda) - P = 0 \]

(11)

\[ \lambda = \sqrt{(c_u c_s)/P}. \]

(12)

Then the optimal \( t_u \) is

\[ t_u^* = \sqrt{(c_s/c_u)} P \]

(13)

and

\[ t_s^* = \sqrt{(c_u/c_s)} P. \]

(14)

Thus the optimal allocation of resources depends on the performance and the cost coefficients of each of the two resources.

**SYSTEM RESOURCES**

System activity is divided into three areas: pre-search, search, and post-search activities. The system cost, \( c_{st} \), of equation (1) can be written

\[ c_{st} = c_{pre-search} + c_{search} + c_{post-search}. \]

(15)

The variables represent the costs per unit of time multiplied by the time used for each of the three activities.

The pre-search system cost per unit of time is given by

\[ c_{pre-search} = \alpha Ch + \beta CPU + \gamma Core. \]

(16)

Here \( Ch \) is the cost of the computer channels, \( CPU \) is the cost of the central processing unit and \( Core \) is the cost of the core storage per unit of time. The value of \( \alpha, \beta, \) and \( \gamma \) represent the utilization rates of each of the components for the pre-search activity.

System search cost per unit of time is a function of the search comparison method employed and the document representation used. Thus

\[ c_{search} = c_{comp. method} + c_{representation}. \]

(17)
Finally, the post-search cost is given in a form analogous to equation (16):

\[ c_{\text{post}} = \alpha \cdot Ch + \beta \cdot CPU + \gamma \cdot Core. \quad (18) \]

Each of the search comparison methods or retrieval rules employed in a retrieval system is presumed to have a cost associated with it. No general statements can be made about the exact formulae for the cost of a comparison method because the cost is highly dependent on the way in which a strategy is implemented in a computerized system. For example, the internal representation of the query and the documents will influence the cost. Nevertheless, it is possible to suggest the form that such an equation might take.

The comparison cost will be a function of the number of terms in the query, the number of logical operators in the query (e.g., 'and', 'or', 'not'), the number of document representations in the file, and the number of words in each representation. Additionally, the cost will depend on the computer resources used: the central processing unit, core storage and the channels. Finally, the amortized cost of programming a particular comparison method will have to be included. For associative searching, the association files need to be constructed, and for comparison on cluster centers, the clustering will have to be performed.

The retrieval system must calculate costs for document representations stored in the system. Total cost for a representation is made up of three components: creation cost, storage cost, and processing cost.

When documents are received at an information center a certain amount of pre-processing is performed before the document can be stored in the system's data bank. For example, the document may have to be indexed, assigned subject classifications, abstracted, etc. In addition, if it is not already in machine readable form, the conversion will have to be performed. All these functions are considered part of the information center's cost of creation of a document surrogate, \( c_{\text{create}} \).

No uniform method exists for accurately determining retrieval system costs or document surrogate costs. Surveys by Landau (6), Olsen (11) and Penner (15) have summarized the cost work that has been performed to date. Leimkuhler and Cooper have proposed the use of standard cost accounting techniques (8) and mathematical models (7) as a solution to the problem. Keith (4) has also outlined a cost model.

The second surrogate cost component is storage cost, \( c_{\text{store}} \). There are a number of variables that determine this cost: the rental cost of the computer storage device, the proportional cost of the control unit for the device, the capacity and utilization of the device, and the number of characters in the representation.

The final component of the surrogate cost function is the processing cost, \( c_p \). Two types of processing are performed in the retrieval system: retrieving records from the file and creating and maintaining the file. In addition, there are three components of the processing cost: the central processing unit cost, the channel cost, and the core storage cost. The basic form of the processing cost equation is the same as in equation (16):

\[ c_{\text{process}} = \alpha \cdot Ch + \beta \cdot CPU + \gamma \cdot Core. \quad (19) \]

While the costs of each of the computer components remain constant in the processing cost equation, the values of \( \alpha, \beta \) and \( \gamma \) vary depending on whether updating or retrieval is being performed.

To summarize, the cost of a document representation is

\[ \text{Representation} = c_{\text{create}} + c_{\text{store}} + c_{\text{process}}. \quad (20) \]

Preliminary analysis suggests that the cost differences between document representations in the same class (e.g., the index terms assigned to document number one and those assigned to document number two) may be so small as to minimize the need for cost computation for each representation of each document. Instead costs could be computed for each representation class. This follows the approach of standard costing suggested earlier (8).

**System Resource Allocation**

When the user begins a dialog with a system employing this resource allocation scheme, he specifies a desired performance level and a budget constraint. Then using the solutions in equations (13) and (14) the system is able to divide the user's fixed budget between system activities and user activities. This section explores possible approaches whereby the system can divide its time between pre-search, search, and post-search activities.

It was postulated earlier in this paper that a relation exists between user time, system time, and system performance. It is also possible to establish a relation between performance, search comparison method, and document representation.

\[ P = f(\text{comparison method, document representation}) \quad (21) \]

Using the equations reflecting representation and comparison method costs per unit time, it is possible to arrive at an optimal choice of document representation and comparison method that minimizes system search cost, \( c_{\text{search}} \), for a given performance level. This is done in a manner similar to that used previously in this section.

Once the comparison method and document representation have been selected, the search cost is determined using equation (17). The search, \( t_{\text{search}} \), is calculated using the average number of documents to be searched or the average number of index entries to be searched. The user is charged based on the average cost figure, and variances are accumulated and at periodic intervals are used to readjust the cost coefficients. In this manner the total search cost,

\[ c_{\text{search}} = t_{\text{search}} \cdot c_{\text{search}}, \]

is determined.
The remaining problem facing the system is to allocate the remaining funds between the pre-search and post-search activities. No precise rules can be given for this. However, it is possible to delimit the values of $t_{pre-s}$ and $t_{post-s}$ that are feasible. For instance the structure of the system may be such that pre-search activity requires a minimum of 'a' time units and cannot use more than 'b' time units no matter what performance is required. Then

$$a \leq t_{pre-s} \leq b$$

Similar bounds could be developed for post-search time requirements. As an additional future stage, it would be desirable to determine if a relation could be established between the system performance and pre- and post-search time allocations.

**User Resources**

The second category of resources that are used in the retrieval process is user resources. The total cost of these factors is

$$c_u t_u$$

where $c_u$ is the cost per unit of user time and $t_u$ is the amount of user time expended for a given query. As before, a distinction is made between the three activities: pre-search, search and post-search. Then the total user cost is given by

$$c_u t_u = c_{pre-u} t_{pre-u} + c_{search-u} t_{search-u} + c_{post-u} t_{post-u}.$$  \hspace{1cm} (23)

Here $c_{pre-u}$, $c_{search-u}$ and $c_{post-u}$ are the per unit costs of the user's time for each activity, and $t_{pre-u}$, $t_{search-u}$ and $t_{post-u}$ are the number of time units of each activity used for a given search.

It should be noted that equation (23) is again a simplification of the actual situation. Post-search activity time for the user is in actuality a function of the amount of time spent in pre-search activity.

In this model it is assumed that when the user is not availing himself of system services, the system can service other users or other jobs. Thus it is assumed that the system is never idle, or if it is the user does not pay for the idle system time. On the other hand, if the system is heavily loaded with other tasks, the user may have to wait for a response to his dialog with the system. This suggests that equation (23) should be modified as follows:

$$c_u t_u = c_{pre-u} (t_{pre-u} + \theta_{pre-s}) + c_{search-u} (t_{search-u} + \theta_{search}) + c_{post-u} (t_{post-u} + \theta_{post-s}).$$  \hspace{1cm} (24)

The variable $\theta$ represents the additional time that the user must wait for the system for each of the activities. For example, search activity will require a small amount of user time perhaps to initiate the searching once the query has been accepted. Then the user will have to wait $\theta_{search}$ units of time until the system completes the search, where

$$\theta_{search} \geq 0.$$  \hspace{1cm} (25)

The values of $c_u$ and $t_u$ in equation (24) are a function of the qualifications of the user, $U$. That is

$$c_u = f(U).$$  \hspace{1cm} (26)

Many different people presumably use an information system. Each person most likely values his time at a certain price. The user cost per unit of time is related to this assessment of value by the user. For example, a senior member of an organization will command a higher salary than a clerk. If both of them use the retrieval system, the allocation of resources between user and system will vary depending on the $c_u$ and $c_s$ values.

Similarly the time that a user spends at the console will depend on his experience with the system as well as his qualifications. Thus

$$t_u = f(U).$$  \hspace{1cm} (27)

It is possible to conceive of a situation in which users with similar $c_u$ values have different $t_u$ values simply due to extensive practice with the system or a more agile mind. Equation (27) is intended to reflect this disparity.

**Cost Model Evaluation**

The cost model that was developed divides the activities involved in the retrieval system operation in several ways. The first division involves an allocation of effort for a given search between the user of the retrieval system and the system itself. That is, either the user can spend time and effort in correctly specifying his query, understanding what kind of material is in the document file, how terms are related in the document file, etc., or a dialog between the user and the system can take place in which this information is established by negotiation. The negotiation process shifts some of the effort from the user to the system. Thus there is a trade off between the cost to the user and the cost to the system for the search. In addition to the division between user and system effort, the model divides the total time during which an interaction is taking place into three parts: pre-search activity, search activity and post-search activity. During the pre-search phase the user negotiates the query with the system; during the search phase the user waits while the system searches the file; and during the post-search phase the system displays the output for the user.

As with all models, the cost model of a literature searching system is a simplified description of the real situation. There are a number of deficiencies in the model. The performance measure that is used in determining the optimal allocation of effort between the user and the system is simplified. The measure only considers performance as a function of user and system time. In all probability, a performance measure is much more complex than this cost model assumes. Another deficiency is that the model has not yet been verified with operating data. Aside from these problems it is believed that the framework that the model presents is a useful way of evaluat-
ing retrieval systems as well as a meaningful method for arriving at an optimal allocation between user and system resources.

References