Quality makes the information market[*]

B. van Gils, H.A. (Erik) Proper, P. van Bommel, Th.P. van der Weide

Institute for Computing and Information Sciences, Radboud University Nijmegen
Toernooiveld 1, 6525 ED Nijmegen, The Netherlands, EU

Abstract. In this paper we consider information exchange via the Web to be an information market. The notion of quality plays an important role on this information market. We present a model of quality and discuss how this model can be operationalized.

This leads us to quality measurement, interpretation of measurements and the associated accuracy. An illustration in the form of a basic quality assessment system is presented.

1 Introduction

The amount of information available to us has been increasing at an explosive rate over the last few years, especially with the massive growth of the Web. Several tools and systems have been developed to help us manage the vast amount of available resources such as indexes, search engines, catalogues and so on. These tools can, to some extent, be seen as information retrieval tools.

Basic background on information retrieval is found in [17] and [16]. The traditional information retrieval paradigm is introduced in Figure 1.
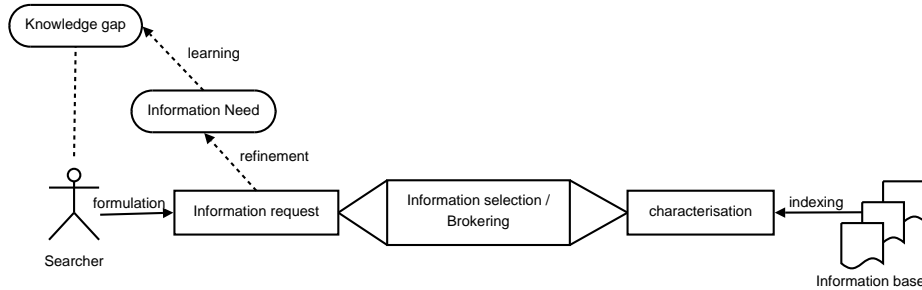

Fig. 1: The information retrieval paradigm

A main challenge in information retrieval is the correct formulation of information requests. See the left part of Figure 1. In the middle and right part, we see brokering or matching, and characterisation.

Characterising supply: Good characterisation of resources is imperative for effective information discovery, as poor characterisations inevitably lead to the retrieval of irrelevant information, or omit relevant information.

Matching demand and supply: The selection of relevant resources for a given query is a well understood problem. The field of information retrieval has developed a number of retrieval models.

Besides traditional performance measures such as precision and recall, a notion of quality in the broader sense is essential in modern information retrieval. Relevant quality aspects are the following. What is the quality of the characterisation of resources? What qualities do resources have? What is the quality of a query? How well is a query formulated and how accurately does it describe the searchers information need? What is the quality of a search engine or match maker? What are its qualities?

Quality has received a lot of attention in the general area of computing and information science. Section 2 contains examples and references to the literature. The specific domain of the Web information market still lacks a proper notion of quality, though. Our research focus is to work towards a generic model of quality of resources and to show how it can be used in the context of the Web. This research on quality is part of a larger project on information supply on the Web. This project has been introduced in [7]. More details about our approach of information supply were presented in for instance [1] and [8].

The organisation of this paper is as follows. In Section 2 we give a survey of the literature on quality in general. Then we discuss our quality model in Section 3. We discuss the operationalisation of our quality model for the Web context in Section 4. Finally, we illustrate an application in terms of an example quality assessment system in Section 5, while Section 6 gives conclusions and directions for future research.

2 Background

From the dictionary definition of quality[1] we learn that the notion of quality has two distinct interpretations: (1) a distinguishing characteristic similar to a property, and (2) inherent or intrinsic excellence, in other words how *good* some artefact is.

---

[1] We have consulted *Webster's third new international dictionary, unabridged* and *Concise Oxford Dictionary*

The term quality has a long history, for example in his work on *the philosophy of nature*, Aristotle used the notion of quality[2]. In his view, quality is the category according to which objects are said to be like or unlike. Other great philosophers such as Descartes, Bacon, Newton, and Galileo oppose to Aristotle's view on quality[3] mainly because they make a distinction between objective qualities and largely subjective qualities.

In e-commerce the notion of quality plays an important role. Two main examples in this respect are the problem of uncertainty with regards to the product/service to be traded and the lack of quality information about the traded assets, see [19] and [11] respectively. These problems are similar in the field of operations management where one is mainly concerned with key dimensions of quality such as product attributes, product performance, warranty, and service availability This is discussed in [10]. A *conformance to specification* approach is very popular in this field, but is criticized in [12] because it would focus too much on the supplier perspective, whereas the consumer perspective would focus on value for money. In [15] the focus is on the ex-post evaluation of quality of information in organizations, based on the ISO-8402 definition of quality:

> The totality of features and characteristics of a product, process or service that bear on its ability to satisfy stated or implicit goals.

Using this definition, in [15] a dual view on quality is proposed. On the one hand, the causal point of view deals with the quality of information, seen as the result of the quality of the process in which it is produced. On the other hand, in the teleological point of view the quality of information is seen as the degree to which it satisfies stated or implicit needs, derived from the situation in which it is used.

In the field of software engineering, the notion of quality plays the role of quality of software, and quality of the software engineering process. In this field the emphasis is mainly on quality attributes. See for instance [18]. Examples of these attributes are safety, security, reliability, resilience, robustness and learnability. Furthermore software quality management can be structured into three principle activities: quality assurance, quality planning, and quality control. See [3, 2, 13, 6] for further discussions on quality in the context of software engineering.

In [5] a discussion on the quality of data on the Web is presented. This discussion starts with the observation that "well-founded and practical approaches to assess or even guarantee a required degree of the quality of data are still missing". In order to overcome this defect the authors propose that a quality algebra be used for dealing with quality issues on the web. Such an algebra is particularly useful for intermediaries on the information market. See for example [7]. In [20] it is posed that user concerns about their perception of the quality of information on the Web continues to be a strong incentive for "the emergence and success

---

[2] http://www.utm.edu/research/iep/a/aristotl.htm and http://www.newadvent.org/cathen/12589c.htm

[3] http://www.ul.ie/~philos/vol1/eustac1.html

of information intermediaries". They can play an important role in the trust relationship between suppliers and consumers, as well as in the control of quality versus price. Last but not least, the approach of [14] is that data quality is the measure of the agreement between the data view presented by an information system and some data in the real world.

3 A model of quality

Upon closer examination, the above definitions and applications of quality show that there are three main views on quality:

Property: the quality properties some artefact may poses.

Excellence: the actual quality of some artefact with regards to some property.

Desirability: the desired qualities (by some actor/buyer/user) of some artefact with regards to some property.

In computing terms, one might think of the first view as a variable, the second view the value that can be assigned to this variable after evaluating the quality of some artefact, while the third view corresponds to the value that can be assigned to this variable when considering the desires of some actor/buyer[4].

In our approach, quality has to be made specific and precise in order to be able to reason about it. We therefore provide a more formal elaboration of the notion of quality. We will do so in two steps. First we discuss quality as excellence, where we will consider both the quality properties (the variables) and the excellence (the value) of some artefact with regard to a property. Then we move on to the desirability of quality properties.

3.1 Quality as excellence

In this subsection we introduce a model for the properties that artefacts can have. In the formalisation that follows we will use the following notation:

$\mathcal{AF}$	Artefact	$\mathcal{PT}$	Property type
$\mathcal{RO}$	Role type	$\mathcal{PD}$	property domain
$\mathcal{FL}$	Fulfillment	$\mathcal{VL}$	Value

Let $\mathcal{AF}$ be the set of all artefacts that may have certain qualities or properties, and let $\mathcal{RO}$ be the set of all roles that these artefacts can fulfill. The combination of an artefact and a role is dubbed a *fulfillment*. So a fulfillment denotes an artefact in a role. Fulfillments are captured by $\mathcal{FL}$. The artefacts and roles that participate in a fulfillment can be found using the functions $\mathsf{Artefact} : \mathcal{FL} \rightarrow \mathcal{AF}$ and $\mathsf{Role} : \mathcal{FL} \rightarrow \mathcal{RO}$ respectively. Since a fulfillment denotes an artefact in a

---

[4] We would like to thank one of the anonymous referees for suggesting this analogy.

role we know that an artefact and a role combination uniquely determines a fulfillment:

Axiom 1 (Unique fulfillment)

$$\mathsf{Artefact}(f_1) = \mathsf{Artefact}(f_2) \ \wedge \ \mathsf{Role}(f_1) = \mathsf{Role}(f_2) \quad \implies \quad f_1 = f_2$$

For a fulfillment $f \in \mathcal{FL}$ we introduce the following notation:

$$\langle a, r \rangle \ \triangleq \ f \quad \text{such that} \quad \mathsf{Artefact}(f) = a \ \wedge \ \mathsf{Role}(f) = r$$

The following example illustrates this. Let *Mug* denoted by a be an artefact that can play two roles. It either plays the role of type *some artefact to drink from* denoted by r_1, or the role of type *art object* denoted by r_2. Both $f_1 = \langle a, r_1 \rangle$ and $f_2 = \langle a, r_2 \rangle$ are fulfillments such that:

$$\begin{array}{ll} \mathsf{Artefact}(f_1) = a & \mathsf{Role}(f_1) = r_1 \\ \mathsf{Artefact}(f_2) = a & \mathsf{Role}(f_2) = r_2 \end{array}$$

Role types can have properties, the value of which are expressed in a property domain. For example, the role type *art object* can have the property type *color* and the values that this property can take are expressed in the domain *RGB-colors*. Let $\mathcal{PT}$ be the set of property types and $\mathcal{PD}$ be the set of property domains. The properties that can be played by a certain role type are given by the function $\mathsf{Props} : \mathcal{RO} \to \wp(\mathcal{PT})$ and the domain in which values of a property can be expressed is given by the function $\mathsf{PrDom} : \mathcal{PT} \to \mathcal{PD}$.

We continue the above mentioned example. Role type *art object* denoted by r_2 can have the property type *color* denoted by p which can be expressed in the domain *RGB-colors* denoted by d. As a consequence, we have:

$$\begin{array}{l} \mathsf{Props}(r_2) \ = \{p\} \\ \mathsf{PrDom}(p) = d \end{array}$$

Note that property types and domains are at the typing level. We still need to assign values to entities having a certain property type. The first step to achieve this is to create a link between $\mathcal{PD}$ and the values from this domain. The set $\mathcal{VL}$ consists of sets of values for a certain domain. In other words, an element from $\mathcal{PD}$ is the *name* of a certain domain and an element of $\mathcal{VL}$ consists of its *values*. The functions $\mathsf{Value} : \mathcal{PD} \to \mathcal{VL}$ and $\mathsf{VlDom} : \mathcal{VL} \to \mathcal{PD}$ are used to find the values of a domain or the name of a set of values respectively. For example, the domain *RGB-colors* denoted by d has the values $v = \{\#000000 \ldots \#FFFFFF\}$. More specifically:

$$\begin{array}{l} \mathsf{Value}(d) \ = v \\ \mathsf{VlDom}(v) = d \end{array}$$

The actual value assignment of a fulfillment (the level of excellence of some artefact in fulfilling a role) having a certain property is given by the function

ValAss : $\mathcal{FL} \times \mathcal{PT} \to \mathcal{VL}$. In the example, the fact that mug a as an art object r_2 has the color p with value *red* is expressed as follows:

$$\mathsf{ValAss}(\langle a, r_2 \rangle, p) = \textit{\#FF0000}$$

We have to ensure that the observations on the instance level do not conflict with the typing level. For example, if a fulfillment is said to have a value assignment for a property, then at least one of the roles of this fulfillment must have this property. If f is a fulfillment, p is a property type and v is a value, we have:

Axiom 2 (Conformance)

$$\mathsf{ValAss}(f, p) = v \quad \Longrightarrow \quad p \in \mathsf{Props}(\mathsf{Role}(f)) \ \wedge \ \mathsf{PrDom}(p) = \mathsf{VlDom}(v)$$

In our framework of quality treatment, the axioms of conformance and unique fulfillment express basic properties of wellformed quality models.

3.2 Quality and desirability

To be able to assess the desirable quality of an artefact for a user, the actual desires of this user must be made explicit. The question is how to do this. One of the main problems is to choose a domain in which quality is expressed. As an example, it does not make sense to say that the quality of an artefact is 24. The notion of quality is, in that respect, similar to the notion of value as discussed in [1]. A value is an abstract notion and can be used to compare artefacts.

Quality, in the sense of desirability, depends on the desires of actors such as people. Here a distinction must be made between hard and soft desires with regard to artefacts. These can be compared, to some extent, to functional and non-functional requirements or hard goals and soft goals in requirements engineering. In requirements engineering one often tries to make soft goals hard. See for instance [4]. In our approach, goals and requirements are considered to be soft if a human opinion is needed for the value assignment. Otherwise, it is considered to be hard. In other words, hardness or softness of a requirement depends on the way of measurement. The following are examples of hard goals and soft goals:

hard goals: price below €20, contents of 25 liters, made of stainless steel.

soft goals: cheap, pretty, low, hard, strong.

Quality in the sense of desirability depends on the requirements of an individual. More specifically, these requirements have to do with value assignments. The quality of a fulfillment increases if properties have *the right value*. Putting it differently, value assignments are constrained. Consider the following example of a requirement for a fulfillment:

The price in euros may not exceed the price of a given cup.

In this example, *price* is a property type which is expressed in the domain €'s. Furthermore, *may not exceed the price of that cup* is a constraint involving an assignment.

Observe that all requirements involve a constraint and a property type. However, some also involve a specific value. We model this as follows. Let $\mathcal{RQ}$ be the set of all requirements and let $\mathcal{CS}$ be the set of all constraints. A requirement has a mandatory property type, a mandatory constraint and an optional expression. Expressions can either be values or fulfillments, as illustrated by the above examples. In terms of our model, we have $\mathcal{EX} \triangleq \mathcal{VL} \cup \mathsf{ValAss}$.

Let $\mathsf{Prop} : \mathcal{RQ} \rightarrow \mathcal{PT}$, $\mathsf{Constr} : \mathcal{RQ} \rightarrow \mathcal{CS}$, and $\mathsf{Expr} : \mathcal{RQ} \rightarrowtail \mathcal{EX}$. In our framework we use the following shorthand notations:

$$r_1 = \langle p, c, e \rangle \triangleq \mathsf{Prop}(r_1) = p \wedge \mathsf{Constr}(r_1) = c \wedge \mathsf{Expr}(r_1) = e$$
$$r_2 = \langle p, c \rangle \triangleq \mathsf{Prop}(r_2) = p \wedge \mathsf{Constr}(r_2) = c$$

This allows us to write $\langle price, <, €10 \rangle$ for the constraint *the price may not exceed €10* and $\langle price, \min \rangle$ for the constraint *the price must be as low as possible*. Note that a requirement with respect to a fulfillment is of a certain actor. Let $\mathcal{AC}$ be the set of actors and let $\mathsf{Req} : \mathcal{AC} \times \mathcal{FL} \rightarrow \wp(\mathcal{RQ})$ yield the requirements of an actor with regard to a fulfillment. For example, we then have:

$$\mathsf{Req}(a, f) = \{r_1, r_2\}$$

This expresses that actor a has requirements r_1 and r_2 with regard to fulfillment f. Here we conclude the discussion of our model of quality. The focus of the next section is on operationalization. In order to have a graphical representation of our quality definitions, we present an object-role model in Figure 2. See for example [9] for details on object-role models.

4 Operationalizing quality

In the previous section we have presented our model for quality which unifies two interpretations of quality. In this section we will shift the focus to operationalizing this model in practice. The ambition of this paper is not to come up with a tool that will determine the quality of a web resource for a searcher. We will merely concentrate on determining which aspects play a role in such a process and how these aspects could be tackled.

4.1 Uncertainty in quality assessment

Since a quality assessment system will measure the quality of a resource for a certain actor, the system should be able to deal with uncertainty with regards
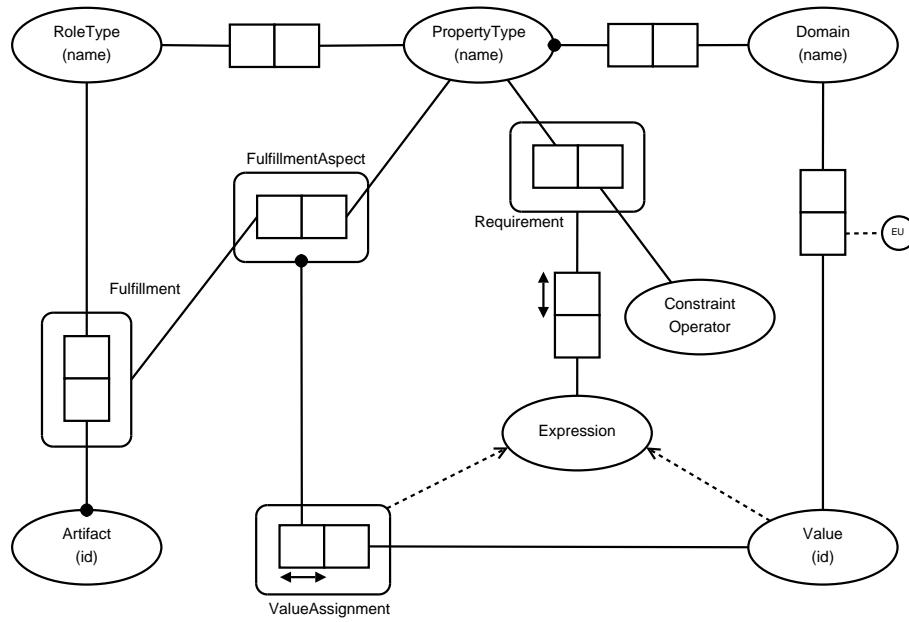
Fig. 2: Object-role model of quality

to the quality. A first kind of uncertainty has to do with the observations and measurements of the system. For example, the fact that a resource has outgoing hyperlinks can be be measured with near 100% certainty. However, the language of a resource is more difficult to measure. This kind of uncertainty is mentioned in the left part of Figure 3.
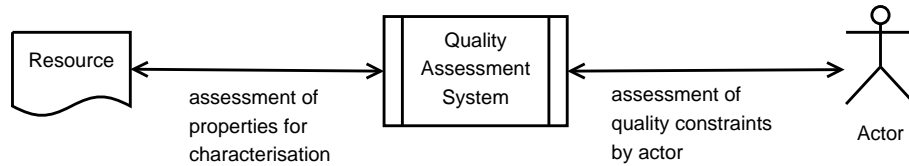

Fig. 3: Uncertainty in quality assessment

In the right part of Figure 3 a second kind of uncertainty is mentioned. This uncertainty deals with the actor for which the quality assessment is made. Consider, for example, the situation in which an actor assesses the quality of resources based on their length, in number of words. Assume that if a text is long then the quality of the resource is considered to be high. Although the number of words

is an objective measure, the adjective *long* is subjective and therefore it will be difficult for the quality assessment system to assert with some level of certainty whether a resource is long or not.

4.2 Linguistic variables

In our quality framework, the concept of a *linguistic variable* is used to describe the fuzzy assessments made by actors [22, 23, 24, 21]. A linguistic variable differs from a numerical variable in that its values are not numbers but words or sentences in some language. For example, the linguistic variable *age* might take *young*, *not young*, *old*, or *not very old* as its values.

More formally, a linguistic variable is defined by a quintuple $\langle \mathcal{X}, T(\mathcal{X}), U, G, M \rangle$ in which $\mathcal{X}$ is the name of the variable. The set $T(\mathcal{X})$ or simply T denotes the termset of $\mathcal{X}$, that is, the set of names of linguistic values being fuzzy variables ranging over U. The rule G is a syntactic rule which usually has the form of a grammar, for generating the names of the values of $\mathcal{X}$. The rule M is a semantic rule for associating with each name X its meaning $M(X)$. The fuzzy restriction is characterized by a membership function $\mu_R : U \rightarrow [0,1]$ which represents the grade of membership with respect to the fuzzy restriction. For example, for a linguistic variable named *age* we could compute that $\mu_{\text{young}}(50) = 0.35$ which expresses how confident we are that an age of 50 years is considered to be young.

In case of resources on the Web, we need a language for expressing qualities of resources as well as domains in U in which these qualities can be expressed. As an example, assume that *importance* of a resource is the only criterion by which the quality is assessed. Using the terminology of Section 3 we then have:

- We are interested in resources which play the role of *webpage*. Let r be such a resource and let $f \in \mathcal{FL}$ with $\mathsf{Artefact}(f) = r$ and $\mathsf{Role}(f) = webpage$.

- We let $\mathcal{X} \in \mathcal{PT}$ be a property type for *importance*, modelled as a linguistic variable. Furthermore we define domain $\mathsf{PrDom}(\mathcal{X}) = PageRank$ and $\mathsf{Value}(PageRank) = \{\text{low}, \text{medium}, \text{high}\}$ which conforms to the termset of the fuzzy variable $\mathcal{X}$. The universe of discourse is the PageRank which can vary from 0 to 10. The membership function is illustrated by Figure 4.

- We define the requirement $R = \langle \mathcal{X}, \text{high} \rangle$ expressing that the importance property of a resource must be high if it is to be judged as having high quality. Now if $a \in \mathcal{AC}$ is the actor for which the quality assessment takes place, we have $\mathsf{Req}(a, f) = \{R\}$.

The membership function in Figure 4 is read as follows. When the quality of a single resource is to be assessed then the actual pagerank can be used to determine whether the importance is high. An actual pagerank can be extracted from engines such as Google. For example, if the pagerank exceeds 9 then the membership function states that we are 100% sure that this resource will have a high importance. For a pagerank of approximately 7 the membership function
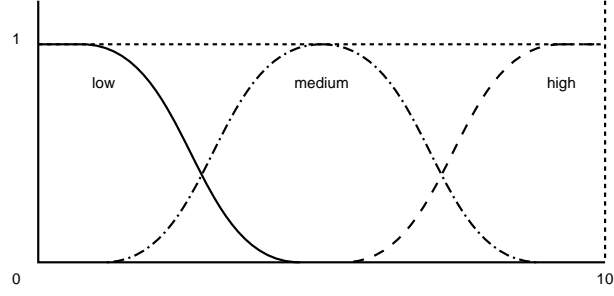
Fig. 4: Membership function for *importance* with PageRank as universe of discourse

states that the degree of membership for *high* and *medium* is approximately the same and equals roughly 0.4. This can be interpreted as being only 40% sure that the quality of the resource is, indeed, high.

4.3   Quality measurement

Actors use requirements and constraints to determine the quality of an artefact. These requirements are often soft in the sense that they can not be measured directly. Some examples are:

- The resource must have a high pagerank.

- The resource must be recent.

In our approach these soft requirements are translated to concrete statements:

- Data resource having attribution (with value "high" AND of type "pagerank")

- Data resource having attribution (with value "recent" AND of type "modification date")

These statements are meaningful under the assumption that *high* and *recent* are fuzzy values which are mapped to their respective hard domains.

Then we define what it means that we measure some property of an artefact to have a value with some degree of certainty. Measurements depend on the situation in which they are done. Measuring the weight of an artefact depends on the location, for instance on the moon or earth. Furthermore, the measuring device is another cause for concern. For example, one thermometer may be less accurate than another. To model this we use a set $\mathcal{SI}$ describing possible relevant situations and $\mathcal{MD}$ describing measuring devices.

Two additional observations are relevant to our discussion here. Firstly, different kinds of measurements can be done:

1. One can attempt to measure the value of some property of an artefact.

2. One can attempt to verify whether the value associated to a property of an artefact satisfies a condition.

So a measurement results in a measured value or in a boolean. Let $\mathcal{MV}$ be the union of all possbible value domains. A measuring device $R \in \mathcal{MD}$ can now be modeled as a function that maps object-situation combinations into values:

$$R = [\mathcal{AF} \times \mathcal{SI}] \rightarrowtail \mathcal{MV}$$

Secondly, we denote a specific measurement with $\mathsf{M}(a, s, d) = v$ where a is the artefact under consideration, s is the present situation, d is the measuring device and v the observed value.

Example 1. Let c be a car. John is driving down the highway somewhere in Europe. Let s denote his situation, that is his current point in space and time. John drives past a police officer who uses a device d which checks the speed of cars. The observation that John is driving at a speed of $125km/h$ is expressed as $\mathsf{M}(c, s, d) = 125km/h$.

4.4 Accuracy of measurements

In this section we consider the accuracy of quality measurements. In this context one must realize that measurements are expressed in a domain and that there are standards for expressing them. For example, speed can be measured in terms of kilometers per hour, weight can be measured in terms of grams, distances in terms of meters and so on. Standards bodies, such as a department of weights and measures, govern these standards. By comparing an actual measurement to a standard measurement one obtains a metric for determining the accuracy of a measurement device. We continue the above example as follows.

Example 2. Let d_s be an approved measuring device for speed, that is it measures exactly according the department of weights and measures. This means that a measurement executed with this device is assumed to be 100% correct. If $\mathsf{M}(c, s, d) = \mathsf{M}(c, s, d_s)$ then we know that John was indeed driving exactly at $125km/h$.

In many cases a small deviation of measurement can be allowed when comparing an actual measurement to a standard measurement. To put it differently, when determining whether an actual measurement is equal to a standard measurement one tests if they are sufficiently equal. We define $\stackrel{\circ}{=}$ to be an operator that determines whether a measurement is sufficiently equal to a standard measurement. In other words, a measurement is accurate or sufficiently equal to a standard measurement if $\mathsf{M}(c, s, d) \stackrel{\circ}{=} \mathsf{M}(c, s, d_s)$.

We relate the above discussion to the uncertainty involved in measurements. This uncertainty is caused by the accuracy of measurement devices and the

many possible situations in which they are used. The following illustrates this. Let d be a measurement device and d_s be a standard measurement device for the same domain. The measurements of device d can be tested against d_s in many but not neccesarily all situations $S \subseteq \mathcal{SI}$. In our framework, the accuracy of d is defined as the average deviation of that device with respect to the situations in which it is tested:

$$\mathsf{Acc}(d) = \frac{\sum_{s \in S} \mathsf{M}(c, s, d) \overset{\circ}{=} \mathsf{M}(c, s, d_s)}{|S|}$$

This accuracy is the basis for defining the measurement uncertainty. That is, if we assert that a property can be measured with a degree of certainty n then we mean that measurements done with this device are correct in $n\%$ of the situations.

4.5 Interpretation of measurements

The uncertainty involved with interpreting measurements is modeled similarly and makes use of linguistic variables. Let $\langle \mathcal{X}, T(\mathcal{X}), U, G, M \rangle$ be a linguistic variable. In the running example $\mathcal{X}$ represents the variable *volume of a mug* with termset $T(\mathcal{X}) = \{\text{big}, \text{medium}, \text{small}\}$. We interpret the membership degree for these linguistic values as the degree of certainty that we have in this specific interpretation of the actual measurement. Let $\mu_t : U \to [0 \dots 1]$ denote the membership degree for the terms in the termset. Consider the following example.

Example 3. Linguistic variable $\mathcal{X}$ denotes volume with termset $\{\text{small}, \text{medium}, \text{big}\}$.

Domain U represents volume in *cc*'s. The following is an example of a linear membership function for the linguistic value *big*:

$$\mu_b(u) = \begin{cases} 0 & u \leq 15 \\ \frac{1}{15}u - 1 & \text{otherwise} \\ 1 & u \geq 30 \end{cases}$$

Now we consider the following question. If the volume of a mug is measured to be $25\,cc$, what are the odds that this mug is considered to be big?

The answer to this question depends on the accuracy of measurements as previously described, but also on the interpretation of the linguistic value *big*. In our approach we interpret the membership degree as certainty of interpretation. This is based on a conversion of the membership degree function to a probability distribution.

5   Example quality assessment system

In this section we will illustrate our quality framework by means of an example quality assessment system. This system is assigned the task to assess the quality of the newsletter of an online news site. The role of this site is informative medium. In terms of our formalism $n \in \mathcal{AF}$ denotes the newsletter and $r \in \mathcal{RO}$ denotes the role played by this site. Furthermore $f = \langle n, r \rangle$ is the fulfillment for this newsletter.

The assessment has to take place for a certain actor $a \in \mathcal{AC}$. Suppose that the actor has three requirements $\mathsf{Req}(f) = \{r_1, r_2, r_3\}$ verbalized as follows:

r_1: Data resource involved in Representation of type "newsletter"
r_2: Data resource of type "Pdf"
r_3: Data resouce having attribution (with value "high" AND of type "importance")

These requirements are embedded in the quality framework as follows:

$r_1 = \langle p_1, c_1, e_1 \rangle$ where p_1 is the property type *representation type*, c_1 is an equality constraint and e_1 is value expression *newsletter*.

$r_2 = \langle p_2, c_2, e_2 \rangle$ where p_2 is the property type *data resource type*, c_2 is an quality constraint and e_2 is value expression *Pdf* which is a data resource type.

$r_3 = \langle p_3, c_3, e_3 \rangle$ where p_3 is the property type *importance*, c_3 is an equality constraint and e_3 value *high*. Note that in this case the system uses a linguistic variable to represent this constraint since *high* is a soft value. The underlying hard domain for importance is chosen to be the pagerank metric.

To be able to make a quality assessment the system uses three measuring devices $d_1, d_2, d_3 \in \mathcal{MD}$, one for each constraint. The three measurements will be done in parallel in one situation $s \in \mathcal{SI}$. Suppose that, based on previous experiences the system knows the following.

d_1 is software tool that is designed with the sole purpose of determining whether a given artefact is a newsletter or not. $\mathsf{Acc}(d_1) = 0.95$ which means that the system is able to correctly judge whether a given artefact is actually a newsletter in 95% of the cases.

d_2 is a tool that checks the data resource type of artefacts. This tool has been trained extensively on all known types and therefore $\mathsf{Acc}(d_2) = 1$.

d_3 is a highly complex tool. It assumes that the PageRank is a good measure for importances of artefacts but knows that this need not always be a 100% correct assumption. Hence suppose $\mathsf{Acc}(d_3) = 0.9$.

The system uses linguistic variables to express the values of constraints. For r_1 and r_2 the membership function is 1 if the condition is met and 0 if it is not. However, for r_3 the situation is a little more complex. The termset for this variable is $\{low, average, high\}$ and the underlying domain $U = [0 \dots 10]$ is the

domain for expressing pagerank. After careful consideration of the user profile of actor a the system chooses the following membership function for linguistic value *high*:

$$\mu_{\text{high}}(u) = \begin{cases} 0 & 0 \leq u \leq 6 \\ \frac{1}{4}u - 1\frac{1}{2} & 6 < u \leq 10 \end{cases}$$

In this example situation s the system makes the following measurements:

$\mathsf{M}(n, s, d_1) = true$ means that the system suggests that s is indeed a newsletter. So membership degree is 1.

$\mathsf{M}(n, s, d_2) = Pdf$ means that the system suggests that s is a *Pdf* file. So the membership degree is 1.

$\mathsf{M}(n, s, d_3) = 9$ means that the observed pagerank for n is 9. The membership degree then is 0.75.

Now the system computes the certainty of the assertion that n is of high quality to actor a as follows:

- $P_{r_1} = 0.95 \times 1 = 0.95$

- $P_{r_2} = 1 \times 1 = 1$

- $P_{r_3} = 0.9 \times 0.75 = 0.675$

Finally the total quality is the multiplication of these three certainties which results in 0.64. The interpretation is that the system is able to assert with 64% certainty that newsletter n is of high quality to actor a.

We are aware of the fact that the example quality assessment system sketched in this section gives a basic illustration of the possibilities of our quality framework. This is sufficient for the purpose of this paper. More complex case studies will be part of future research.

6   Conclusions and future research

The notion of quality plays an important role on the Web, as we rely more and more on information gathered on the Web to perform our day to day tasks. This is why the focus of our project is on aptness-based search rather than topic-based search. Not only topic, but other factors should be taken into account as well when searching the Web. In the current paper we have focused on aptness of Web resources in general, and on the notion of quality in particular.

The paper gives an overview of how the quality notion is used in different fields. Also, we have presented a model which explains what quality is and how the quality of an asset for a certain actor can be measured. This model is sufficiently expressive but still needs more work. The fuzzy-logic approach using linguistic variables provides a straightforward way to deal with quality on the Web. We elaborated our approach in an example quality assessment system.

In future research we aim at more complex case studies. On the one hand, our quality framework needs a more extensive validation. On the other hand, we plan to apply the framework in different domains, such as scientific search, medical information management, geographic applications, and bioinformatics. In the area of technology we see the application of XML-based quality management as a challenge to overcome the heterogeneity on the Web.

References

1. P. van Bommel, B. van Gils, H.A. Proper, M. van Vliet, and Th.P. van der Weide. The information market – its basic concepts and its challenges. In *Web information systems engineering (WISE), New York*, volume 3806 of *Lecture Notes in Computer Science*, pages 577–583. Springer-Verlag, November 2005.
2. F.P. Brooks Jr. No silver bullet: essence and accidents of software engineering. *IEEE Computer*, 20(4):10–19, April 1987.
3. G.B. Davis and M.H. Olson. *Management Information Systems: Conceptual Foundations, Structure and Development*. McGraw–Hill, New York, USA, 1985.
4. P Donzelli and B. Bresciani. Improving requirements engineering by quality modelling – a quality-based requirements engineering framework. *Journal of Research and Practice in Information Technology*, 36(4), November 2004.
5. Michael Gertz, M. Tamer Özsu, Gunter Saake, and Kai-Uwe Sattler. Report on the dagstuhl seminar: data quality on the web. *SIGMOD Rec.*, 33(1):127–132, 2004.
6. T. Gilb. *Principles of software engineering management*. Addison Wesley, Reading, Massachusetts, USA, 1988.
7. B. van Gils, H.A. (Erik) Proper, and P. van Bommel. A conceptual model of information supply. *Data & Knowledge Engineering*, 51:189–222, 2004.
8. B. van Gils, H.A. (Erik) Proper, P. van Bommel, and Th.P. van der Weide. Transformations in information supply. In *Workshop of the 16th Conference on Advanced Information Systems Engineering (CAiSE), Riga*, pages 60–78, June 2004.
9. T.A. Halpin. *Information Modeling and Relational Databases, From Conceptual Analysis to Logical Design*. Morgan Kaufmann, San Mateo, California, USA, 2001.
10. M. Harrison. *Principles of operations management*. Pitman, London, United Kingdom, EU, 1996.
11. V. Lala, A. Arnold, S.G. Sutten, and L. Guan. The impact of relative information quality of e-commerce assurance seals on internet purchasing behavior. *International Journal of Accounting Information Systems*, 3(4):237–253, December 2002.
12. K.C. Laudon and J.P. Laudon. *Management Information Systems, International Edition*. Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1996.
13. Steve McConnell. *Code complete, a practical handbook of software construction*. Microsoft Press, Redmond, Washington, USA, 2 edition, 2004.
14. Ken Orr. Data quality and systems theory. *Commun. ACM*, 41(2):66–71, 1998.
15. G. John van der Pijl. Quality of information and the goals and targets of the organization. In *Computer personnel research conference on reinventing IS, Alexandria, USA*, pages 165–172, 1994.
16. C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, United Kingdom, EU, 1975.
17. G.E Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, New York, USA, 1983.

18. I. Sommerville. *Software Engineering.* Addison Wesley, Reading, Massachusetts, USA, 1989.

19. E. Turban, J. Lee, D. King, and H.M. Chung. *Electronic Commerce, a managerial perspective.* Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1999.

20. C. Vishik and A.B. Whinston. Knowledge sharing, quality, and intermediation. In *Conference on work activities coordination and collaboration, San Francisco*, pages 157–166, 1999.

21. L. Zadeh. From computing with numbers to computing with words - from manipulation of measurements to manipulation of perceptions. *International Journal of Applied Mathematics and Computer Science*, 12:307–324, 2002.

22. L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning – i. *Information Science*, 8:199–249, 1975.

23. L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning – ii. *Information Science*, 8:301–357, 1975.

24. L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning – iii. *Information Science*, 9:301–357, 1975.