The Relationship Between Relevance and Reliability in Internet-based Information and Retrieval Systems

Daniel E. O'Leary
3660 Trousdale Parkway
University of Southern California
Los Angeles, CA 90089-1421

oleary@rcf.usc.edu
213-740-4856

Abstract

In this paper, internet information retrieval systems are modeled as an intermediate step between users of the system and the original or expected information. For example, use of a search engine will generate a "report" providing a list of URLs and a corresponding brief description that may or may correctly describe the label being searched. In particular, there is a probability that the report of the content (x#) and the actual content (x) are not the same, where Pr(x#|x) is used to define the reliability of the system.

If reliability is not considered in internet systems, non optimal behavior may be promulgated in both the design and use of those systems. This suggests that system reliability is an important issue to be studied.

The basic model of information relevance in the information retrieval process is reviewed, where the precision is a function, in part, of the recall and fallout rate. Then that relevance model is extended to include reliability due to the intermediate nature of the system and the implications of that reliability are studied. Reliability is found to have a major impact on precision and fallout rates. Thus, decisions on the design of internet information and retrieval systems also are impacted.

1. Introduction

There has been substantial analysis of the quality of information systems, primarily in the information retrieval literature, with an analysis of concepts such as precision, recall and fallout, discussed further below. As a result, the primary focus of information retrieval has been on the relevance of the information retrieved. However, with internet-based information systems there is an additional concern that has received little attention in the information retrieval community. Oftentimes databases on the internet are administered locally, where administrators do not have substantial expertise. Such local control of databases suggests that the information may not be as reliable as if they were expert actors involved throughout.
In addition, some of the databases on the internet are perhaps at best only casually maintained. Some home pages are specifically not truthful in their content, in order to get the user to pursue some activity. As a result, there is a concern for the reliability of the information retrieved on internet-based systems.

On the internet we see a number of similar scenarios. Requests are made of search engines and addresses and descriptors are provided to the users. The responses to the queries are "reports" that may or may not differ from the actual data. Similarly, home pages have links that provide the visitor with "reports" of data that may or may not be what the links indicate that they are.

As an example, using a major search engine yields the following information

"Electronic Commerce Course Cases Page -- http://www.usc.edu/dept/sba/atisp/AI/IJISAFM/call-for.htm (Score 82, Size 2K) International Journal of Intelligent Systems in Accounting, Finance and Management (IJISAFM): Call for Papers . Editor: Daniel E. O'Leary . University of Southern California . 3660 Trousdale (See also Similar Pages)"

In actuality, the real address for the label "Electronic Commerce Course Cases Page" is

http://www.usc.edu/dept/sba/atisp/ec/cases/case2.htm.

In this case, information was supplied by the originator of the particular home page. That data was then placed into the search engine for use on future searches. For some reason, whether it was because of the information submitted or the processing of submitted information the search engine has incorrect descriptor and URL information. In this case the actual information on

Electronic Commerce Course Cases is not the same as the report contained in the search engine.

Another type of lack of reliability is deliberate mislabeling the user of the database. For example, as noted in his speech to the *AAAI Fall Symposium Workshop on Knowledge Navigation,* Tom Gruber's home page has the misleading verbiage

"Nude Photos
In response to user requests.."

(http://www-ksl.stanford.edu/people/gruber/index.html)

As a final example, hyperlinks on a home page can be mislabeled bringing the visitor to locations different than expected. For example, one of the hyperlinks located at http://www.usc.edu/dept/sba/atisp/AI/research/ai-reeng.htm, labeled

"Call for Papers -- International Journal of Intelligent Systems in Accounting, Finance and Management -- Forthcoming"

is wrong, linking to

http://www.usc.edu/dept/sba/atisp/AI/AI-Bus/sigaibus.htm,

rather than,

http://www.usc.edu/dept/sba/atisp/AI/IJISAFM/call-for.htm.

These examples indicate that there is a "reliability" concern with information retrieved on the internet.

In order to address the issue of reliability, internet information retrieval systems are modeled as an intermediate step between the original "idealized" information to be retrieved, the actual information which is available to be retrieved and the user of the information. Ultimately, information retrieved is a computer-based "report" using the available information in its database. That information is a representation of the original material. The retrieved information is only a representation of the original text. That representation may be a URL, a few key words, the system's version of an abstract, the system's version of the entire text or the system may not actually have a version of the text.

In any case, the relationship between the available version and the original text will be used to model the reliability of the system. Reliability is introduced by distinguishing between the report of information available to the system and the actual direct investigation of the original documents or sources of information by the user. In particular, if the system's version is x# and the actual version is x, then Pr(x#|x) will be used to represent the reliability of the representation. That probability is not 1 for a number of reasons, including errors in content, errors in indexing, errors in identification, errors in query formulation, errors in statement of what the material is and other issues.

Accordingly, the purpose of this paper is to introduce reliability of the representation into information in order to determine the impact on traditional information retrieval measures and to help establish alternative optimal strategies that account for reliability.

1.1 Contributions

This paper has a number of contributions. First, the model integrates reliability into the classic information retrieval model of relevance, precision, recall and fallout. Second, it is found that even small changes in reliability can have a material effect on those information retrieval measures. In one example presented later in the paper, decreasing reliability from 1 to .99 leads to a decrease in precision of over 25% and an increase in fallout by 198%. Third, since classic models of information retrieval characterize relevance as a critical point decision, and since reliability influences the value that is compared to the critical point, then by not including reliability into the decision making process, it is found that non-optimal decisions will be made. Monotonicity results are used to study the behavior around the critical point due to reliability. Fourth, accounting for reliability has a relatively small cost: in the simplest case only one additional parameter beyond the classic model needs to be estimated. Fifth, reliability on the internet is characterized so that we can study its impact on internet-based information systems. Finally, the paper finds that reliability is an important component in the design and development of internet-based information systems.

2. Reliability Issues in Internet Information Systems

There are a number of sources of concern for the reliability of an information system, i.e., to believe that Pr(x#|x) is not always equal to one. Those sources derive from the process of establishing content and retrieving information and include, indexing documents, forming queries, and searching the files.

Content of home pages and the corresponding links on those pages may contain errors. Errors can occur in the descriptor or in the address and descriptor match, as noted in the above examples. If content is scanned then there generally is a reliability of about 80-90%, but definitely less than one, indicating reliability issues even when full text is used.

On the internet, generally pages are indexed with keywords, locally for search algorithms. This step consists of assigning content identifiers to homepages or documents either manually by human indexers or automatically by web crawlers. In either case, errors or omissions can be introduced into the system, because of the frailties of humans and software.

In the first case, experimental studies, e.g., Zunde and Dexter, (1969) have found that different indexers frequently will index the same document differently. In addition, the same indexer will index the same document differently at different times. This can lead to different characterization according to different search engines. The information may be keyed into the system incorrectly (e.g., Bourne, 1963), the information may not be verified and errors may be introduced in the maintenance process. Further, since development is local there is often no control over the quality of the relationship between the descriptors and the actual content of a page or document.

In the second case of web crawlers and other robot approaches, machines are good, but they are less than perfect. "Automatic indexing," designed to increase reliability and speed of indexing, is the process of assigning content identifiers with the aid of a computer (e.g., Salton 1986 & Salton and McGill, 1983) Automatic indexing is limited to the ability of machines to process natural language. Jacobs (1991) reports that the best such systems have an 80% accuracy.

In internet and retrieval systems there generally is a query of the system to find the necessary information. This is the process of choosing particular content identifiers and using them to identify text that meets desired specifications. This process typically is done either by the user or a human intermediary. In either case, as noted by Rau (1991), even the formation of a query for a company name can take many different forms. In particular Rau (1991) notes that, although there are a number of heuristics used to formulate queries (e.g., capital letters for names, capturing instances of "Incorporated," etc.), heuristics are by definition approximate solution generation processes. In addition, the use of an intermediary can put another level of uncertainty and error in the process of query formation. The intermediary might misunderstand the user's needs or may not be as expert in the specific domain. Some of the search engines indicate that the primary reason for unsuccessful searches are spelling errors.

The search of the file by the computer, comparing the query formulation with the indexed document representations, is the actual preparation of a "report" for the user by the system. Because these searches are done by the computer, there is little opportunity for error at this stage of the process. However, programming bugs would still facilitate the opportunity for a lack of reliability at this stage.

Modelers of document retrieval processes have not taken reliability into account in the development of analytic models. The classic description of the information retrieval model is concerned with relevance, not reliability. Thus, direct modeling of reliability can be an important extension of the information retrieval model. For example, Kochen (1974, p. 158), discussed a relevance-based information retrieval model that assumes that "A document is retrieved only if it contains each of the words w_1, ..., w_m." This statement assumes that if the original document contains each of the words w_1, ..., w_m, then the computer model of that document also contains those words. In addition, if the plan is to retrieve those words, then the actually implementation includes those words. That is, it implicitly assumes that there is direct access to the actual documents that are then searched, without error, are then found to contain the relevant words w_1, ..., w_m, and are then given directly to the ultimate user. The model does not take into account the secondary or intermediate structure of document retrieval systems. These systems provide only representations of selected documents -- not the actual documents. Thus, Kochen's statement can be paraphrased as "A document is selected to be retrieved only if a representation of that document is in the system and that representation contains each of the words, w_1, ..., w_m." In addition, rather than words w_i, there might be reference to representations of the word w_i, with words $w_{i,1}$, ..., $w_{i,ti}$

3. Performance Measures and a Relevance Model of Retrieval

This paper couches its analysis in a classic information retrieval model (Marion and Kuhns, 1960 & Verhoeff, 1961), that allows extension to the examination of the impact of reliability. That model assumes a retrieved data item is either relevant (R) to the user or it is not relevant (NR). The model also assumes then, if "we have a set of documents" (Verhoeff, 1961, p. 557]) that the criteria for selection indicate either the document data item should be selected (S) or not selected (NS).

There are a number of performance measures used by system designers to measure the effectiveness of information retrieval in the context of this model. These measures include the "precision," "recall," and "fallout." Precision also is referred to as "acceptance rate," recall sometimes is referred to as "hit rate" and the fallout often is called "type II error" or "false drop"(Swets, 1963 & Kochen, 1974).

Let Pr(a) be the probability of a and let Pr(a and b) be represented as Pr(a,b). In the context of the model used in this paper, the precision is equal to Pr (R|S), the recall is equal to Pr (S|R) and type II error is Pr(S|NR). Salton (1986) has noted that in a situation where there is an inverse trade-off between the precision and recall, users tend to favor precision maximizing searches. The rationale behind this choice is that, particularly in very large databases, these types of searches would yield a smaller, yet relevant set of documents. As a result, the primary focus of this paper is on the precision, but, recall and fallout also are examined.

Associated with the design of an information retrieval system are some costs and values to the user. Using the cost notation of Swets 1963, V1 is the value to the user of retrieving a relevant item; V2 is the value of not retrieving an irrelevant item; K1 is the cost of retrieving a nonrelevant item; K2 is the cost of failing to retrieve a relevant item. The user incurs V2 because by not retrieving the item the user does not lose time investigating the item. K1 is a cost because the user will spend time investigating a non relevant item. K2 is an opportunity cost of not examining a relevant item. These costs and values are summarized in table 1.

Costs of Information Retrieval --- Table 1

	Relevant	Not Relevant
Select	K_1	V_1
Not Select	K_2	V_2

Verhoeff et al. (1961) developed a model that indicates that the information system retrieval system is maximized if the probability of relevance (Pr (R)) is greater than a critical probability PCR= (K1 + V2)/(K1 + K2 + V1 + V2). That model is developed as follows. Let p equal the probability that the item is relevant and (1-p) equal the probability that the item is not relevant. The critical probability at which the costs and benefits of retrieving and not retrieving are equal is as follows:

$$p*V1 - (1-p)*K1 = -p*K2 +(1-p)*V2$$

Thus, the above relationship holds. This result is not new, but the critical point nature of the process is important to the results established later in the paper.

However, if only the prior probability (Pr (R)) is used then that ignores the direct search of the database by the user. Bayes' Theorem can be used to relate the posterior probability, the precision and in general Pr(R|x), to the prior probability that the item is relevant, P'=Pr(R) (prior to our observation of x, the direct inspection of the system yielding indications that we should select or not select the item).

$$P = Pr(R|x)$$
$$Pr(R|x) = Pr(R, x)/Pr(x)$$
$$Pr(R|x) = [Pr(x|R)*P']/[Pr(x|R)*P' + Pr(x|NR)*(1-P')].$$
$$Pr(R|x) = P'/[P'+(1-P')*L(x)], \quad (1)$$
$$\text{where, } L(x) = Pr(x|NR)/Pr(x|R). \quad (2)$$

Thus, for x = S, the precision Pr(R|S) is related to L(S)= Pr(S|NR)/Pr(S|R), which is the ratio of the fallout to recall. L(S) is the ratio analyzed in Swets (1963).

The retrieval information changes the prior probability (P'= Pr (R)) that the item is relevant to yield P=Pr(R|x). If Pr (R|x) exceeds the critical value then it is desirable to retrieve the data item. Thus, if there is reason to suppose that L(x) is understated or overstated for any reason, such as reliability, then the cutoff nature of the process can lead to inappropriate decision being made. This is the basis of the discussion in the remainder of the paper.

4. Reliability of Evidence

Unfortunately, in the above model, as noted in Verhoeff et al. (1961), "we assume that the inquirer expects a certain reference list, namely the one he would have procured had he himself probed the documents in the set." However, the information retrieval system reports on information it has retrieved from its database, it does not retrieve perfectly from the entire set of feasible source documents. Thus, the information system reports a value IS (i.e., the information system suggests that the data item be selected) -- it does not provide S. Alternatively, the information system reports a value of INS (not selected) rather than NS. Mathematically, the distinction between the report of the evidence (x#) from the system and the actual occurrence in the original data (x) can be introduced into the probability Pr(R|x#) by introducing it into the only factor that includes

the variable x, L(x). This is done in Lemma 1. Let x'= "not x."

Lemma 1 (Based on Schum and Du Charme, 1971)

L(x#)= [Pr(x#|(NR, x))Pr(x|NR)+Pr(x#|(NR, x'))Pr(x'|NR)]/[Pr(x#|(R, x))Pr(x|R)+Pr(x#|(R, x'))Pr(x'|R)]

Proof

L(x#)= Pr (x#|NR)/Pr (x#|R)= [Pr(x#, NR)/Pr (NR)]/[Pr (x#, R)/Pr (R)]

= [{Pr(x#, x, NR)+Pr(x#, x', NR)}/Pr (NR)]/[{Pr(x#, x, R)+Pr(x#, x', R)}/Pr (R)]

= [{Pr(x#|(NR, x))Pr(x|NR)Pr(NR)+Pr(x#|(NR, x'))Pr(x'|NR)Pr(NR)}/Pr (NR)]/ [{Pr(x#|(R, x))Pr(x|R)Pr(R)+Pr(x#|(R, x'))Pr(x'|R)Pr(R)}/Pr (R)]

= [Pr(x#|(NR, x))Pr(x|NR)+Pr(x#|(NR, x'))Pr(x'|NR)]/ [Pr(x#|(R, x))Pr(x|R)+Pr(x#|(R, x'))Pr(x'|R)]

The factors in L(x#) that relate to Pr(x#|.) reflect what Schum and Du Charme (1971) refer to as the reliability of the reported evidence. If Pr(x#|(NR, x)) = 1 and Pr(x#|(NR, x')) = 0 and if Pr(x#|(R, x)) = 1 and Pr(x#|(R, x')) = 0 then L(x) = L(x#). The report would be 100% reliable. Then the model in Lemma 1 would reduce to model (2). However, if Pr(x#|(NR, x')) and Pr(x#|(R, x')) are not zero and/or Pr(x#|(NR, x)) and Pr(x#|(R, x)) are less than one, then there is nonzero probability that the reported value is dependent on either the relevance (R) of the item, the actual value of the occurrence (x), or both.

4.1 Reliability Assumptions

The model in Lemma 1 has four different reliability parameters. In order to make the discussion more tractable and to focus on the impact of reliability, two special cases of Pr(x#|.) will be analyzed. This is done to reduce the number of parameters in order to facilitate the discussion and so that the impact and importance of reliability as a concept can be assessed. Then a detailed investigation of the implications of the introduction of reliability into the model (1) is made.

The first assumption on reliability is that the probability distribution of the reported version of x, x#, is not dependent on probability distribution of the status of whether or not a document is relevant (NR or R) and that Pr(x#|x) is symmetric, i.e., Pr(x#|x) = Pr(x#'|x')= r. In this case there is a certain amount of confusion as to whether x or x' actually occurs. These probabilities are summarized in table 2. The impact of this assumption is summarized in Theorem 1.

SYMMETRIC RELIABILITY PROBABILITIES*

Reported Value	Actual Value** S	Actual Value** NS
IS	r	1-r
INS	1-r	r

Table 2

* With symmetric reliability the probabilities are independent of whether the system is R (relevant) or NR (not relevant).

** S and NS refer to the states " select the data item" and "do not select the data item," assuming that there is direct access to the original documents or system access to perfect representation of the original text.

*** IS and INS refer to the states "information system indicates that data item should be selected" and "information system suggests that data item not be selected."

Theorem 1

If reliability is symmetric and $1 \geq r \geq 0$, then

L(x#) = L(r,x)= [r*Pr(x|NR) + (1-r)*(Pr(x'|NR)]/[r*Pr(x|R) + (1-r)*(Pr(x'|R)]

Proof By Lemma 1,

L(x#) = [Pr(x#|(NR, x))Pr(x|NR)+Pr(x#|(NR, x'))Pr(x'|NR)]/[Pr(x#|(R, x))Pr(x|R)+Pr(x#|(R, x'))Pr(x'|R)]

If we assume that the reported version of x, x#, is not dependent on whether the data is relevant or not relevant then

L(x#)= [Pr(x#|x)Pr(x|NR)+Pr(x#|x')Pr(x'|NR)]/[Pr(x#|x)Pr(x|R)+Pr(x#|x')Pr(x'|R)]

Since reliability is symmetric r= Pr(x#|x) and 1-r=Pr(x#|x').//

Thus, in the case of symmetric reliability, Pr(R|x#) = P'/[P' + (1-P') * L(x#)]. The precision of the information retrieval system becomes, Pr(R|IS)= P'/[P'+(1-P')* L(IS)]. The recall becomes Pr(IS|R)= ([r*Pr(S|R) + (1-r)*(Pr(NS|R)]. The fallout rate becomes Pr(IS|NR)= [r*Pr(S|R)+(1-r)*(Pr(NS|R)]

The behavior of L(x#) and Pr(R|x#), as a function of the reliability level is illustrated in an example later in the paper. However, it is important to note that L(x#) is highly sensitive to reliability. Small changes in reliability can have a major impact on L(x#) and, thus, Pr(R|x#).

4.2 The Level of Reliability at With Pr(R|x#) Independent of x#

The value of L(x#)=1 is of particular interest, because if L(x#)=1 then Pr(R|x#)=P'. At that value, Pr(R|x#) is independent of x#, the reported recommendation. In this case, whether or not a data item is relevant is independent of any report by the system, because the system provides results that are completely uncertain. Thus, there is concern with investigating when L(x#) = 1.

If the reliability is symmetric then L(x#)=1 at r=.5. At that value of r there is maximum uncertainty about the reliability of the report. Since the report is completely unreliable there is no reason to revise the prior probability. On the other hand, if reliability is asymmetric then L(x#)=1 when $r_1 = r_2 = k$, for $1 \geq k > 0$.

Alternatively, L(x#) will be one if Pr(x|R) = Pr(x|NR). In this case the recall and the fallout rate are the same. That would mean that there is complete uncertainty in the knowledge of whether we would select an item given that it is relevant or not relevant -- an unlikely state of the world.

5. The Impact of Reliability on Recall and Fallout Rate

The reliability model introduced in section 3 has an impact on both the recall and the fallout rate. Only the symmetric probabilities are discussed because the asymmetric case is similar. Let the recall at reliability r be expressed as H(r) and the fallout rate be F(r).

Theorem 3 -- Recall

Let r" and r' be two different reliability levels, r">r'.

(a) If Pr(S|R)>.5 then H(r") > H(r').

(b) If Pr(S|R)<.5 then H(r") < H(r').

Proof

(a) Proof by contradiction. Assume that Pr(S|R)>.5 and H(r")≤H(r'). Thus,

r"*Pr(S|R) + (1-r")*Pr(NS|R)≤r'*Pr(S|R) + (1-r')*Pr(NS|R)

r"*[2Pr(S|R) -1] ≤ r'*[2Pr(S|R) -1]

But, since r" > r' and Pr (S|R)>.5 there is a contradiction.

(b) Similar to part a.

Theorem 4 -- Fallout Rate

Let r" and r' be two different reliability levels, r">r'.

(a) If Pr(S|NR)>.5 then F(r") > F(r').
(b) If Pr(S|NR)<.5 then F(r") < F(r').

Proof --Similar to theorem 3.

These two theorems indicate that by not taking into account the reliability of the information retrieval system, the fallout rate and the recall can be underestimated or overestimated. Assuming that the reliability was not accounted for, i.e., r=1, indicates that for Pr(S|R)<.5 the recall and, thus, the quality of the system on this dimension is understated. Alternatively, assuming that the reliability was not accounted for, indicates that for Pr(S|R)<.5 the fallout rate is understated and the quality of the system on this dimension overstated.

6. Impact of Reliability on Precision--Symmetric Case

Assuming a symmetric reliability model, the relationship between recall and fallout establishes the impact of reliability changes on precision. The results of this section indicate that in some cases, precision (generally, Pr(R|x#)) increases as reliability decreases and in other cases decreases as reliability decreases. In particular, if the probability of the fallout is less than or equal to the recall and more generally, Pr(x|NR) ≤ Pr(x|R) then under symmetric reliability, as r increases that means that there is decreasing reliability on Pr (x'|NR) and Pr (x'|R), based on Theorem 1. Since Pr(x|NR) ≤ Pr(x|R) that means that L(x#) will decrease. Since L(x#) is in the denominator of Pr(R|x#) that means that P(R|x#) will increase. Further,

P(R|x#) as a function of the reliability, is found to be monotonically increasing or decreasing. These results are summarized in Lemma 3 and Theorem 5. Let r" and r' be two different reliability values and define monotonicity as follows.

Lemma 3--Recall Greater Than or Equal to Fallout

If $Pr(x|NR) \leq Pr(x|R)$ and $r" \geq r'$ then $L(r",x) \leq L(r',x)$, that is L(r,x) is monotone decreasing in r.

Proof

Proof by contradiction. Assume that $r" > r'$ and $L(r",x) \geq L(r',x)$. Let $p_1=Pr(x|NR)$ and $p_2=Pr(x|R)$.

$$[r"*p_1 + (1-r")*(1-p_1)]/[r"*p_2 + (1-r")*(1-p_2)] \geq [r'*p_1 + (1-r')*(1-p_1)]/[r'*p_2 + (1-r')*(1-p_2)]$$

$$[r"*p_1 + (1-r")*(1-p_1)]*[r'*p_2 + (1-r')*(1-p_2)] \geq [r'*p_1 + (1-r')*(1-p_1)]*[r"*p_2 + (1-r")*(1-p_2)]$$

$$r'*p_2*(1-p_1)*(1-r") + r"*p_1*(1-p_2)*(1-r') \geq r'*p_1*(1-p_2)*(1-r") + r"*p_2*(1-p_1)*(1-r')$$

$$r'*p_2 + r"*p_1 \geq r'*p_1 + r"*p_2$$

$$r'*(p_2-p_1) \geq r"*(p_2-p_1)$$

But $r" \geq r'$ so there is a contradiction and $L(r",x) \leq L(r',x)$

Theorem 5--Recall Greater Than or Equal to Fallout

If $Pr(x|NR) \leq Pr(x|R)$ and $r" \geq r'$ then $Pr(R|(r",x)) \geq Pr(R|(r',x))$, i.e., Pr(R|(r,x)) is monotonically increasing in r.

Proof

If $Pr(x|NR) \leq Pr(x|R)$ and $r" \geq r'$ then $L(r",x) \leq L(r',x)$. Thus, $Pr(R|(r",x)) \geq Pr(R|(r',x))$.

The results in Theorem 5 indicate that Pr(R|(r,x)) is monotonically increasing in r. If an information retrieval model is used that does not account for reliability, then that assumes r=1. Thus, when r<1, Pr(R|x#) is assumed to be higher than it actually is, i.e., the precision is overestimated if reliability is not accounted for. By assuming that r=1, the user may be retrieving data elements that are not relevant under the model (1). Thus, it is clear that it is important to include reliability in information retrieval models, otherwise more data items may be investigated than is warranted statistically.

Similar results can be developed for the case where the recall is less than or equal to the fallout, and more generally, $Pr(x|NR) \geq Pr(x|R)$. The results are summarized in Lemma 4 and Theorem 7. In contrast to Theorem 6, however, Pr(R|(r,x)) is monotonically decreasing in r.

Lemma 4--Recall Less Than or Equal to Fallout

If $Pr(x|NR) \geq Pr(x|R)$ and $r" \geq r'$ then $L(r",x) \geq L(r',x)$, that is L(r,x) is monotonically increasing in r.

Proof -- Similar to Lemma 3.

Theorem 6--Recall Less Than or Equal to Fallout

If $Pr(x|NR) \geq Pr(x|R)$ and $r" \geq r'$ then $Pr(R|(r",x)) \leq Pr(R|(r',x))$, i.e., Pr(R|(r,x)) is monotonically decreasing in r.

Proof -- Similar to Theorem 5.

The results in Theorem 6 indicate that Pr(R|(r,x)) is monotonically decreasing in r. If the information retrieval model does not take into account reliability, then this assumes r=1. Thus, when r<1, Pr(R|x#) is assumed to be lower than it actually is. By assuming that r=1, the decision maker may not be retrieving data items that are relevant under the model (1).

7. Example

Reliability has a large impact. An example was developed to illustrate the impact of reliability on the precision (Pr(R|IS), the recall (Pr(IS|R)) and the fallout rate (Pr(IS|NR)) and is illustrated in Tables 3. The results indicate that, assuming symmetric reliability, there is a substantial impact on precision due to reliability. In the first case, movement of r from 1.00 to .99, yielded a change of 26.3% in Pr(R|IS), while a drop in r from 1.00 to .90, lead to a change of 73.4%. In the second example, the results were not as strong but still material. A drop in r from 1.00 to .99 caused a change of 7.6% and decreasing r from 1.00 to .90 lead to a change of 39.5%. However, the changes in Pr(IS|R) and Pr(IS|NR) are much less severe. In general, these latter two rates seem to change at about roughly the same rate as the change in reliability.

Example - Symmetric Reliability --- Table3

Pr(S|NR)	Pr(S|R)	Pr(R)	r	Pr(R|IS)	Pr(IS|R)	Pr(IS|NR)
.005	.2	.1	1.00	.816	0.20	0.0050
.005	.2	.1	0.99	.601	0.21	0.0149
.005	.2	.1	0.95	.319	0.23	0.0545
.005	.2	.1	0.90	.217	0.26	0.1040
.005	.2	.1	0.50	.100	0.50	0.5000
.005	.2	.1	0.25	.088	0.65	0.7475
.005	.2	.1	0.00	.082	0.80	0.9950
.02	.2	.3	1.00	.811	0.20	0.0050
.02	.2	.3	0.99	.749	0.21	0.0296
.02	.2	.3	0.95	.592	0.23	0.0680
.02	.2	.3	0.90	.490	0.26	0.1160
.02	.2	.3	0.50	.300	0.50	0.5000
.02	.2	.3	0.25	.273	0.65	0.7400
.02	.2	.3	0.00	.259	0.80	0.9800

8. Some Implementation Considerations

In order to implement the classic model (Marion and Kuhns, 1960 & Verhoeff et al., 1961) in this paper, one needs to estimate both costs and values, and probabilities. The reliability based model requires only one extra parameter if the symmetric model is used and only two extra parameters if the asymmetric model is used.

8.1 Costs and Values

The first concern of the classic and reliability models is the development of the costs and values summarized in table 1. It is these costs that provide the critical ratio, P_{CR}. Only if P_{CR} exceeds that critical value should the selection be made. Since these also are developed for the relevance model, they are not discussed further.

8.2 Probabilities

In this paper, there is concern with estimating four basic probabilities: prior probabilities (P(R)), posterior probabilities and reliabilities. Estimating prior probabilities is required for both the classic and reliability model.

Since only the reliabilities are beyond the classic relevance model (Marion and Kuhns, 1960 and Verhoeff et al. 1961), they will be the focus of the remainder of this discussion. In this paper reliability is represented as a probability relationship Pr(x#|x). Estimates of these probabilities can be developed from empirical data gathered in the process of developing information retrieval systems. Complaints, or the lack of complaints to the database administrator are one likely source of reliability data.

Alternatively, experiments or field studies could be used to establish expected reliability calculations for generic information retrieval systems. For a sample of text retrievals, comparisons between the actual and representation can be made in a specific system. Such sampling would need to consider reliability effects from each of the causes elicited in the introduction and section 2. As a result, reliability statistics can be developed for different error sources.

In any case the development of reliability estimates is an empirical issue. It can be addressed using experiments or actual data from operations.

9. Summary and Conclusions

This paper has argued that it is important to view internet information systems as intermediaries between the actual (or expected) data and the requester of an information retrieval, and that the process should be modeled to reflect the "reporting" nature of information retrieval. Since reports cannot always be perfectly accurate, there is a reliability issue. This paper characterized reliability as the relationship between the actual data items and the report of

the data items by the system. Accordingly, reliability was measured by Pr(x#|x).

Reliability characterizes information search results relationships between "reports" on searches to users and the actual information, or between "reported" hyperlink information and actual hyperlinks established.

Initially, the classic relevance-based model of the information retrieval process was presented. In that model precision is a function of both the recall and fallout. In addition, that discussion characterized precision as a cutoff point decision.

The classic model was then extended to include the reliability of the system. The special case of symmetric reliability was investigated, however, the results could be extended to an asymmetric case also. Implications of the basic model were explored. An example illustrated that the introduction of reliability has a major impact of precision probabilities. The impact of reliability on recall and fallout rate was less severe, but still significant.

The classic model indicates that the choice of precision is a cutoff point decision, and reliability significantly affects precision. As a result, in the development of information retrieval systems, not just relevance needs to be considered, but also reliability. Reliability and relevance are not independent considerations. Monotonicity results were developed that showed that non-optimal decisions can be made around the critical point if we do not account for reliability.

The impact of reliability is substantial. Even small changes in reliability levels can have a large impact on relevance levels.

When compared to the classic model, the only new information required are the reliability parameters. Reliability parameters can be gathered from existing data based on system usage or from experiments designed to solicit that information. As a result, it appears that accounting for reliability is not a costly proposition.

References

Bourne, C., 1963. *Methods of Information Handling*, New York, John Wiley & Sons.

Gaughan, E., 1968. *Introduction to Analysis*, Belmont, California, Brooks/Cole.

Jacobs, P., 1991. "From Parsing to Database Generation: Applying Natural Language Systems," *Proceedings of the Seventh Annual Conference on Artificial Intelligence Applications*, Miami Beach, Fl, February , Washington, IEEE, pp. 18-24.

Kochen, M., 1974. *Principles of Information Retrieval*, Los Angeles, Ca., Melville Publishing Company.

Maron, M., and Kuhns, J., 1960. "On Relevance, Probabilistic Indexing and Information Retrieval," *Journal of the Association of Computing Machinery*, Volume 7, Number 3, pp. 216-244.

Rau, L., 1991. "Extracting Company Names from Text," *Proceedings of the Seventh Annual Conference on Artificial Intelligence Applications*, Miami Beach, Fl, February, Washington, IEEE, pp. 29-32.

Salton, G., , 1986. "Another Look at Automatic Text-Retrieval Systems," *Communications of the ACM*, Volume 29, Number 7, July, pp. 648-656.

Salton, G. and McGill, M.,, 1983. *Introduction to Modern Information Retrieval*, New York, McGraw-Hill.

Schum, D. and Du Charme, W., 1971. "Comments on the Relationship Between the Impact and the Reliability of Evidence," *Organizational Behavior and Human Performance*, Volume 6, pp. 111-131.

Swets, J., 1963. "Information-Retrieval Systems," *Science*, Volume 141, July 19, pp. 245-250.

Verhoeff, J., Goffman, W., and Belzer, J.,1961. "Inefficiency of the Use of Boolean Functions for Information Retrieval," *Communications of the Association of Computing Machinery*, Volume 4, pp. 557-558 and p. 594.

Zunde, P. and Dexter, M.,, 1969. "Indexing Consistency and Quality," *American Documentation*, Volume 20, Number 3, July, pp. 259-264.