Chapter 3

Information Quality Assessment

Information quality assessment is the process of evaluating if a piece of information meets the information consumer's needs in a specific situation [Nau02, PLW02]. Information quality assessment involves measuring the quality dimensions that are relevant to the information consumer and comparing the resulting scores with the information consumer's quality requirements. Information quality assessment is rightly considered difficult [Nau02] and a general criticism within the information quality research field is that, despite the sizeable body of literature on conceptualizing information quality, relatively few researchers have tackled the problem of quantifying information quality dimensions [NR00, KB05].

This chapter will give an overview about different types of information quality assessment metrics and discuss their applicability within the context of web-based information systems. Afterwards, the concept of quality-based information filtering policies is introduced.

3.1 Assessment Metrics

An *information quality assessment metric* is a procedure for measuring an information quality dimension. Assessment metrics rely on a set of *quality indicators* and calculate an assessment score from these indicators using a *scoring function*. Assessment metrics are heuristics that are designed to fit a specific assessment situation [PWKR05a, WZL00]. The types of information which may be used as quality indicators are very diverse. Beside of information to be assessed itself, scoring functions may rely on meta-information about the circumstances in which information was created, on background information about the information provider, or on ratings provided by the information consumer herself, other information consumers, or domain experts.

Figure 3.1 shows an abstract view on an information exchange situation. All types of information that may be used as quality indicators are shaded gray. The range of employable scoring functions is also diverse [PLW02]. Depending on the quality dimension to be assessed and the chosen quality indicators, scoring functions range from simple comparisons, like "assign true if the quality indicator has a value greater than X", over set functions, like "assign true if the indicator is in the set Y", aggregation functions, like "count or sum up all indicator values", to more complex statistical functions, text-analysis, or network-analysis methods.

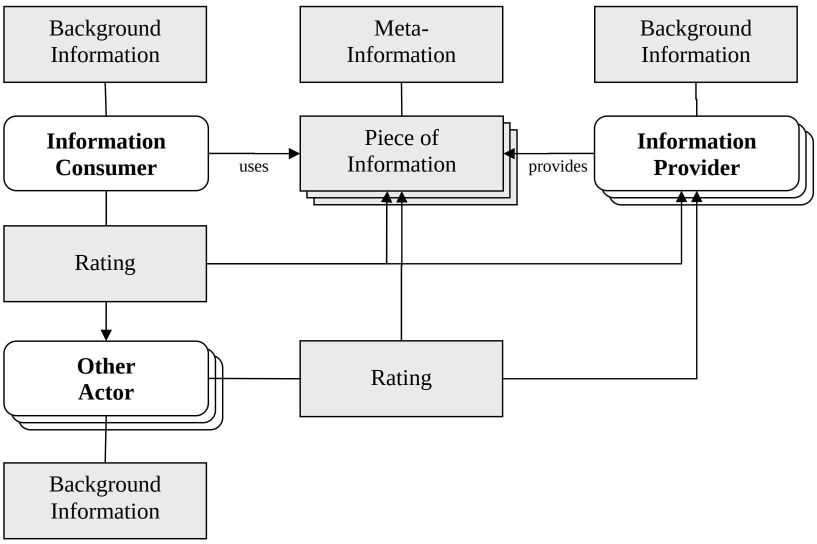

Figure 3.1: Abstract view on an information exchange situation.

Information quality assessment metrics can be classified into three categories according to the type of information that is used as quality indicator:

Content-Based Metrics use information to be assessed itself as quality indicator. The methods analyze information itself or compare information with related information.

Context-Based Metrics employ meta-information about the information content and the circumstances in which information was created, e.g., who said what and when, as quality indicator.

Rating-Based Metrics rely on explicit ratings about information itself, information sources, or information providers. Ratings may originate from the information consumer herself, other information consumers, or domain experts.

Quality indicators were chosen as classification dimension due to the fact that the applicability of different assessment metrics is determined by the quality indicators that are available in a specific assessment situation. The following sections will give an overview of each category.

3.1.1 Content-Based Metrics

An obvious approach to assess the quality of information is to analyze the content of information itself. The following section will discuss different types of assessment metrics which rely on information content as quality indicator. The metrics fall into two groups: Metrics that are applicable to formalized information and metrics for natural language text and loosely-structured documents, such as HTML pages.

Formalized Information

A basic method to assess the quality of a piece of formalized information is to compare its value with a set of values that are considered acceptable. For instance, a metric to assess the believability of a sales offer could be to check if the price lies above a specific boundary. If the price is too low, the offer might be considered bogus and therefore not believable.

Very often, a set of formalized information objects contains objects that are grossly different or inconsistent with the remaining set. Such objects are called *outliers* and there are various statistical methods to identify outliers [KNT00]. Within the context of information quality assessment, outlier detection methods can be used as heuristics to assess quality dimensions such as accuracy or believability. Outlier detection methods can be classified into three categories [HK01]: Distance-based, deviation-based and distribution-based methods. Distance-based methods consider objects as outliers which differ more than a predefined threshold from the centroid of all objects. Which method is used to calculate centroids depends on the scales of the variables describing the objects [KNT00]. For instance, for a set of objects described by a single ratio-scaled variable the centroid can be calculated as the median or arithmetic mean. Deviation-based methods identify outliers by examining the main characteristics of objects within a set of objects. Using a dissimilarity function, outliers are identified as the objects whose removal from the set results in the greatest reduction of dissimilarity in the remaining set [HK01]. Distribution-based outlier detection methods assume a certain distribution or probabilistic model for a given set of values (e.g. a normal distribution) and identify values as outliers which deviate from this distribution [HK01].

Natural Language Text

Various text analysis methods can be employed to assess the quality of natural language texts or loosely-structured documents, such as HTML pages. In general, text analysis methods derive assessment scores by matching terms or phrases against a document and/or by analyzing the structure of the document. Within deployed web-based information systems, text analysis methods are used to assess the relevancy of documents, to detect spam, and to scan websites for offensive content.

Seen from an information quality assessment perspective, the complete Information Retrieval [GF04] research field is concerned with developing assessment metrics for a single quality dimension: Relevancy. In general, information retrieval methods consider a document to be relevant to a search request if search terms appear often and/or in prominent position in the document. An example of a well known information retrieval method is the vector space model [GF04]. Within the vector space model, documents as well as search requests are represented as vectors of weighted terms. The vector space model can be combined with different indexing functions to extract content-bearing terms from documents and to assign weights to these terms. Indexing functions are usually optimized for a specific class of documents. For instance, in the case of HTML pages, an indexing function might assign higher weights to terms that appear in the document title or in headings within a document.

In contrast to information retrieval methods which aim at identifying relevant content, spam detection methods try to identify irrelevant content. Two general types of spam content appear on the public Web, in discussion forums, or as feedback comments within weblogs: Content that advertises some product or service and content that aims at misleading search engines into giving certain websites a higher rank (*link-spam*) [GGM05]. The first type of spam can be detected by matching content against a black-list of suspicious terms and by classifying content as spam if these terms occur with a certain frequency or form certain patterns within the content. Link-spam cannot be detected using a term black-list as it contains arbitrary terms related to the content of the target page. Thus, Ntoulas et al. propose to identify link-spam by employing statistical approaches that rely on the number of different words in a page, or in the page title, the average length of words, the amount of anchor text, the fraction of stop-words within a page and the fraction of visible content [NNMF06]. A similar approach that is optimized towards weblogs is proposed by Narisawa et al. [NYIT06].

Automated text analysis methods are generally imprecise. Approaches to increase their precision include the usage of stemming algorithms, which reduce words to their root form [Hul96], and the usage of thesauri and ontologies for the disambiguation of terms [VFC05]. Text analysis-based assessment metrics are often combined with context- or rating-based assessment metrics in order to increase precision.

3.1.2 Context-Based Metrics

Context-based information quality assessment metrics rely on meta-information about information itself and about the circumstances in which information was created as quality indicator. Web-based information systems often capture basic provenance meta-information, such as the name of an information provider and date on which information was created. Other systems record more detailed meta-information like abstracts, keywords, language or format identifiers. Meta-information is often included directly into web content, for instance in the case of HTML documents, using `<meta>` tags in the `<head>` section of a document.

This section gives an overview about standards for representing meta-information about web content. Afterwards, it discusses how meta-information can be used as quality indicator to assess different quality dimensions.

Meta-Information Standards

Several standards have evolved for representing common types of meta-information about web content. Publishing meta-information according to these standards reduces ambiguity and enables the automatic processing of meta-information. The following section gives an overview of popular standards for representing general purpose meta-information, licensing information, and for classifying web content.

Dublin Core Element Set. The Dublin Core Element Set is a widely used standard for representing general purpose meta-information, such as author and creation date [ISO03a]. The standard has been developed by the Dublin Core Meta Data Initiative[1] in order to facilitate the discovery of electronic resources. The Dublin Core Element Set consists of 15 meta-information attributes. The standard does not restrict attribute values to a single fixed set of terms, but allows different vocabularies to be used to express attribute values. Table 3.1 gives an overview of the Dublin Core Element Set. The right column contains a definition for each attribute and refers to standards that are recommended by the

[1] http://dublincore.org/ (retrieved 09/25/2006)

Element	*Definition and recommended value formats*
Title	A name given to the resource. Value format: Free text.
Creator	An entity primarily responsible for creating the content of the resource. Value format: Name as free text.
Subject	A topic of the content of the resource. Value formats: Library of Congress Subject Headings (LCSH) [Ser06], Medical Subject Headings (MeSH) [Nat06], Dewy Decimal Classification (DDC) [Onl03].
Description	An account of the content of the resource. Value format: Free text.
Publisher	An entity responsible for making the resource available. Value format: Name as free text.
Contributor	An entity responsible for making contributions to the content of the resource. Value format: Name as free text.
Date	The date when the resource was created or made available. Value Format: W3C-DTF [WW97].
Type	The nature or genre of the content of the resource. Value Format: DCMI Type Vocabulary [DCM04].
Format	The physical or digital manifestation of the resource. Value Format: MIME-Type [FB96].
Identifier	An unambiguous reference to the resource within a given context. Value Formats: String or number conforming to a formal identification system, such as Uniform Resource Identifier (URI) [BLFM98], Digital Object Identifier (DOI) [Int06b], or International Standard Book Number (ISBN) [Int01].
Source	A Reference to a resource from which the present resource is derived. Value Formats: String or number conforming to a formal identification system.
Language	A language of the intellectual content of the resource. Value Formats: RFC 3066 [Alv01], ISO 639 [ISO98].
Relation	A reference to a related resource. Value Format: String or number conforming to a formal identification system.
Coverage	The extent or scope of the content of the resource. Value formats for spacial locations: Thesaurus of Geographic Names [Get06], ISO 3166 [ISO97]. Value formats for temporal period: W3C-DTF [WW97]
Rights	Information about rights held in and over the resource. Value Format: No recommendation.

Table 3.1: The Dublin Core Element Set [ISO03a].

Dublin Core specification [ISO03a] to express attribute values. Dublin Core is widely used for annotating HTML documents. The standard is also widespread in content management systems, library-, and museum information systems.

Creative Commons Licensing Schemata. Information providers might annotate web content with licensing information in order to restrict the usage of content or to dedicate content to the public domain. A widely used schema for expressing licensing information is the Creative Commons[2] term set. Creative Commons licenses enable copyright holders to grant some of their rights to the public while retaining others through a variety of licensing and contract schemes. Using the Creative Commons term set, copyright holders can, for instance, allow content to be copied, distributed, and displayed for personal purposes but prohibit its commercial use.

ICRA Content Labels. Internet Content Rating Association (ICRA)[3] has created a content description system which allows information providers to self-label their content in categories such as nudity, sex, language, violence, and other potentially harmful material. For each category, there is a fixed set of terms indicating different types of potentially offensive content. ICRA labels can be scoped to contexts such as art, medicine, or news as an information consumer might consider a piece of content containing depictions of nudes only acceptable within a medical context. ICRA labels are used by content-filtering software to block web pages that users prefer not to see, whether for themselves or their children.

Using Meta-Information for Quality Assessment

Meta-information can be used as quality indicator for assessing several quality dimensions. The following section matches information quality dimensions with meta-information attributes and outlines exemplary assessment metrics.

Relevancy. Meta-information attributes such as `title`, `description` and `subject` classify and summarize content. The values of these attributes can be used as indicators for assessing whether content is relevant for a specific task. For instance, a basic metric to assess the relevancy of

[2] http://creativecommons.org/ (retrieved 09/25/2006)
[3] http://www.icra.org/ (retrieved 09/25/2006)

information is to count the occurrence of relevant terms within these attributes. A more sophisticated approach could use an information retrieval method, like the vector space model [GF04], and assign a higher weight to terms that appear within the meta-information attributes. Information consumers might want to use web content for commercial purposes and therefore only consider content as relevant that fulfills certain licensing requirements. In these cases, licensing meta-information, like Creative Commons labels, can be used to distinguish relevant from irrelevant content.

Believability. An important indicator for assessing the believability of web content is meta-information about the identity of the information provider. That is, assumptions about the believability of information providers are extended to information they provide. An example for a simple heuristic to assess the believability of information is to check whether an information provider is contained in a list of trusted providers. Other meta-information that might influence believability are the identities of the `contributors` and the `publisher` of information as well as the source from which information is retrieved.

Timeliness. The obvious indicator for assessing timeliness is the information creation date.

Understandability. A prerequisite for an information consumer to understand web content is that the content is expressed in a language that the information consumer understands. Therefore meta-information about the `language` of web content can be used to assess principal understandability.

Offensiveness. Information consumers might regard sexually explicit or violent web content as offensive. If the content is labeled with Internet Content Rating Association labels, these labels can be used to assess whether content should be regarded as offensive or not.

Beside of relying solely on meta-information, information quality assessment metrics can also combine meta-information with background information about the application domain. A metric for assessing the believability dimension could, for instance, be based on the role of an information provider in the application domain ("Prefer product descriptions published by the manufacturer over descriptions published by a vendor" or "Disbelieve everything a vendor says about its competitor."), his membership in a specific group ("Believe only information from authors working for certain companies.") or his former activities ("Believe only information from authors who have already published several times on a topic." or "Believe only reports from stock analysts whose former predictions proved correct to a certain percentage.").

3.1.3 Rating-Based Metrics

Rating-based information quality assessment metrics rely on explicit or implicit ratings as quality indicators. Many web-based information systems employ feedback mechanisms and allow information consumers to rate content. Examples of well-known websites that use content rating are Amazon[4] where users rate the quality of book reviews, Google Finance[5] where users rate the quality of discussion forum postings, and Slashdot[6] where users rate technical news and comments about these news.

The design of rating systems has been widely studied in computer science. Jøsang et al. present a survey of deployed rating systems and analyze current development trends within this area [JIB06]. A second survey is presented by Mui et al. [MMH02]. As rating systems can be seen as on-line versions of classical off-line surveys, all principles of sound empirical evaluations, that have been developed within the empirical social sciences [Sim78], also apply to ratings systems.

Seen from an abstract perspective, rating-based quality assessment involves two processes: The acquisition of ratings and the calculation of assessment scores from these ratings. Figure 3.2 gives an overview about the elements of both processes. In order to successfully employ rating systems for quality assessment, it is important to understand the interplay of the elements.

The Rating Process

There are various options to design the rating process. The different options can be outlined along three dimensions: What is rated? Who is rating? Which rating schema is used? The design choices along these dimensions are determined by the concrete application situation.

What is rated? The first dimension is the object to be rated. Ideally, information quality ratings should be as fine grained as possible, meaning that individual facts should be rated. But as it is not practicable within

[4]http://www.amazon.com (retrieved 09/25/2006)
[5]http://finance.google.com/finance (retrieved 09/25/2006)
[6]http://slashdot.org (retrieved 09/25/2006)

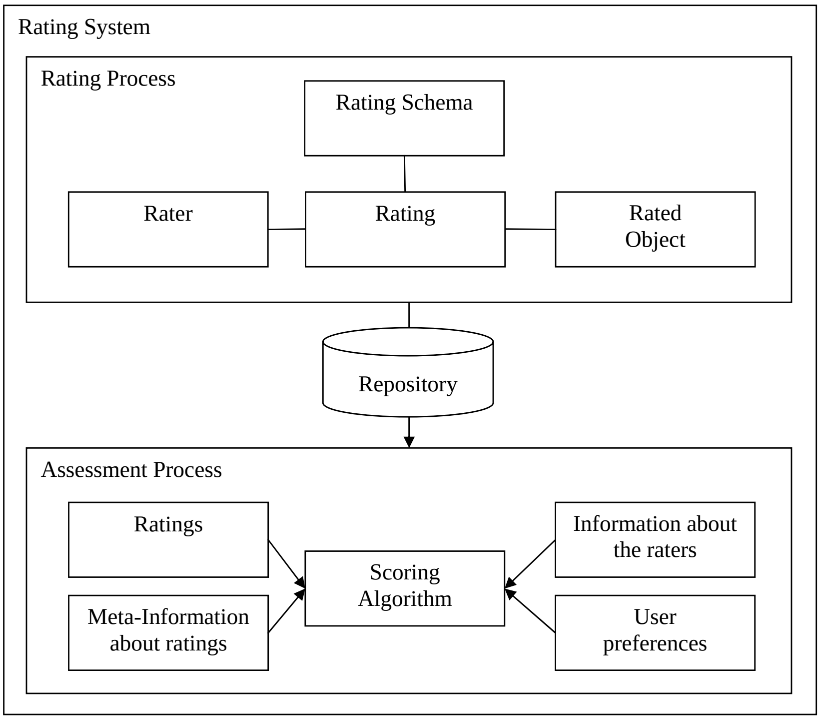

Figure 3.2: Elements of a rating system.

most application scenarios to require raters to rate individual facts, rating systems alternatively collect ratings about all information within a document or data source, e.g. a website or a database [Nau02], or ratings about the general ability of an information provider to provide quality information [GM06]. These approaches are based on the assumption that data sources and information providers generally provide information of similar quality. An assumption which is clearly questionable as data sources may contain information from multiple providers and as the expertise of a single information provider varies from topic to topic.

Who is rating? Ratings can be provided by the information consumer herself, other information consumers, or domain experts. The quality of acquired ratings largely depends on the expertise of the raters. Public websites often allow users to anonymously rate content. This is problematic as the quality of acquired ratings is uncertain [FR98]. An alternative to anonymous ratings is to allow only registered users to rate content. This enables background information about raters to be used to assess the quality of ratings. Ideally, raters should be domain experts. Authors with a management perspective on information quality assessment usually recommend having experts review information and to compensate them for their work [WZL00, Red01]. In the context of web-based systems it is problematic to compensate experts, as in most cases the business models of public websites do not provide for compensations. Therefore, quality is traded for quantity by allowing anonymous ratings. An exception are commercial rating agencies like CyberPatrol[7] or Net Nanny[8] which build their business model on rating the offensiveness of websites.

Which rating schema is used? A schema for capturing information quality ratings defines a set of questions and specifies the ranges of possible answers. Rating schemata have to find a balance between the requirements of sound empirical evaluations and the willingness of raters to spend time on filling questionnaires. In the case of public websites, this often leads to extremely minimalistic rating schemata, like Amazon's "Was this review helpful to you?" with a yes-or-no answer.

The significance of rating-based information quality assessment depends on the availability of a sufficiently large number of ratings and the quality of these ratings. Both, availability and quality of ratings, can be problematic due to several problems associated with the rating process:

Subjectivity. Ratings are subjective. This is especially true for contextual quality dimensions such as relevancy, understandability, or believability and in situations where the perception of quality depends on the information consumer's individual taste.

Unfair Ratings. Raters might try to influence rating systems by providing unfair ratings [Del00]. Raters can provide unjustified high or unjustified low ratings (*bad-mouthing*) and can try to flood rating systems with more than the legitimate number of ratings (*ballot stuffing*). Rating systems can try to confine ballot stuffing by clearly identifying raters and by increasing the effort required to generate multiple pseudonyms [FR98]. In addition, the systems can try to identify and exclude ratings that are likely to be unfair from the calculation of the assessment score [Lev02]. Approaches to detecting unfair ratings rely either on statistical analysis [Del00] or on additional ratings about the trustworthiness of raters [GM06, CDC02].

[7] http://www.cyberpatrol.com/ (retrieved 09/25/2006)
[8] http://www.netnanny.com/ (retrieved 09/25/2006)

Motivation of the Rater. Most websites that collect information quality ratings do not provide direct incentives to raters. Therefore, the motivation of the raters is a fundamental problem as there is no rational reason for providing ratings, and a potential for free-riding by letting the others do the rating [JIB06]. An alternative to providing material incentives, like discounts or cash is to try to motivate users with immaterial incentives in the form of status or rank. This approach is successfully practiced by Amazon[9], which awards *Top Reviewer* badges to active reviewers and maintains a *Top Reviewer* list[10].

An alternative to collecting explicit ratings is to treat other types of information as implicit ratings. This approach is very successfully used by Google[11], which relies on external links pointing to a web page to assess the relevancy of the page [PBMW98]. Other approaches rely on user activities and treat content which is frequently accessed as especially relevant [EM02, LYZ03].

The Assessment Process

Within the assessment process, a scoring function calculates assessment scores form the collected ratings. The scoring function decides which ratings are taken into account and might assign different weights to ratings. Scoring functions should fulfill the following requirements [DFM00]: They should be capable to deal with subjective ratings. They should be robust against unfair ratings and ballot-stuffing attacks. The calculation should be comprehensible, so that information consumers can check assessment results. Designing scoring functions is a popular research topic and various authors have proposed different algorithms. Approaches to classifying the proposed algorithms are presented by Zhang et al. [ZYI04] and Ziegler [Zie05]. The following section gives an overview of several, popular classes of scoring functions:

Simple Scoring Functions. An example of a very simple form of computing assessment scores is to sum the number of positive and negative ratings separately, and to keep a total score as the positive minus the negative score. This scoring function is used by eBay[12] to assess the

[9] http://www.amazon.com (retrieved 09/25/2006)
[10] http://www.amazon.com/exec/obidos/tg/cm/top-reviewers-list/-/1/002-3019175-1616068 (retrieved 09/25/2006)
[11] http://www.google.com (retrieved 09/25/2006)
[12] http://www.ebay.com (retrieved 09/25/2006)

trustworthiness of traders. The utility of the function for the specific use case and its implications on the behavior of traders have been empirically examined by Resnik [RZ02]. A slightly more advanced algorithm is to compute the score as the average of all ratings. This principle is used by numerous commercial websites, such as Amazon[13] or Google Finance[14]. Other models in this category compute a weighted average of all ratings, where the rating weight is determined by factors such as the age of the rating or the reputation of the rater [JIB06]. The advantage of simple scoring algorithms is that anyone can understand the principle behind the score. Their disadvantage is that they do not deal very well with subjective or unfair ratings.

Collaborative Filtering. An approach to overcome the subjectivity of ratings is collaborative filtering [ZMM99, CS01]. Collaborative filtering algorithms select raters that have rated other objects similar as the current user and take only ratings from these raters into account. A criticism against collaborative filtering algorithms is that they take a too optimistic world view and assume all raters to be trustworthy and sincere. Thus, collaborative filtering algorithms are vulnerable to unfair ratings [JIB06].

Web-of-Trust Algorithms. A class of algorithms that are capable of dealing with subjective as well as unfair ratings are web-of-trust algorithms. The algorithms are based on the idea to use only ratings from raters who are directly or indirectly known by the information consumer. The algorithms assume that members of a community have rated each other's trustworthiness, or in the case of information quality assessment, each other's ability to provide quality information. For determining assessment scores, the algorithms search for paths of ratings connecting the current user with the information provider and take only ratings along these paths into account. An example of a web-of-trust algorithms is TidalTrust proposed by Golbeck [GM06]. The TidalTrust algorithm is described in detail in Section 9.6.2.

Flow Models. A further approach to deal with subjectivity and unfair ratings are flow models [JIB06]. The models start with an equal initial score for each member of a community. Based on the ratings of community members for each other, the initial values are increased or decreased. Participants can only increase their score at the cost of others. The process of transferring scoring points from one participant

[13] http://www.amazon.com (retrieved 09/25/2006)
[14] http://finance.google.com/finance (retrieved 09/25/2006)

to the next is iteratively repeated until the variations shrink close to an equilibrium. This leads to a flow of scoring points along arbitrary long and possible looped chains across the network. Examples of flow models are Google's PageRank algorithm [PBMW98], the Appleseed algorithm [ZL04], and Advogato's reputation scheme [Lev02].

It is difficult for an information system to explain the results of web-of-trust and flow model-based scoring functions to the user. This lack of traceability might be one of the reason why most major websites use simple scoring functions.

The preceding sections gave an overview of different information quality assessment metrics. The next section critically discusses the accuracy of assessment results. Afterwards, it is shown how multiple assessment metrics are combined in a concrete application situation.

3.2 Accuracy of Assessment Results

Information quality assessment results should be as accurate as possible or should, at least, be accurate enough to be useful for the information consumer. But, as many practitioners state, information quality assessment results are usually imprecise [NR00, KB05, Pie05]. This is due to several practical as well as theoretical problems associated with information quality assessment:

Quantification of Information Quality Dimensions. Concepts like believability, relevancy, or offensiveness are rather abstract and there are no agreed-upon definitions for these concepts [WSF95, KB05]. Therefore, it is also difficult to find theoretically convincing quantifications for them [PWKR05b]. One theoretical approach to operationalize information quality dimensions by grounding their definitions in an ontology was presented by Ward and Wang [WW96], but has not been taken up by the research community so far. In practice, information quality assessment metrics are usually designed in an ad-hoc manner, often relying on trial-and-error, to fit a specific assessment situation. Assessment metrics therefore have to be understood as heuristics [PWKR05a, WZL00].

Subjectivity of Information Quality Dimensions. Information quality dimensions like relevancy or understandability are subjective and scores for these dimensions can only be precise for individual information consumers, never for an entire group. Thus, assessing scores for subjective quality dimensions requires input from the information consumer. The willingness of an information consumer to provide such input depends on the relevancy of the assessment results for the information consumer. Practical experience from web-based systems shows that information consumers are often only willing to spend very little time on providing input [Nau02]. It is therefore often necessary to employ assessment metrics that are less precise but minimize the input required from the information consumer.

Availability of Quality Indicators. The applicability of information quality assessment metrics is determined by the availability of the quality indicators that are required by a metric. Which quality indicators are available depends on the concrete assessment situation. Within a closed company setting, it is often possible to install binding procedures to capture quality-related meta-information within normal business processes. For instance, information systems within companies often require users to properly authenticate themselves, allowing information to be clearly associated to a user. Within a Web setting, where information providers are autonomous, is is not possible to install binding procedures for capturing quality-related meta-information. Therefore, meta-information is often not available or incomplete. In the worst case, information quality assessment metrics can only rely on the URL from which information was retrieved, the retrieval date, and the information itself, as quality indicators.

Quality of Quality Indicators. A further problem is the quality of indicators themselves. Ratings may, for instance, be inaccurate or biased as the rater might not be an expert on the rated topic or as she might try to mislead the rating system by providing unfair ratings [Lev02, FR98]. A further example for the varying quality of quality indicators is meta-information that is included into web documents such as HTML pages. Some information providers misuse meta-information in an attempt to mislead search engines into listing them at a prominent position in the search results. Other information providers do not bother with providing meta-information about their documents at all.

Because of these problems, information quality assessment often has to trade accuracy for practicability [Nau02]. A factor which relativizes the imprecision of assessment results is that information consumers are often satisfied with approximate answers in the context of web-based systems. For them, the utility of the Web lies in this vast amount of accessible information

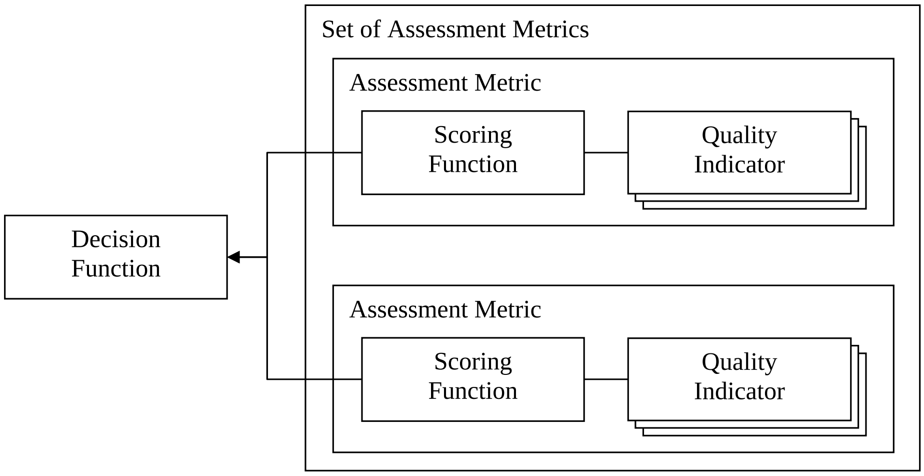

Figure 3.3: Elements of a quality-based information filtering policy.

and the benefit of having access to a huge information-base is often higher than the costs of having some noise in the answers [Nau01].

Therefore, the goal of practical information quality assessment in the context of web-based information systems is to find heuristics which can be applied in a given situation and that are sufficiently precise to be useful for the information consumer.

3.3 Information Filtering Policies

This section establishes the concept of quality-based information filtering policies. Quality-based information filtering policies are heuristics for deciding whether to accept or reject information to accomplish a specific task. The decision whether to accept of reject information is a multi-criteria decision problem [Tri04, YH06].

A quality-based information filtering policy consists of a set of assessment metrics, for assessing the quality dimensions that are relevant for the task at hand and a decision function which aggregates the resulting assessment scores into an overall decision whether information satisfies the information consumer's quality requirements. Each assessment metric relies on a set of quality indicators and specifies a scoring function to calculate an assessment score from these indicators. The decision function weights assessment scores depending on the relevance of the different quality dimensions for the task at hand. Figure 3.3 illustrates the elements of a quality-based information filtering policy.

Information consumers may choose a wide range of different policies to decide whether to accept or reject information. The following section describes the factors that influence policy selection. Afterwards, two example policies that an investor could apply to assess the quality of analyst reports are developed.

3.3.1 Policy Selection

An information consumer who wants to determine whether to accept or reject information has to answer the following questions: Which information quality dimensions are relevant in the context of the task at hand? Which information quality assessment metric should be used to assess each dimension? How should the assessment results be compiled into an overall decision whether to accept or reject information?

The relevance of the different quality dimensions is determined by the task at hand. The choice of suitable assessment metrics for specific quality dimensions is restricted by several factors:

Availability of Quality Indicators. Whether an assessment metric can be used in a specific situation depends on the availability of the quality indicators that are required by the metric. Assessing quality dimensions like timeliness is possible in many cases, as the required quality indicators are often available. Accessing other dimensions like accuracy or objectivity often proves difficult, as it might involve the information consumer or experts verifying or rating information.

Quality of Quality Indicators. The choice of assessment metrics is also influenced by the quality of the available quality indicators. If an information consumer is in doubt about the quality of certain indicators, he might prefer to chose a different assessment metric which relies on other indicators.

Understandability. The key factor for an information consumer to trust assessment results is his understanding of the assessment metric. Therefore, relatively simple, easily understandable, and traceable assessment metrics are often preferred.

Subjective Preferences. The information consumer might have subjective preferences for specific assessment metrics. He might, for example, consider specific quality indicators and scoring functions more reliable then others. Thus, there is never a single best policy for a specific task, as the subjectively best policy differs from user to user.

3.3.2 Example Policies

Financial information portals like Wall Street Journal Online[15], Bloomberg[16], Yahoo Finance[17] and Google Finance[18] enable investors to access a multitude of financial news, analyst reports, and postings from investment related discussion forums. Investors using these portals face several information quality problems:

1. Different information providers have different levels of knowledge about specific markets and companies. Therefore, judgments from certain sources are more accurate than judgments from other sources.

2. Different information providers have different intentions. Companies tend to present their own activities and future prospects as positively as possible. The emitters of funds and other securities often also publish market analysis and may be tempted to use the analysis for marketing their products. Investors may try to influence quotes by launching rumors or biased information.

3. The huge amount of accessible information obscures relevant information.

This section develops two alternative policies that an investor could apply to assess the quality of analyst reports provided by a financial information portal. Let us assume that the investor is interested in reports about three different companies. As he speaks English and German, he is willing to accept reports in both languages. The portal provides the following meta-information about each report: Ticker symbol of the stock covered in the report, publication date of the report, language of the report, author of the report, affiliation of the author. The portal enables investors to rate analysts and provides these ratings to its users.

For the task of selecting high quality analyst reports, the investor might consider the following information quality dimensions as equally relevant: Believability, relevancy, timeliness, and understandability. In the light of the available quality indicators, the investor might decide to use the assessment metrics shown in Table 3.2 to access the different dimensions. The chosen metrics translates into the following overall filtering policy: "Accept only research reports that cover one of the three stocks, are written in German or

[15] http://online.wsj.com/ (retrieved 09/25/2006)
[16] http://quote.bloomberg.com/ (retrieved 09/25/2006)
[17] http://finance.yahoo.com/ (retrieved 09/25/2006)
[18] http://finance.google.com/ (retrieved 09/25/2006)

Quality Dimension	Relevancy
Quality Indicator	Stock ticker symbol
Scoring Function	Return true if the report covers one of the three relevant stocks, otherwise false.
Quality Dimension	Timeliness
Quality Indicator	Publication date
Scoring Function	Return true if date lies within the last month, otherwise false.
Quality Dimension	Understandability
Quality Indicator	Meta-information about the language of the report.
Scoring Function	Check if language is German or English.
Quality Dimension	Believability
Quality Indicator	Author of the report, ratings about the analyst from other investors.
Scoring Function	True, if an analyst has received more positive than negative ratings, oterhwise false.
Decision Function	Accept a report if all four scores are true.

Table 3.2: Policy for selecting analyst reports.

Quality Dimension	Believability
Quality Indicator	Affiliation of the author
Background Knowledge	List of trustworthy analyst houses
Scoring Function	Return true if the author works for a trustworthy analyst house, false otherwise.

Table 3.3: Alternative metric for assessing the believability dimension.

English, are not older than a month and have been written by analysts who have received more positive than negative ratings from other investors."

Our investor might be uncertain about the quality of the ratings, as he might not trust the judgments of the other users of the portal. Based on his past experience, he might prefer analyst reports which originate from certain analyst houses. Therefore, he could decide to choose an alternative metric to assess the believability dimension and require the authors of reports to work for one of the analyst houses that he considers trustworthy. This alternative heuristic is shown in Table 3.3.

3.4 Summary

This chapter gave an overview of different metrics to assess the quality of information. The metrics were classified according to the type of information that is used as quality indicator into three categories: Content-, context- and rating-based information quality assessment metrics.

Afterwards, the concept of quality-based information filtering policies was established. Quality-based information filtering policies combine several assessment metrics with a decision function in order to determine whether information satisfies the information consumer's quality requirements for a specific task. Information consumers use a wide range of different filtering policies. Which policy is selected depends on the task at hand, the available quality indicators, the quality of these indicators, and the subjective preferences of the information consumer.

Section 3.2 identified several factors that restrict the accuracy of information quality assessment. Two important factors, in the context of web-based information systems, are the availability and the quality of quality indicators. Because of the decentralized nature of the Web and the autonomy of information providers, meta-information about web content is often incomplete. The quality of ratings is often uncertain, as the expertise of raters is unknown in many cases.

The imprecision of assessment results is relativized by the fact that users of web-based information systems are accustomed to tolerate a certain amount of minor quality information. For them, the benefit of having access to a huge information-base is often higher than the costs of having some noise in the answers [Nau01]. Therefore, the goal of practical information quality assessment is to find heuristics which can be applied in a given situation and that are sufficiently precise to be useful from the perspective of the information consumer.