ELSEVIER

Decision Support Systems 27 (1999) 319–327

Decision Support Systems

www.elsevier.com/locate/dsw

Internet-based information and retrieval systems

Daniel E. O'Leary [*]

Marshall School of Business, University of Southern California, Los Angeles, CA 90089-1421, USA

Abstract

There is limited reliability of internet-based information systems. For example, Internet search engines provide results that have limited reliability and data available on the Internet is limited in its reliability. As a result, the purpose of this paper is to elicit sources of the lack of reliability, develop a model that can be used to study the impact of reliability and propose some solutions to mitigate reliability issues. The model couches Internet data as an ''intermediary report.'' For example, use of a search engine will generate an intermediary ''report'' providing a list of relevant universal resource locators (URL) and a corresponding brief description that may or may not correctly describe the label being searched. This ''report'' structure is used to model Internet information and retrieval systems as an intermediate step between users of the system and the *original* or expected information. 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. Reliability is found to have an impact on precision and fallout rates. Alternatives are proposed to mitigate the impact of this lack of reliability. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Internet; Information retrieval; Reliability

1. Introduction

The purpose of this paper is to investigate issues of reliability of internet-based information and retrieval systems, including data such as ''home pages''. The paper elicits issues that cause a concern for reliability. Then this paper investigates one model of that reliability to find that in some settings reliability has a major impact on issues such as relevance. A probability model is used to study the impact of unreliable information on traditional information retrieval measures, in order to help establish alternative optimal strategies that account for reliability. In addition, this paper briefly explores some potential approaches to mitigating particular reliability issues.

There has been substantial analysis of the quality of information systems, with an investigation 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. Oftentimes databases on the Internet are administered locally, where administrators have limited resources and do not have accountability for the quality of the information. As a result of these and other concerns, Internet information retrieval systems can result in

[*] Tel.: +1-213-740-4856; e-mail: oleary@rcf.usc.edu

PII: S0167-9236(99)00054-8

information that may not be as *reliable* as data in other settings.

As an example, using LYCOS yielded the following erroneous information

Electronic Commerce Course Cases Page
International Journal of Intelligent Systems in Accounting, Finance and Management (IJISAFM): Call for Papers Editor: Daniel E. O'Leary Univ. http://www.usc.edu/dept/sba/atisp/AI/IJISAFM/call-for.htm

In actuality, the description is one of a call for papers in artificial intelligence and the real hyperlink address for the label ''Electronic Commerce Course Cases Page'' is *http://www.usc.edu/dept/sba/atisp/ec/cases/case2.htm*. For some reason, whether it was because of the information submitted or the processing of submitted information, the search engine has incorrect descriptor and universal resource locator (URL) information. In this case the actual information on ''Electronic Commerce Course Cases Page'' is not the same as the report contained in the search engine.

As a result, in order to address the issue of reliability, internet information retrieval systems are modeled as an intermediate step between the *original* ''correct'' information to be retrieved (e.g., the ''correct'' descriptor and URL for electronic commerce course page), and the actual information which is available to be retrieved and the user of the information (e.g., the wrong information about the call for papers). Ultimately, information retrieved is a computer-based ''report'' using the *available* information in its database. The available 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. 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.

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 correct version 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.

2. Limited reliability of internet information systems

There are a number of reasons why Internet information systems can have limited reliability, i.e., why Pr(x#|x) is not 1, including errors, ontological differences, asymmetric knowledge, developer motives, limited resources and expertise, limited accountability and changes in referenced pages.

2.1. Errors on web pages

Perhaps the most apparent source of reliability problems in Internet information systems is errors. These errors can be generated through typing, mapping the directories incorrectly and other types of errors. Errors can be made in data available on web pages; errors can be made in the links from one page to another. In the following example, hyperlinks on a home page are apparently mislabeled bringing the visitor to a different location than expected. At the web page *http://www.usc.edu/dept/sba/atisp/AI/research/ai-reeng.htm* there is a link to a ''Call for Papers...'', however, that link does not connect to a general call for papers, but instead is connected to the AAAI Special Interest Group on AI in Business, *http://www.usc.edu/dept/sba/atisp/AI/AI-Bus/sigaibus.htm.*

2.2. Errors with search engines

As seen in the ''Electronic Commerce'' example in the introduction, errors can penetrate search engines. Errors can be made in search engine queries; errors can be made in information reported to search engines; and search engines can make errors indexing pages. Internet search engines index pages using a number of approaches, such as keywords provided by developers or by human indexers, or generated by web crawlers. In any case, errors or omissions can be introduced into the system, because of the frailties of humans and software. Such limitations can include erroneous information provided to search engines or problems with processing information submitted to search engines.

2.3. Ontological differences

Internet databases cover a wide range of topics, possibly leading to different definitions of terms by users and developers, ''ontological differences.'' For example, as noted in his speech to the 1995 *AAAI Fall Symposium Workshop on Knowledge Navigation*, Tom Gruber's home page has the ''misleading'' (different ontology?) verbiage ''*Nude Photos In response to user requests . . .* '' (*http://www-ksl.stanford.edu/people/gruber/index.html . . .* as of 4/15/98). The link leads to some pictures of cats (*http://www-ksl.stanford.edu/people/gruber/photos/* and *http://www-ksl.stanford.edu/people/gruber/photos/10-c-portrait.gif*), probably not what most visitors would be expecting.

Users may have an alternative meaning for terms established in the ontology. For example, many people might think that a CPA could refer to a ''certified public accountant,'' while others (according to the ''Style Guide'' of the IEEE Computer Society) would think it refers to ''computer press association.'' Experimental studies, such as Zunde and Dexter [7], have found that different human indexers frequently will index the same document differently. As a result, the same page is likely to be indexed differently by different developers. In addition, the same indexer will index the same document differently at different times. Such differences can be the result of different ontologies and can result in descriptors not meaning what investigators expect.

Further, some search engines employ an ontology to aid the search. As noted by Yahoo! a query can result in multiple different meanings depending on the ontology. For example,

> A Yahoo! Web site match is listed within the category that contains it. This offers you the opportunity to go directly to the site, or to click on the category for a list of related sites. If you're not searching for something specific, we suggest the latter. Again, Yahoo! categories can yield a number of worthwhile options.

Alternative ontologies can result in differences between x and x#.

2.4. Asymmetric knowledge between database user and developer

As noted by Brown et al. [1], there is asymmetric knowledge between suppliers and users of information on the Internet. Asymmetric knowledge allows a developer to deliberately ''mislabel'' homepage links in order to accomplish particular objectives, such as guiding a user to see particular home page material or advertising. Such deliberate mislabeling would result in reliability concerns where x was not the same as x#.

2.5. Developer motives

There are a wide range of motives attributed to the generation of web pages, e.g., in order to provide some set of information or just as fun. Further, as seen by the many visitor counters that typically are on home pages, a number of developers apparently want to attract visitors to their pages. Brown et al. [1] suggests that information can be misrepresented in order to generate visitors to web pages. Such misrepresentation could result in reliability concerns where x was not the same as x#.

2.6. Limited resources and expertise

Many Internet databases are developed and maintained by a single developer or by small companies. Since there may not be supporting organizations, developers can face resource constraints. Further, developers are limited to their own expertise and learning capabilities. These limitations can manifest themselves as limitations to the reliability of the information.

2.7. Limited accountability

Developers of internet-based databases have limited accountability for the databases that they generate. Use of internet-based databases often is a buyer or user beware setting. Information is frequently available at no cost, but there is no guarantee as to the quality or reliability of the information. As a result, even when errors are found, there may be no incentives to fix the errors. Accordingly, this can influence the reliability of the system.

2.8. Changes in referenced pages

Pages need to be updated over time. As pages are updated, their relationship to other pages can change and the pages that they reference may change. Since there is not likely to be a reciprocal arrangement between page developers to send information about page changes to developers of pages who reference their page, addresses and content can change without the referencing page developer knowing. As a result, in this setting, reliability can be impacted because of changes outside the control of the referencing page developer.

3. Performance measures and a relevance model of information retrieval

This paper couches its analysis in a classic information retrieval model [6], 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 that, if ''we have a set of documents ([6], 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'' [2,5].

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 [3] 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. The primary focus here is on the precision, but, recall and fallout also are examined.

Table 1
Costs of information retrieval

	Relevant	Not relevant
Select	V1	K1
Not select	K2	V2

Associated with the design of an information retrieval system are some costs and values to the user. Using the cost notation of Swets [5], 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 non-relevant 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.

Verhoeff et al. [6] 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 $\mathrm{PCR} = (\mathrm{K1} + \mathrm{V2})/(\mathrm{K1} + \mathrm{K2} + \mathrm{V1} + \mathrm{V2})$. 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 $p\mathrm{V1} - (1-p)\mathrm{K1} = -p\mathrm{K2} + (1-p)\mathrm{V2}$. 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, $P' = \Pr(\mathrm{R})$ (prior to our observation of x, the direct inspection of the system yielding indications that we should select or not select the item).

$$P = \Pr(\mathrm{R}|\mathrm{x})$$

$$\Pr(\mathrm{R}|\mathrm{x}) = \Pr(\mathrm{R},\mathrm{x})/\Pr(\mathrm{x})$$

$$\Pr(\mathrm{R}|\mathrm{x}) = [\Pr(\mathrm{x}|\mathrm{R})P']/[\Pr(\mathrm{x}|\mathrm{R})P' + \Pr(\mathrm{x}|\mathrm{NR})(1-P')].$$

$$\Pr(\mathrm{R}|\mathrm{x}) = P'/[P' + (1-P')L(\mathrm{x})], \qquad (1)$$

$$\text{where, } L(x) = \Pr(x|NR)/\Pr(x|R). \tag{2}$$

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 by Swets [5].

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. Extension of this model provides the basis of the discussion in the next section.

4. Modeling reliability

Unfortunately, in the above model, as noted in Verhoeff et al. [6], ''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, as noted in the ''Electronic Commerce'' example discussed above, 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 *s*elected) — it does not provide S. Alternatively, the information system reports a value of INS (*n*ot *s*elected) 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, $L(x)$. Let $x' =$ ''not x.''

Lemma 1 (Based on Schum and Du Charme [4]) $(x\#) = [\Pr(x\#|(NR,x))\Pr(x|NR) + \Pr(x\#|(NR,x'))\Pr(x'|NR)]/[\Pr(x\#|(R,x))\Pr(x|R) + \Pr(x\#|(R,x'))\Pr\ x'|R)]$.

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 [4] 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 non-zero 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, one special case of $\Pr(x\#|.)$ will be analyzed. The 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

Table 2
Symmetric reliability probabilities[a]

Reported value[c]	Actual value[b]	
	S	NS
IS	r	$1-r$
INS	$1-r$	r

[a] With symmetric reliability the probabilities are independent of whether the system is R (relevant) or NR (not relevant).
[b] S and NS refer to the states ''*s*elect the data item'' and ''do *n*ot *s*elect the data item,'' assuming there is direct access to original documents or system access to perfect representation of original.
[c] IS and INS refer to the states ''*i*nformation system indicates that data item should be *s*elected'' and ''*i*nformation system suggests that data item *n*ot be *s*elected.'').

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. Although the remainder of this paper is concerned with symmetric probabilities, the results can be extended to asymmetric and other types of probabilities.

Theorem 1 *If reliability is symmetric and* $1 \geq r \geq 0$, *then,* $L(x\#) = L(r,x) = [rPr(x|NR) + (1-r)(Pr(x'|NR)] / [rPr(x|R) + (1-r)(Pr(x'|R)]$

Proof by Lemma 1, $L(\mathrm{x}\#) = [\Pr(\mathrm{x}\#|(\mathrm{NR,x}))\Pr(\mathrm{x}|\text{-}\mathrm{NR}) + \Pr(\mathrm{x}\#|(\mathrm{NR,x'}))\Pr(\mathrm{x'}|\mathrm{NR})]/[\Pr(\mathrm{x}\#|(\mathrm{R,x}))\Pr(\mathrm{x}|\mathrm{R}) + \Pr(\mathrm{x}\#|(\mathrm{R,x'}))\Pr(\mathrm{x'}|\mathrm{R})]$If we assume that the reported version of x, x#, is not dependent on whether the data is relevant or not relevant then $L(\mathrm{x}\#) = [\Pr(\mathrm{x}\#|\mathrm{x})\Pr(\mathrm{x}|\mathrm{NR}) + \Pr(\mathrm{x}\#|\mathrm{x'})\Pr(\mathrm{x'}|\mathrm{NR})]/[\Pr(\mathrm{x}\#|\mathrm{x})\Pr(\mathrm{x}|\mathrm{R}) + \Pr(\mathrm{x}\#|\mathrm{x'})\Pr(\mathrm{x'}|\mathrm{R})]$Since reliability is symmetric $r = \Pr(\mathrm{x}\#|\mathrm{x})$ and $1 - r = \Pr(\mathrm{x}\#|\mathrm{x'})$.//

Thus, in the case of symmetric reliability, $\Pr(\mathrm{R}|\mathrm{x}\#) = P'/[P' + (1-P')L(\mathrm{x}\#)]$. The precision of the information retrieval system becomes, $\Pr(\mathrm{R}|\mathrm{IS}) = P'/[P' + (1-P')L(\mathrm{IS})]$. The recall becomes $\Pr(\mathrm{IS}|\mathrm{R}) = ([r\Pr(\mathrm{S}|\mathrm{R}) + (1-r)(\Pr(\mathrm{NS}|\mathrm{R})]$. The fallout rate becomes $\Pr(\mathrm{IS}|\mathrm{NR}) = [r\Pr(\mathrm{S}|\mathrm{R}) + (1-r)(\Pr(\mathrm{NS}|\mathrm{R})]$. The behavior of $L(\mathrm{x}\#)$ and $\Pr(\mathrm{R}|\mathrm{x}\#)$, as a function of the reliability level is illustrated in an example later in the paper. However, it is important to note $L(\mathrm{x}\#)$ and, thus, $\Pr(\mathrm{R}|\mathrm{x}\#)$ are highly sensitive to reliability.

4.2. 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. Let the recall at reliability r be expressed as $H(r)$ and the fallout rate be $F(r)$.

Theorem 2 — recall *Let* r'' *and* r' *be two different reliability levels,* $r'' > r'$.*(a) If* $Pr(S|R) > 0.5$ *then* $H(r'') > H(r')$, *(b) If* $Pr(S|R) < 0.5$ *then* $H(r'') < H(r')$.

Proof(a) Proof by contradiction. Assume that $\Pr(\mathrm{S}|\mathrm{R}) > 0.5$ and $H(r'') < H(r')$. Thus,$r''\Pr(\mathrm{S}|\mathrm{R}) + (1 - \mathrm{r}'')\Pr(\mathrm{NS}|\mathrm{R}) \leq r'\Pr(\mathrm{S}|\mathrm{R}) + (1 - r')\Pr(\mathrm{NS}|\mathrm{R})r''[2\Pr(\mathrm{S}|\mathrm{R}) - 1] \leq r'[2\Pr(\mathrm{S}|\mathrm{R}) - 1]$But, since $r'' > r'$ and $\Pr(\mathrm{S}|\mathrm{R}) > 0.5$ there is a contradiction.(b) Similar to part a.//

Theorem 3 — fallout rate *Let* r'' *and* r' *be two different reliability levels,* $r'' > r'$.*(a) If* $Pr(S|NR) > 0.5$ *then* $F(r'') > F(r')$. *(b) If* $Pr(S|NR) < 0.5$ *then* $F(r'') < F(r')$.

Proof — Similar to Theorem 2.//

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 $\Pr(\mathrm{S}|\mathrm{R}) < 0.5$ 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 $\Pr(\mathrm{S}|\mathrm{R}) < 0.5$ the fallout rate is understated.

4.3. Impact of reliability on precision

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, $\Pr(\mathrm{R}|\mathrm{x}\#)$) 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, $\Pr(\mathrm{x}|\mathrm{NR}) \leq \Pr(\mathrm{x}|\mathrm{R})$ then under symmetric reliability, as r increases that means that there is decreasing reliability on $\Pr(\mathrm{x'}|\mathrm{NR})$ and $\Pr(\mathrm{x'}|\mathrm{R})$, based on Theorem 1. Since $\Pr(\mathrm{x}|\mathrm{NR}) < \Pr(\mathrm{x}|\mathrm{R})$ that means that $L(\mathrm{x}\#)$ will decrease. Since $L(\mathrm{x}\#)$ is in the denominator of $\Pr(\mathrm{R}|\mathrm{x}\#)$ that means that $\mathrm{P}(\mathrm{R}|\mathrm{x}\#)$ will increase. Further, $\mathrm{P}(\mathrm{R}|\mathrm{x}\#)$ as a function of the reliability, is found to be monotonically increasing or decreasing. These results are summarized in Lemma 2 and Theorem 4.

Lemma 2 — recall greater than or equal to fallout*Let r'' and r' be two different reliability. If Pr(x|NR) ≤ Pr(x|R) and $r'' \geq r'$ then L(r'',x) ≤ L(r',x), that is L(r,x) is monotone decreasing in r.*

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

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

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

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

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

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

But $r'' \geq r'$ so there is a contradiction and $L(r'',\mathrm{x}) \leq L(r',\mathrm{x})$//

Theorem 4 — recall greater than or equal to fallout*If Pr(x|NR) ≤ Pr(x|R) and $r'' \geq r'$ then Pr(R|(r'',x)) ≥ Pr(R|(r',x)), Pr(R|(r,x)) is monotone increasing in r.*

In Theorem 4, 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 when the recall is less than or equal to the fallout, and more generally, Pr(x|NR) ≥ Pr(x|R). The results are summarized in Lemma 3 and Theorem 5. In contrast to Theorem 4, however, Pr(R|(r,x)) is monotonically decreasing in r.

Lemma 3 — recall less than or equal to fallout*If Pr(x|NR) ≥ Pr(x|R) and $r'' \geq r'$ then L(r'',x) ≥ L(r',x), that is L(r,x) is monotonically increasing in r.*

Proof — Similar to Lemma 2.//

Theorem 5 — recall less than or equal to fallout*If Pr(x|NR) ≥ Pr(x|R) and $r'' \geq r'$ then Pr(R|(r'',x)) ≤ Pr(R|(r',x)), i.e., Pr(R|(r,x)) is monotonically decreasing in r.*

Proof — similar to Theorem 4.//

The results in Theorem 5 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).

4.4. Example

An example was developed to illustrate the impact of reliability on precision (Pr(R|IS), recall (Pr(IS|R)) and fallout rate (Pr(IS|NR)), as illustrated in Table 3. The results indicate that, assuming symmetric reliability, there is a substantial impact on precision due to reliability. Movement of r from 1.00 to 0.99, 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%.

Table 3
Example — symmetric reliability

Pr(S|NR)	Pr(S|R)	Pr(R)	r	Pr(R|IS)	Pr(IS|R)	Pr(IS|NR)
0.005	0.2	0.1	1.00	0.816	0.20	0.0050
0.005	0.2	0.1	0.99	0.601	0.21	0.0149
0.005	0.2	0.1	0.95	0.319	0.23	0.0545
0.005	0.2	0.1	0.90	0.217	0.26	0.1040
0.005	0.2	0.1	0.50	0.100	0.50	0.5000
0.005	0.2	0.1	0.25	0.088	0.65	0.7475
0.005	0.2	0.1	0.00	0.082	0.80	0.9950
0.02	0.2	0.3	1.00	0.811	0.20	0.0050
0.02	0.2	0.3	0.99	0.749	0.21	0.0296
0.02	0.2	0.3	0.95	0.592	0.23	0.0680
0.02	0.2	0.3	0.90	0.490	0.26	0.1160
0.02	0.2	0.3	0.50	0.300	0.50	0.5000
0.02	0.2	0.3	0.25	0.273	0.65	0.7400
0.02	0.2	0.3	0.00	0.259	0.80	0.9800

5. Approaches to mitigating reliability issues

Section 4 illustrates the impact of reliability. This section provides an example that illustrates development of probabilities for reliability models. In addition, the development of these probability estimates can give insight into some potential technology-based approaches that can provide a basis for improving that reliability.

There is substantial information available that can facilitate an empirical development of the probabilities for the models developed in this paper. For example, consider the case discussed earlier with the link from ''Electronic Commerce Course Cases Page'' to *http: / / www.usc.edu / dept / sba / atisp / AI / IJISAFM / call-for.htm.* An empirical analysis of an ''Excite'' search found that of the first 50 entries for the request ''electronic commerce'' 19 out of 50 had ''e'' and ''c'' as adjacent letters in the URLs and 40 out of 50 had either the adjacent letters ''e'' and ''c'' or ''electronic commerce'' in the short description. Similarly, an ''Excite'' search on ''artificial intelligence'' found the adjacent letters ''a'' and ''i'' or ''artificial intelligence'' in 35 out of 50 directories and descriptors. Using information of this type we can begin to generate estimates of some of the probabilities in the above model that the above URL is a pointer to an ''Electronic Commerce Course Cases Page.''

However, this information is not just useful to generate probabilities. In addition, this same information can be used to generate tools to further mitigate reliability problems, e.g., such as provide the necessary domain intelligence for an intelligent agent who could review URLs to hypothesize whether or not they met retrieval criteria. Those same descriptive terms provide insight into whether or not the label ''electronic commerce'' corresponds to the description and URL. In particular, a knowledgeable assistant or intelligent agent would look at the directory and the URL address and draw at least two conclusions. First, the address probably relates to ''AI,'' an abbreviation for artificial intelligence. In any case, it does not seem to relate to electronic commerce. Second, the file name is an apparent abbreviation of a call for papers. As a result, rather than ''cases'' it seems to be a ''request for papers''. Further, an agent could examine the description to determine that it apparently does not relate to electronic commerce, but instead does relate to artificial intelligence, substantiating the hypothesis created by the term ''AI'' in the directory structure. Technology has rapidly progressed so that now there are many intelligent agent shells available that could be used to exploit and narrow the search through this kind of information (*http: / / www.primenet.com / terry / New_Home_Page / ai_info / intelligent_agents.html* and *http: / / www.botspot.com / site_map /*).

Finally, as noted earlier in the paper, some of the reliability problems are a function of the lack of existence of an ontology that clearly defines its terms or interlinking ontologies that accommodate multiple ontologies. As a result, additional research in ontologies can provide an important capability to increase $\Pr(x\#|x)$ by mitigating ontology ambiguity. As with the use of intelligent agents, the topics being searched and the directory structures of the references being given, can guide the choice of the ontologies. In the ''electronic commerce'' example, ontologies for ''electronic commerce,'' ''AI'' and ''artificial intelligence'' provide a starting point of analysis.

6. Summary, contributions and extensions

This paper has argued that reliability is a critical aspect of Internet-based information systems. A number of sources of problems with reliability were elicited and reviewed. Internet information systems were modeled as intermediary report between the actual data and the system's representation of that data, in order to capture the ''reporting'' nature of information retrieval. Since reports cannot always be perfectly accurate, there is a reliability issue. Reliability was characterized as a probability relationship between the actual data items and the report of the data items by the system, $\Pr(x\#|x)$. Reliability characterizes information search relationships between ''reports'' (e.g., searches) and the actual information, or between ''reported'' hyperlink information and actual hyperlinks.

This paper has a number of contributions. First, the paper elicits reliability issues with Internet information systems. Second, the model integrates reliability into the classic information retrieval model of relevance, precision, recall and fallout. Third, it is found that even small changes in reliability can have a material effect on those information retrieval measures. Fourth, 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. Fifth, accounting for reliability has a relatively small cost: in the simplest case only one additional parameter beyond the classic model needs to be estimated. Sixth, 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 and provides some other approaches that can mitigate reliability problems.

This paper can be extended in a number of directions. First, the set of categories that are at the root of reliability issues in Internet systems could be further analyzed. For example, a sample of pages could be assessed for reliability problems and the relative frequency of problems from different categories, e.g., ''Errors on Web Pages.'' Such a study would give insight into the source of reliability problems in Internet information systems. Second, some of the sources of errors, e.g., ''changes in referenced pages'', suggest organizational relationships that could improve reliability. For example, since changes that others make can influence the reliability, that suggests that reliability can be improved by facilitating communication between those that reference each others pages. Given a list of developers that have referenced a page, when that page was changed its developers could contact those that reference it indicating the nature of the change. In so doing, ''reliability relationships'' could be established. Third, the mathematical model could be extended based on alternative assumptions, e.g., asymmetric probabilities. This would lead to additional results relating reliability and traditional information retrieval measures. Fourth, approaches to mitigating reliability could be extended and implemented. For example, an intelligent agent could be implemented using some of the specific knowledge discussed here.

Acknowledgements

An earlier version of this paper was presented at WITS 97. The author wishes to acknowledge the comments of the anonymous referees and the participants at the presentation. In addition, the author would like to thank the anonymous referees for their comments on a revised version of this paper. The author would like to thank an anonymous WITS referee for suggesting a section on how some of the problems discussed in Section 2 could be addressed. The author would like to thank an anonymous referee for the suggestion of discussing the extensions in greater detail.

References

[1] C. Brown, L. Gasser, D. O'Leary, A. Sangster, AI on WWW: supply and demand agents, IEEE Expert 10 (4) (1995) 50–55.

[2] M. Kochen, Principles of Information Retrieval, Melville Publishing, Los Angeles, CA, 1974.

[3] G. Salton, Another look at automatic text-retrieval systems, Communications of the ACM 29 (7) (1986) 648–656.

[4] D. Schum, W. Du Charme, Comments on the relationship between the impact and the reliability of evidence, Organizational Behavior and Human Performance 6 (1971) 111–131.

[5] J. Swets, Information-retrieval systems, Science 141 (1963) 245–250.

[6] J. Verhoeff, W. Goffman, J. Belzer, Inefficiency of the use of Boolean functions for information retrieval, Communications of the ACM, Vol. 4, 1961, pp. 557–558 and p. 594.

[7] P. Zunde, M. Dexter, Indexing consistency and quality, American Documentation 20 (3) (1969) 259–264.

Daniel E. O'Leary is a Professor in the Marshall School of Business at the University of Southern California. Dan received his PhD from Case Western Reserve University, his MBA from the University of Michigan and his BS from Bowling Green State University. Professor O'Leary has published papers in a wide range of journals including Communications of the ACM, IEEE Intelligent Systems, IEEE Computer and Management Science.