H.4 Evaluation of Retrieval Systems

How can the quality of a retrieval system be described quantitatively? Is search engine X better than search engine Y? What aspects determine the success of a retrieval system? Answering these questions is the task of evaluation research. Croft, Metzler and Strohman (2010, 297) define "evaluation" as follows:

> Evaluation is the key to making progress in building better search engines. It is also essential to understanding whether a search engine is being used effectively in a specific application.

Evaluation regards both system criteria and user assessments. Tague-Sutcliffe (1996, 1) places the user at the center of observation:

> Evaluation of retrieval systems is concerned with how well the system is satisfying users not just in individual cases, but collectively, for all actual and potential users in the community. The purpose of evaluation is to lead to improvements in the information retrieval process, both at a particular installation and more generally.

Retrieval systems are a particular kind of information system. Hence, we must take into account evaluation criteria for information systems in general as well as for retrieval systems in particular.

A Comprehensive Evaluation Model

We introduce a comprehensive model that allows us to span a theoretical framework for all aspects of the evaluation of retrieval systems (similarly: Lewandowski & Höchstötter, 2008, 318; Saracevic, 1995). The model takes into account different dimensions of evaluation. The methods of evaluation derive from various scientific disciplines, including information systems research, marketing, software engineering, and—of course—information science.

A historical point of origin for the evaluation of information systems in general is the registration of technology acceptance (Davis, 1989), which uses subdimensions (initially: "ease of use" and "usefulness", later supplemented by "trust" and "fun") in order to measure the quality of the information system's technical make-up (dimension: IT system quality). In the model proposed by DeLone and McLean (1992), the technical dimension is joined by that of information quality. Information quality concentrates on the knowledge that is stored in the system. The dimension of knowledge quality consists of the two subdimensions of document quality (more precisely: documentary reference unit quality) and the surrogates (i.e. the documentary units) derived from these in the information system. DeLone and McLean (2003) as well as Jennex and Olfman (2006) expand the model via the dimension of service quality. When analyzing IT service quality, the objective is to inspect the services offered by

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

the information system and the way they are perceived by the users. The quality of a retrieval system thus depends upon the range of functions it offers, on their usability (Nielsen, 2003), and on the system's effectiveness (measured via the traditional values of Recall and Precision). In an overall view, we arrive at the following four dimensions of the evaluation of retrieval systems (Figure H.4.1):

- IT service quality,
- Knowledge quality,
- IT system quality,
- Retrieval system quality.

All dimensions and subdimensions make a contribution to the usage or non-usage—i.e., the success or failure—of retrieval systems.

Evaluation of IT Service Quality

The service quality of a retrieval system is described, on the one hand, via the processes that the user must implement in order to achieve results. On the other hand, it involves the attributes of the services offered and the way these are perceived by the user.

Suitable methods for registering the process component of an IT service include the sequential incident technique and the critical incident technique. In the sequential incident technique (Stauss & Weinlich, 1997), users are observed while working through the service in question. Every step of the process is documented, which produces a "line of visibility" of all service processes—i.e., displaying the service-creating steps that are visible to the user. If the visible process steps are known, users can be asked to describe them individually. This is the critical incident technique (Flanagan, 1954). Typical questions posed to users are "What would you say is the primary purpose of X?" and "In a few words, how would you summarize the general aim of X?"

A well-known method for evaluating the attributes of services is SERVQUAL (Parasuraman, Zeithaml, & Berry, 1988). This method works with two sets of statements: those that are used to measure expectations about a service category in general (EX) and those that measure perceptions (PE) about the category of a particular service. Each statement is accompanied by a seven-point scale ranging from "strongly disagree" (1) to "strongly agree" (7). For the expectation value, one might note that "In retrieval systems it is useful to use parentheses when formulating queries" and ask the test subject to express this numerically on the given scale. The corresponding statement when registering the perception value would then be "In the retrieval system X, the use of parentheses is useful when formulating queries." Here, too, the subject specifies a numerical value. For each item, a difference score Q = PE – EX is defined. If, for instance, a test subject specifies a value of 1 for perception after having noted a 4 for expectation, the Q value for system X with regard to the attribute in question will be 1 – 4 = -3.

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

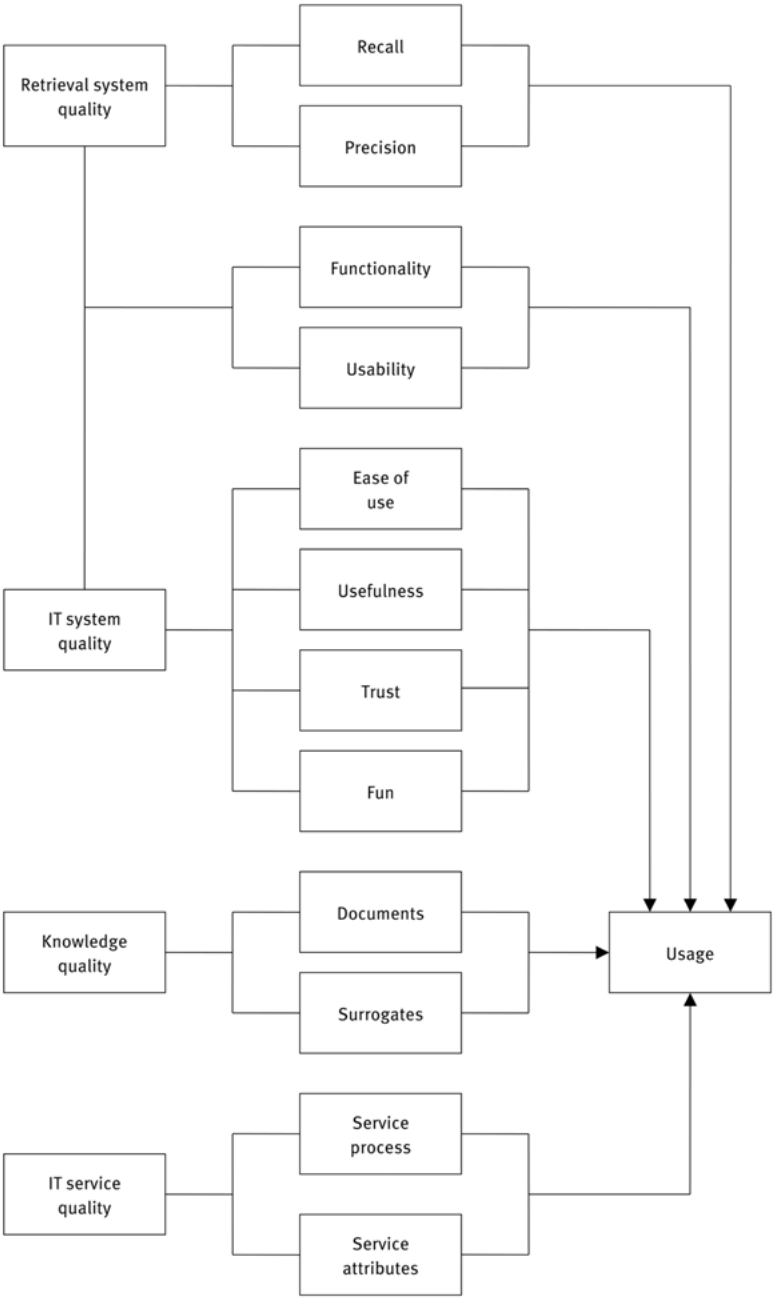

Figure H.4.1: A Comprehensive Evaluation Model for Retrieval Systems. Source: Modified Following Knautz, Soubusta, & Stock, 2010, 6.

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

Parasuraman, Zeithaml and Berry (1988) define five service quality dimensions (tangibles, reliability, responsiveness, assurance, and empathy). This assessment is conceptualized as a gap between expectation and perception. It is possible to adopt SERVQUAL for measuring the effectiveness of information systems (Pitt, Watson, & Kavan, 1995). In IT SERVQUAL, there are problems concerning the exclusive use of the difference score and the pre-defined five quality dimensions. It is thus possible to define separate quality dimensions that are more accurate in answering specific research questions than the pre-defined dimensions. The separate dimensions can be derived on the basis of the critical processes that were recognized via sequential and critical incident techniques. It was suggested to not only apply the difference score, but to add the score for perceived quality, called SERVPERF (Kettinger & Lee, 1997), or to work exclusively with the perceived performance scoring method. If a sufficient amount of users were used as test subjects, and if their votes were, on average, close to uniform, SERVQUAL would seem to be a valuable tool for measuring the quality of IT systems' attributes.

In SERVQUAL, both the expectations and the perceptions of users are measured. Customers value research (McKnight, 2006) modifies this approach. Here, too, the perception values are derived from the users' perspective, but the expectation values are estimates, on the part of the developers or operators of the systems, as to how their users will perceive the attribute in question. The difference between both values expresses "irritation", i.e. the misunderstandings between the developers of an IT service and their customers.

Evaluation of Knowledge Quality

Knowledge quality involves the evaluation of documentary reference units and of documentary units. The quality of the documents that are depicted in an information service can vary significantly, depending on the information service analyzed. Databases on scientific-technological literature (such as the ACM Digital Library) contain scientific articles whose quality has generally already been checked during the publication process. This is not the case in Web search engines. Web pages or documents in sharing services (e.g. videos on YouTube) are not subject to any process of evaluation. The information quality of such documents is extremely hard to quantify. Here we might consider subdimensions such as believability, objectivity, readability or understandability (Parker, Moleshe, De la Harpe, & Wills, 2006).

The evaluation of documentary units, i.e. of surrogates, proceeds on three subdimensions. If KOSs (nomenclatures, classification systems or thesauri) are used in the information service, these must be evaluated first (Ch. P.1). Secondly, the quality of indexing and of summarization is evaluated via parameters such as the indexing depth of a surrogate, the indexing effectiveness of a concept, the indexing consistency of surrogates, or the informativeness of summaries (Ch. P.2). Thirdly, the update

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

speed is evaluated, both in terms of the first retrieval of new documents and in terms of updating altered documents. Both aspects of speed express the database's freshness.

Evaluation of IT System Quality

The prevalent research question of IT system quality evaluation is "What causes people to accept or reject information technology?" (Davis, 1989, 320). Under what conditions is IT accepted and used (Dillon & Morris, 1996)? Davis' empirical surveys lead to two subdimensions, perceived usefulness and perceived ease of use (Davis, 1989, 320):

> *Perceived usefulness* is defined ... as "the degree to which a person believes that using a particular system would enhance his or her job performance." ... *Perceived ease of use*, in contrast, refers to "the degree to which a person believes that using a particular system would be free of effort."

Following the theory of reasoned action, Davis, Bagozzi and Warshaw (1989, 997) are able to demonstrate that "people's computer use can be predicted reasonably well from their intentions" and that these intentions are fundamentally influenced by perceived usefulness and perceived ease of use. The significance of the two subdimensions is seen to be confirmed in other studies (Adams, Nelson, & Todd, 1992) that draw on their correlation with the respective information system's factual usage. In the further development of technology acceptance models, it is shown that additional subdimensions join in determining the usage of information systems. On the one hand, there is the "trust" that users have in a system (Gefen, Karahanna, & Straub, 2003), and on the other hand, the "fun" that users experience when using a system (Knautz, Soubusta, & Stock, 2010). The trust dimension is particularly significant in e-commerce systems, while the fun dimension is most important in Web 2.0 environments. Particularly critical aspects of retrieval systems' usefulness are their ability to satisfy information needs and the speed with which they process queries. Croft, Metzler and Strohman (2010, 297) describe this as the effectiveness and efficiency of retrieval systems:

> Effectiveness ... measures the ability of the search engine to find the right information, and efficiency measures how quickly this is done.

We will return to the question of measuring effectiveness when discussing Recall and Precision in the below.

When evaluating IT system quality, questionnaires are used. The test subjects must be familiar with the system in order to make correct assessments. For each subdimension, a set of statements is formulated that the user must estimate on a 7-point

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

scale (from “extremely likely” to “extremely unlikely”). Davis (1989, 340), for instance, posited “using system X in my job would enable me to accomplish tasks more quickly” to measure perceived usefulness, or “my interaction with system X would be clear and understandable” for the aspect of perceived ease of use. In addition to the four subdimensions (usefulness, ease of use, trust, and fun), it must be asked if and how the test subjects make use of the information system. If one asks factual users (e.g. company employees on the subject of their intranet usage), estimates will be fairly realistic. A typical statement with regard to registering usage is “I generally use the system when the task requires it.” When test subjects are confronted with a new system, estimates are hypothetical. It is useful to calculate how the usage values correlate with the values of the subdimensions (and how the latter correlate with one another). A subdimension’s importance rises in proportion to its correlation with usage.

Evaluation of Retrieval System Quality I: Functionality and Usability

Retrieval systems provide functions for searching and retrieving information. Depending on the retrieval system’s purpose (e.g. a general search engine like Google vs. a specialized information service like STN International), the extent of the functions offered can vary significantly. When evaluating functionality, the object is the “quality of search features” (Lewandowski & Höchstötter, 2008, 320). We differentiate between the respective ranges of commands for search, for push services, and for informetric analyses (Stock, 2000). Table H.4.1 shows a list of functionalities typically to be expected in a professional information service.

Table H.4.1: Functionality of a Professional Information Service. Source: Modified Following Stock, 2000, 26.

Steps	Functionality
Selection of Databases	Database Index Database Selection – precisely one database – selecting database segments – selection across databases
Looking for Search Arguments	Browsing a Dictionary Presentation of KOSs – verbal – graphic – display of paradigmatic relations for a concept – display of syntagmatic relations for a concept Statistical Thesaurus Dictionary of Synonyms (for full-text searches)

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

	– thesaurus (in the linguistic sense)
	Dictionary of Homonyms
	– dialog for clarifying homonymous designations
Search Options	Field-Specific Search
	– search within fields
	– cross-field search in the basic index
	Citation Search
	– search for references (“backward”)
	– search for citations (“forward”)
	– search for co-citations
	– search for bibliographic coupling
	Grammatical Variants
	– upper/lower case
	– singular/plural
	– word stem
	Fragmentation
	– to the left, to the right, in the center
	– number of digits to be replaced (precisely n digits; any number of digits)
	Set-Theoretical Operators
	– Boolean operators
	– parentheses
	Proximity Operators
	– directly neighboring (with and without a regard for term order)
	– adjacency operator
	– grammatical operators
	Frequency Operator
	Hierarchical Search
	– search for a descriptor including its hyponyms on the next n levels
	– search for a descriptor including its hyperonyms on the next n levels
	– search for a descriptor including all related terms
	Weighted Retrieval
	Cross-Database Search
	– duplicate detection
	– duplicate elimination
	Reformulation of Search Results into Search Arguments
	– mapping in the same field
	– mapping into another field
	– mapping with change of database
Display and Output	– hit list
	– sorting of search results
	– marking of surrogates for sorting and output
	– output of reports in tabular form
	– output of the surrogates in freely selectable format
	– particular output formats (e.g. CSV or XML)
	Ordering of Full Texts
	– provision in the original format

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

	– link to the digital version
	– link to document delivery services
Push Services	– creation of search profiles
	– management of search profiles
	– delivering of search results
Informetric Analysis	– rankings
	– time series
	– semantic networks
	– information flow graphs

How does a retrieval system present itself to its users? Is it intuitively easy to use? Such questions are addressed by usability research (Nielsen, 2003). “Usable” retrieval systems are those that do not frustrate the user. This view is shared by Rubin and Chisnell (2008, 4):

> (W)hen a product or service is truely usable, *the user can do what he or she wants to do the way he or she expects to be able to do it, without hindrance, hesitation, or questions*.

A common procedure in usability tests is task-based testing (Rubin & Chisnell, 2008, 31). Here an examiner defines representative tasks that can be performed using the system and which are typical for such systems. Such a task for evaluating the usability of a search engine might be “Look for documents that contain your search arguments verbatim!” Test subjects should be “a representative sample of end users” (Rubin & Chisnell, 2008, 25). The test subjects are presented with the tasks and are observed by the examiner while they perform them. For instance, one can count the links that a user needs in order to fulfill a task (in the example: the number of links between the search engine’s homepage to the verbatim setting). An important aspect is the difference between the shortest possible path to the goal and the actual number of clicks needed to get there. The greater this difference is, the less usable the corresponding system function will be. An important role is played by the test users’ abandonment of search tasks (“can’t find it”) and by their exceeding the time limit. Click data and abandonment frequencies are indicators for the quality of the navigation system (Röttger & Stock, 2003).

It is useful to have test subjects speak their thoughts when performing the tasks (“thinking aloud”). The tests are documented via videotaping. Use of eye-tracking methods provides information on which areas of the screen the user concentrated on (thus possibly overlooking a link). In addition to the task-based tests, it is useful for the examiner to interview the subjects on the system (e.g. on their overall impression of the system, on screen design, navigation, or performance).

Benchmarks for usability tests are generally set at a minimum of ten test subjects and a corresponding number of at least ten representative tasks.

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

Evaluation of Retrieval System Quality II: Recall and Precision

The specifics of retrieval systems are located in their significance as IT systems that facilitate the search and retrieval of information. The evaluation of retrieval system quality is to be found in the measurements of Recall and Precision (Ch. B.2) as well as in metrics derived from these (Baeza-Yates & Ribeiro-Neto, 2011, 131-176; Croft, Metzler, & Strohman, 2010, 297-338; Harman, 2011; Manning, Raghavan, & Schütze, 2008, 139-161). Tague-Sutcliffe (1992) describes the methodology of retrieval tests. Here are some of the questions that must be taken into consideration before the test is performed:

- To test or not to test? Which innovations (building on the current state of research) are aimed for in the first place?
- What kind of test? Should a laboratory test be performed in a controlled environment? Or are users observed in normal search situations (Ch. H.3)?
- How to operationalize the variables? Each variable to be tested (users, query, etc.) must be described exactly.
- What database to use? Should an experimentation database be built from scratch? Can one draw on pre-existing databases (such as TReC)? Or should a “real-life” information service be analyzed?
- What kind of queries? Should informational, navigational, transactional, etc. queries (Ch. F.2) be consulted? Can such different query types be mixed among one another?
- Where to get queries? Where do the search arguments come from? How many search atoms and how many Boolean operators are used? Should further operators (proximity operators, numerical operators, etc.) be tested?
- How to process queries? It is necessary for the search processes to run in a standardized procedure (i.e. always under the same conditions). Likewise, the test subjects should have at least a similar degree of pre-knowledge.
- How to design the test? How many different information needs and queries are necessary in order to achieve reliable results? How many test subjects are necessary?
- Where do the relevance judgments come from? One must know whether or not a document is relevant for an information need. Who determines relevance? How many independent assessors are required in order to recognize “true” relevance? What do we do if assessors disagree over a relevance judgment?
- How many results? In a search engine that yield results according to relevance, it is not possible (or useful) to analyze all results. But where can we place a practicable cut-off value?
- How to analyze the data? Which effectiveness measurements are used (Recall, Precision, etc.)?

Effectiveness measurements were already introduced in the early period of retrieval research (Kent, Berry, Luehrs, & Perry, 1955). In the Cranfield retrieval tests, Clever-

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

don (1967) uses Recall and Precision throughout. Cranfield is the name of the town in England where the tests were performed.

Table H.4.2: Performance Parameters of Retrieval Systems in the Cranfield Tests. Source: Cleverdon, 1967, 175.

	Relevant	Non-relevant	
Retrieved	a	b	a + b
Not retrieved	c	d	c + d
	a + c	b + d	a + b + c + d = N

Table H.4.2 is a four-field schema that differentiates between relevance/non-relevance and retrieved/not retrieved, respectively. The values for *a* through *d* stand for numbers of documentary units. To wit, *a* counts the number of retrieved relevant documentary units, *b* the number of found irrelevant DUs, *c* the number of missed relevant DUs, and, finally, *d* is the number of not retrieved irrelevant DUs. *N* is the total number of documentary units in the information service. Recall (R) is the quotient of *a* and *a + c*; Precision (P) is the quotient of *a* and *a + b*; finally, Cleverdon (1967, 175) introduces the fallout ratio as the quotient of *b* and *b + d*.

Even though Cleverdon (1967, 174-175) works with five levels of relevance, a binary view of relevance will prevail eventually: a documentary unit is either relevant for the satisfaction of an information need or it is not.

It is possible to unite the two effectiveness values Recall and Precision into a single value. In the form of his E-Measurement, van Rijsbergen (1979, 174) introduces a variant of the harmonic mean:

$$E = 1 - \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}$$

α can assume values between 0 and 1. If α is greater than 0.5, greater weight will be placed on Precision; if α is smaller than 0.5, Recall will be emphasized. In the value α = 0.5, effectiveness is balanced on P and R in equal measure. The greatest effectiveness is reached by a system when the E-value is 0.

Since 1992, the TReC conferences have been held (in the sense of the Cranfield paradigm) (Harman, 1995). TReC provides experimental databases for the evaluation of retrieval systems. The test collection is made up of three parts:

- the documents,
- the queries,
- the relevance judgments.

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

The documents are taken from various sources. They contain, for instance, journalistic articles (from the "Wall Street Journal") or patents (from the US Patent and Trademark Office). The queries represent information needs. "The topics were designed to mimic a real user's need, and were written by people who are actual users of a retrieval system" (Harman, 1995, 15). Table H.4.3 shows a typical TReC query.

Table H.4.3: Typical Query in TReC. Source: Harman, 1995, 15.

Query	
Number:	066
Domain:	Science and Technology
Topic:	Natural Language Processing
Description:	Document will identify a type of natural language processing technology which is being developed or marketed in the U.S.
Narrative:	A relevant document will identify a company or institution developing or marketing a natural language processing technology, identify the technology, and identify one or more features of the company's product.
Concept(s):	1. natural language processing; 2. translation, language, dictionary, font; 3. software applications
Factors:	Nationality: U.S.

The relevance judgments are made by assessors. For this, they are given the following instruction by TReC (Voorhees, 2002, 359):

> To define relevance for the assessors, the assessors are told to assume that they are writing a report on the subject of the topic statement. If they would use any information contained in the document in the report, then the (entire) document should be marked relevant, otherwise it should be marked irrelevant. The assessors are instructed to judge a document as relevant regardless of the number of other documents that contain the same information.

Each document is evaluated by three assessors in TReC. Inter-indexer consistency is at an average of 30% between all three assessors, and at just under 50% between any two assessors (Voorhees, 2002). This represents a fundamental problem of any evaluation following the Cranfield paradigm: relevance assessments are highly subjective.

In order to calculate the Recall, we need values for *c*, i.e. the number of documents that were not retrieved. We can only calculate the absolute Recall for an information service after having analyzed all documents for relevance relative to the queries. This is practically impossible in the case of large databases. TReC works with "relative Recall" that is derived from the "pooling" of different information services. In the TReC experiments, there are always several systems to be tested. In each of them, the queries are processed and ranked according to relevance (Harman, 1995, 16-17):

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

> The sample was constructed by taking the top 100 documents retrieved by each system for a given topic and merging them into a pool for relevance assessment. ... The sample is then given to human assessors for relevance judgments.

The relative Recall is thus dependent upon the pool, i.e. on those systems that just happen to be participating in a TReC experiment. Results for the relative Recall from different pools cannot be compared with one another.

An alternative method for calculating Recall is Availability. This method stems from the evaluation of library holdings and calculates the share of all successful loans relative to the totality of attempted loans (Kantor, 1976). The assessed quantity is known items, i.e. the library user knows the document he intends to borrow. Transposed to the evaluation of retrieval systems, known items, i.e. relevant documents (e.g. specific URLs in the evaluation of a search engine) are designated as the test basis. The documents were not retrieved via the search engine and they are available at the time of analysis. Relevant queries are then constructed from the documents that must make the latter searchable in a retrieval system. Now the queries are entered into the system that is to be tested and the hit list analyzed. The question is: are the respective documents ranked at the top level of the SERP (i.e. among the first 25 results)? The Availability of a retrieval system is the quotient of the number of retrieved known items and the totality of all searched known items (Stock & Stock, 2000).

When determining Precision in systems that yield their results in a relevance ranking, one must determine a threshold value up to which the search results are analyzed for relevance. This cut-off value should reflect the user behavior to be observed. From user research, we know that users seldomly check more than the first 15 results in Web search engines. Here, the natural decision would thus be to set the threshold value at 15. Does it make sense to use the Precision measurement? In the following example, we set the cut-off value at 10. In retrieval system A, let the first five hits be relevant, and the second five non-relevant. The Precision of A is 5 / 10 = 0.5. In retrieval system B, however, the first five hits are non-relevant, whereas those ranked sixth through tenth are relevant. Here, too, the Precision is 0.5. Intuitively, however, we think that system A works better than system B, since it ranks the relevant documentary units at the top.

A solution is provided by the effectiveness measurement MAP (mean average precision) (Croft, Metzler, & Strohman, 2010, 313):

> Given that the average precision provides a number for each ranking, the simplest way to summarize the effectiveness of rankings from multiple queries would be to average these numbers.

MAP calculates two average values: average Precision for a specific query on the one hand, and the average values for all queries on the other. We will exemplify this via an example (Figure H.4.2). Let the cut-off value be 10. In the first query, five documents (ranked 1, 3, 6, 9 and 10) are relevant. The average Precision for Query 1 is 0.62.

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

The second query leads to three results (ranked 2, 5 and 7) and an average Precision of 0.44. MAP is the arithmetic mean of both values, i.e. (0.62 + 0.44) / 2 = 0.53. The formula for calculating MAP is:

$$\mathrm{MAP} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} P(R_{kj})$$

$P(R_{kj})$ is the Precision for the ranking position *k*, *m* is the cut-off value (or—if the hit list is smaller than the cut-off value—the number of hits) and *n* is the number of analyzed queries.

Ranking for Query 1 (5 relevant documents)

Rank	1	2	3	4	5	6	7	8	9	10
relevant?	r	nr	r	nr	nr	r	nr	nr	r	r
P	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

Average Precision: (1.0 + 0.67 + 0.5 + 0.44 + 0.5) / 5 = 0.62

Ranking for Query 2 (3 relevant documents)

Rank	1	2	3	4	5	6	7	8	9	10
relevant?	nr	r	nr	nr	r	nr	r	nr	nr	nr
P	0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

Average Precision: (0.5 + 0.4 + 0.43) / 3 = 0.44

Mean Average Precision: (0.62 + 0.44) / 2 = 0.53

Cut-off: 10; r : relevant; nr: not relevant; P: precision

Figure H.4.2: Calculation of MAP. Source: Modified Following Croft, Metzler, & Strohman, 2010, 313.

Su (1998) suggests gathering user estimates for entire search result lists. “Value of search results as a whole is a measure which asks for a user’s rating on the usefulness of a set of search results based on a Likert 7-point scale” (Su, 1998, 557). Alternatively, it is possible to have a SERP be rated by asking the test subjects to compare this list with another one. The comparison list, however, contains the search results in a random ranking. This way, it can be estimated whether a hit list is better than a random ranking of results.

Evaluation of Evaluation Metrics

Precision, MAP, etc.—there are a lot of metrics for measuring the effectiveness of retrieval systems. Della Mea, Demartini, Di Gaspero and Mizzaro (2006) list more than 40 effectiveness measurements known in the literature. Are there “good” meas-

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

urements? Can we determine whether measurement A is “better” than measurement B? What is needed is an evaluation of evaluation in information retrieval (Saracevic, 1995). Metrics are sometimes introduced “intuitively”, but without a lot of theoretical justification. Why, for instance, are the individual Precision values of the relevant ranking positions added in MAP? Saracevic (1995, 143) even goes so far as to term Recall a “metaphysical measure”.

A huge problem is posed by uncritical usage of the dichotomous 0/1 view of relevance. In relevance research (Ch. B.3), this view is by no means self-evident, as a gradual conception of relevance can also be useful. Saracevic (1995, 143) remarks:

> (T)he findings from the studies of relevance on the one hand, and the use of relevance as a criterion in IR evaluations, on the other hand, have no connection. A lot is assumed in IR evaluations by use of relevance as the sole criterion. *How justifiable are these assumptions?*

Furthermore, the assessors’ relevance assessments are notoriously vague. To postulate results on this basis, e.g. that the Precision of a retrieval system A is 2% greater than that of system B, would be extremely bold.

What is required is a calibration instrument on which we can measure the evaluation measurements. Sirotkin (2011) suggests drawing on user estimates for entire hit lists and to check whether the individual evaluation measurements match these. Such matches are dependent upon the cut-off value used, however. For instance, the traditional Precision for a cut-off value of 4 is “better” than MAP, whereas the situation is completely reversed when the cut-off value is 10 (Sirotkin, 2011).

“The evaluation of retrieval systems is a noisy process” (Voorhees, 2002, 369). In light of the individual evaluation metrics’ insecurities, it is recommended to use as many different procedures as possible (Xie & Benoit III, 2013). In order to glean useful results, it is thus necessary to process all dimensions of the evaluation of retrieval systems (Figure H.4.1).

Conclusion

- A comprehensive model of the evaluation of retrieval systems comprises the four main dimensions IT service quality, knowledge quality, IT system quality and retrieval system quality. All dimensions make a fundamental contribution to the usage (or non-usage) of retrieval systems.
- The evaluation of IT service quality comprises the process of service provision as well as important service attributes. The process component is registered via the sequential incident technique and the critical incident technique. The quality of the attributes is measured via SERVQUAL. SERVQUAL uses a double scale of expectation and perception values.
- Knowledge quality is registered by evaluating the documentary reference units and the documentary units (surrogates) of the database. The information quality of documentary reference units is hard to quantify. Surrogate evaluation involves the KOSs used, the indexing, and the freshness of the database.

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

- Evaluation of IT system quality builds on the technology acceptance model (TAM). TAM has four subdimensions: perceived usefulness, perceived ease of use, trust, and fun. The usefulness of retrieval systems is expressed in their effectiveness (in finding the right information) and in their efficiency (in doing so very quickly).
- Elaborate retrieval systems have functionalities for calling up databases, for identifying search arguments, for formulating queries, for the display and output of surrogates and documents, for creating and managing push services, and for performing informetric analyses. A system's range of commands is a measurement for the quality of its search features. The quality of the system's presentation, as well as that of its functions, is measured via usability tests.
- The traditional parameters for determining the quality of retrieval systems are Recall and Precision. Both measurements have been summarized into a single value: the E-Measurement. Experimental databases for the evaluation of retrieval systems (such as TReC) store and provide documents, queries, and relevance judgments. The pooling of different databases provides an option for calculating relative Recall (relative always to the respective pool). Calculating Precision requires the fixing of a cut-off value. In addition to traditional Precision, MAP (mean average precision) is often used.
- As there are many evaluation metrics, we need a calibration instrument to provide information on "good" measuring procedures, "good" cut-off values, etc. However, as of yet no generally accepted calibration method has emerged.

Bibliography

Adams, D.A., Nelson, R.R., & Todd, P.A. (1992). Perceived usefulness, ease of use, and usage of information technology. A replication. MIS Quarterly, 16(2), 227-247.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). Modern Information Retrieval. The Concepts and Technology behind Search. 2nd Ed. Harlow: Addison-Wesley.

Cleverdon, C. (1967). The Cranfield tests on index language devices. Aslib Proceedings, 19(6), 173-192.

Croft, W.B., Metzler, D., & Strohman, T. (2010). Search Engines. Information Retrieval in Practice. Boston, MA: Addison Wesley.

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(1), 319-340.

Davis, F.D., Bagozzi, R.P., & Warshaw, P.R. (1989). User acceptance of computer technology. A comparison of two theoretical models. Management Science, 35(8), 982-1003.

Della Mea, V., Demartini, L., Di Gaspero, L., & Mizzaro, S. (2006). Measuring retrieval effectiveness with Average Distance Measure. Information – Wissenschaft und Praxis, 57(8), 433-443.

DeLone, W.H., & McLean, E.R. (1992). Information systems success. The quest for the dependent variable. Information Systems Research, 3(1), 60-95.

DeLone, W.H., & McLean, E.R. (2003).The DeLone and McLean model of information systems success. A ten-year update. Journal of Management Information Systems, 19(4), 9-30.

Dillon, A., & Morris, M.G. (1996). User acceptance of information technology. Theories and models. Annual Review of Information Science and Technology, 31, 3-32.

Flanagan, J.C. (1954). The critical incident technique. Psychological Bulletin, 51(4), 327-358.

Gefen, D., Karahanna, E., & Straub, D.W. (2003). Trust and TAM in online shopping. An integrated model. MIS Quarterly, 27(1), 51-90.

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

Harman, D. (1995). The TREC conferences. In R. Kuhlen & M. Rittberger (Eds.), Hypertext – Information Retrieval – Multimedia. Synergieeffekte elektronischer Informationssysteme (pp. 9-28). Konstanz: Universitätsverlag.

Harman, D. (2011). Information Retrieval Evaluation. San Rafael, CA: Morgan & Claypool.

Jennex, M.E., & Olfman, L. (2006). A model of knowledge management success. International Journal of Knowledge Management, 2(3), 51-68.

Kantor, P.B. (1976). Availability analysis. Journal of the American Society for Information Science, 27(5), 311-319.

Kent, A., Berry, M., Luehrs, F.U., & Perry, J.W. (1955). Machine literature searching. VIII: Operational criteria for designing information retrieval systems. American Documentation, 6(2), 93-101.

Kettinger, W.J., & Lee, C.C. (1997). Pragmatic perspectives on the measurement of information systems service quality. MIS Quarterly, 21(2), 223-240.

Knautz, K., Soubusta, S., & Stock, W.G. (2010). Tag clusters as information retrieval interfaces. In Proceedings of the 43rd Annual Hawaii International Conference on System Sciences (HICSS-43), January 5-8, 2010. Washington, DC: IEEE Computer Society Press (10 pages).

Lewandowski, D., & Höchstötter, N. (2008). Web searching. A quality measurement perspective. In A. Spink & M. Zimmer (Eds.), Web Search. Multidisciplinary Perspectives (pp. 309-340). Berlin, Heidelberg: Springer.

Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge: Cambridge University Press.

McKnight, S. (2006). Customers value research. In T.K. Flaten (Ed.), Management, Marketing and Promotion of Library Services (pp. 206-216). München: Saur.

Nielsen, J. (2003). Usability Engineering. San Diego, CA: Academic Press.

Parasuraman, A., Zeithaml, V.A., & Berry, L.L. (1988). SERVQUAL. A multiple-item scale for measuring consumer perceptions of service quality. Journal of Retailing, 64(1), 12-40.

Parker, M.B., Moleshe, V., De la Harpe, R., & Wills, G.B. (2006). An evaluation of information quality frameworks for the World Wide Web. In 8th Annual Conference on WWW Applications. Bloemfontein, Free State Province, South Africa, September 6-8, 2006.

Pitt, L.F., Watson, R.T., & Kavan, C.B. (1995). Service quality. A measure of information systems effectiveness. MIS Quarterly, 19(2), 173-187.

Röttger, M., & Stock, W.G. (2003). Die mittlere Güte von Navigationssystemen. Ein Kennwert für komparative Analysen von Websites bei Usability-Nutzertests. Information – Wissenschaft und Praxis, 54(7), 401-404.

Rubin, J., & Chisnell, D. (2008). Handbook of Usability Testing. How to Plan, Design, and Conduct Effective Tests. 2nd Ed. Indianapolis, IN: Wiley.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval (pp. 138-146). New York, NY: ACM.

Sirotkin, P. (2011). Predicting user preferences. In J. Griesbaum, T. Mandl, & C. Womser-Hacker (Eds.), Information und Wissen: global, sozial und frei? Proceedings des 12. Internationalen Symposiums für Informationswissenschaft (pp. 24-35). Boizenburg: Hülsbusch.

Stauss, B., & Weinlich, B. (1997). Process-oriented measurements of service quality. Applying the sequential incident technique. European Journal of Marketing, 31(1), 33-65.

Stock, M., & Stock, W.G. (2000). Internet-Suchwerkzeuge im Vergleich. 1: Retrievaltest mit Known Item Searches. Password, No. 11, 23-31.

Stock, W.G. (2000). Qualitätskriterien von Suchmaschinen. Password, No. 5, 22-31.

Su, L.T. (1998). Value of search results as a whole as the best single measure of information retrieval performance. Information Processing & Management, 34(5), 557-579.

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

Tague-Sutcliffe, J.M. (1992). The pragmatics of information retrieval experimentation, revisited. Information Processing & Management, 28(4), 467-490.

Tague-Sutcliffe, J.M. (1996). Some perspectives on the evaluation of information retrieval systems. Journal of the American Society for Information Science, 47(1), 1-3.

van Rijsbergen, C.J. (1979). Information Retrieval. 2nd Ed. London: Butterworths.

Voorhees, E.M. (2002). The philosophy of information retrieval evaluation. Lecture Notes in Computer Science, 2406, 355-370.

Xie, I., & Benoit III, E. (2013). Search result list evaluation versus document evaluation. Similarities and differences. Journal of Documentation, 69(1), 49-80.

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15

Bereitgestellt von | Karl-Franzens-Universitaet/Universitaetsbibliothek Graz
Angemeldet
Heruntergeladen am | 08.06.17 15:15