

Tailoring Data Quality Models using Social Network Preferences

Ismael Caballero¹, Eugenio Verbo², Manuel Serrano¹, Coral Calero¹, Mario Piattini¹

¹ University of Castilla–La Mancha
Grupo Alarcos - Institute of Information Technologies & Systems
Pº de la Universidad 4, 13071 Ciudad Real (Spain)
{Ismael.Caballero, Manuel.Serrano, Coral.Calero, Mario.Piattini}@uclm.es

² Indra Software Labs
R&D Department of Indra Software Labs
Ronda Toledo s/n, 13004 Ciudad Real (Spain)
emverbo@indra.es

Abstract. To succeed in their tasks, users need to manage data with the most adequate quality levels possible according to specific data quality models. Typically, data quality assessment consists of calculating a synthesizing value by means of a weighted average of values and weights associated with each data quality dimension of the data quality model. We shall study not only the overall perception of the level of importance for the set of users carrying out similar tasks, but also the different issues that can influence the selection of the data quality dimensions for the model. The core contribution of this paper is a framework for representing and managing data quality models using social networks. The framework includes a proposal for a data model for social networks centered on data quality (DQSN), and an extensible set of operators based on soft-computing theories for corresponding operations.

Keywords: data quality, data quality dimensions, quality model, social networks.

1 Introduction

Users need documents containing data with an adequate level of quality to succeed in their tasks. The concept of data quality (typically used as synonymous with information quality [1]) provides users with a set of useful fundamentals for assessing the quality of data contained in relevant documents and hence for computing the overall data quality of the document.

Let us assume that it is possible to define classes of users that access and use documents with the same data quality requirements [2]. In addition, let us generalize the term “user” by employing “stakeholder” instead, to mean any agent (person, system or process) involved in the use of documents. To study the level of data quality of a document, and depending on the stakeholders’ role on data, data quality is assessed with regard to a data quality model composed of a set of data quality

dimensions [3]. It is important to highlight that typically, for the sake of simplicity, data quality models present data quality dimensions as having all the same importance for the assessment. But as Wang and Strong demonstrate in [3], not all of them are indeed equally important. And this is the aim of our research: to study what foundations are appropriate for determining how important each one of data quality dimensions is for an assessment scenario, and how various issues can influence to the level of perceived importance of the dimensions. As we also want to make operative our findings, we propose the technological support for conducting such studies.

To be more precise, our interest is focused on how to manage the “weight” of each data quality dimension in the assessment. Initially, this weight would be provided by each one of the stakeholders participating in the assessment scenario. This weight represents his or her perception of the relative importance of a specific data quality dimension with respect to the remaining ones in the model. What we want to set forth is that, even when planning assessments for diverse scenarios, different data quality models could be required. Anyway, we posit that working-alone or set of stakeholders carrying out similar tasks would take into consideration practically the same data quality dimensions for use with the set of different applications needed for their tasks across varying scenarios. The results of the research could be used as a starting point for new assessment efforts by establishing the corresponding benchmarking.

So the simplest way to get a data quality model, containing different data quality dimensions and their associated weights, would consist of asking to the group of stakeholders to decide firstly about the dimensions, and secondly about the perceived level of importance for each data quality dimensions for the data quality model to be used in the assessment. After this, the results should be synthesized, taking into account the different perceptions. Since representations of preferences are typically subjective, it would be preferable to provide stakeholders with a set of linguistic variables to model this subjectivity, instead of numeric values as it has been done in other researching works related to data quality measurement. The development of the foundations for desired calculus is going to be based on soft-computing and computing with words [4] as it is presented in section 2.

To give sufficient automated computational support to our study, and taking into account the definition of social network provided by [5], we have decided to coin the term **Data Quality Social Network (DQSN)** as a set of stakeholders connected by social relationships who share a data quality model for assessing data quality of documents used when developing a task. We intend to use a DQSN and associated analysis for each assessment scenario in order to manage the set of stakeholders, their perceptions and other future researchable issues, like how to manage the dependence among data quality dimensions.

The main contribution of this paper, therefore, is an extensible framework for managing data quality models, based on the overall perception of the level of importance for the different data quality dimensions identified by the members of a DQSN.

The remainder of the paper is structured as follows: Section 2 shows the main foundations of the proposed framework. In the third section, an example of application is introduced. Finally, the fourth section shows several conclusions and outlines our intended future work.

2 The Framework

This section presents the main components of the framework for establishing a Data Quality Social Network (DQSN). The main aim of a DQSN is to make a data quality model which is based on collaboration. This data quality model will be composed of all those data quality dimensions and their corresponding weights. These in turn correspond to the overall perception of the levels of importance of the data quality dimensions which are of interest for stakeholders. This data quality model is associated with the task at hand for the stakeholders, although it is intended to be shareable and interoperable for many other applications. For instance, it can be used for filtering or ranking documents in order to discriminate those that are not of a high enough quality to be used, or for edition purposes, in order to guide designers to optimize their results.

The framework consists of two main elements:

- A Data Model for representing and managing information about the members of the DQSN and their relationships, so as to support annotations regarding to the perception of the level of importance of data quality dimensions made by the members of the DQSN. The data model observes the concept related to data quality measurement issues. In addition, this Data Model might be easily extensible to satisfy those possible future researchable issues.
- An extensible set of Operators based on soft-computing, for making the corresponding calculations, like those to synthesize the overall level of importance of the individual data quality dimensions of a model.

It is important to realize that the framework presented in this paper does not observe the entire process of measurement of data quality dimensions for each one of the entities identifiable in the documents being assessed. See [6-8] for a broader explanation about measuring data quality under the hypothesis considered in this paper.

2.1 A Data Model for DQSN

The main aim of depicting a Social Network is to enable the option of introducing Social Networking Analysis techniques as a way to further extend the framework. In this framework, the corresponding support for annotations is provided, aimed at collecting the perceptions (or preferences) of the stakeholder about each one of the data quality dimensions used to assess the data quality of documents based on their backgrounds and experiences. This implies gathering and conveniently storing information about the elements that participate in the scenario [9].

The information we consider important to manage is grouped into different classes and relationships of the data model presented. As we seek to enable usage by machines, and interoperability between applications, we decided to implement the data model by means of an OWL ontology, given that Semantic Web and Social Networks models support one another: on one hand, the Semantic Web enables online and explicitly represented social information; on the other hand, social network provides knowledge management in which users “outsource” knowledge and beliefs

via the social network. OWL has been chosen instead of any other ontology language because it is the most complete language for the representation of Web data [10]. Instances of this ontology with gathered data are going to be registered by using RDF, since this is widely accepted and it is the basis for Semantic Web applications [11].

The vocabularies used to implement some of the concepts of our data model are:

- **Dublin Core (DC)** [12]. DC emerged as a small set of descriptors for describing such metadata about documents, like creator or publishers.
- **Friend of a Friend (FOAF)** [13], is the basis for developing the Social Network, since it provides terms for describing personal information [5].
- **Software Measurement Ontology (SMO)** implemented from the software measurement ontology proposed by García et al. in [14]. It contains a Software Measurement Ontology describing the most important concepts related to the measurement of the software artifacts. It is available at <http://alarcos.inf-cr.uclm.es/ontologies/smo>.
- **Data Quality Measurement Ontology (DQMO)** implemented from the Data Quality Measurement Information Model (DQMIM) proposed by [6]. This DQMIM describes the different concepts regarding data quality measurement issues.

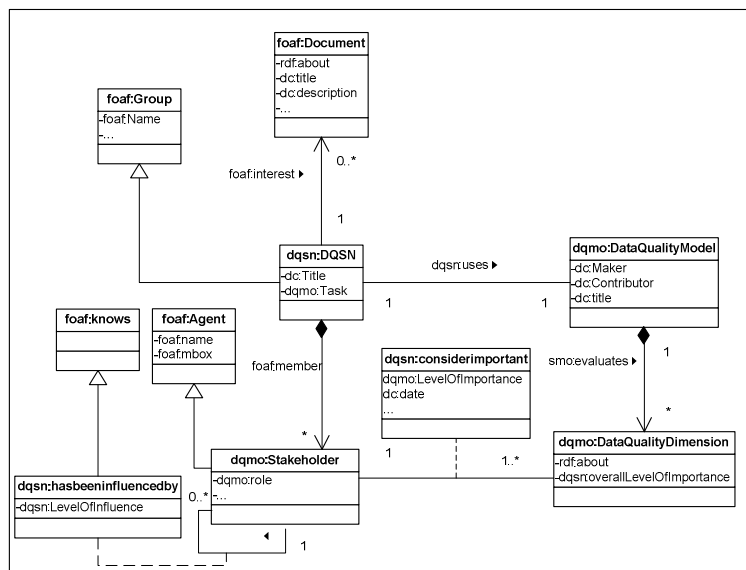


Fig. 1. Data Model for Supporting DQSN

The classes that we have identified are (see Fig. 1):

- **dqsn:DQSN**. Acronym for Data Quality Social Network, which is an aggregation of Agents. This is the class of which a data quality social network is made up and it consists of an aggregation of different dqmo:Stakeholders.
- **dqmo:Stakeholder**. This class is intended to represent a first approach to the stakeholders, who can be the information consumers [1].

- **dqmo:DataQualityDimension.** This class represents the Data Quality Dimensions (see [7, 15]) whose levels of importance are going to be managed. This class has a property named **dqsn:overallLevelOf-Importance**, whose main aim is to store the synthesizing level of importance calculated as the overall opinion of all stakeholders implied in the definition of the data quality model. This is calculated by means of the soft-computing operators described in section 2.2.
- **dqmo:DataQualityModel.** This is used to represent information about the Data Quality Model which is defined for each document under assessment. It contains a set of data quality dimensions, each one with its overall perceived level of importance.
- **foaf:Document.** This represents the document whose data quality is required to be assessed by using the data quality model.

On the other hand, we have identified several properties which allow us to establish relationships between the different classes. Since we needed to complete the semantic of some properties, we have extended them with sub-properties. UML has been used to represent them and some of the following relationships are depicted by means of association classes to keep their meaning. The properties that have been incorporated, represented as relationships or association classes in Fig. 1, are:

- **foaf:interest**, this is used to relate a data quality social network to the set of documents that the members of the network are currently interested in.
- **dqsn:hasbeeninfluencedby**, it is based on the property *foaf:knows*. Its main objective is to represent how a stakeholder has been influenced in his or her perception of the level of importance of data quality dimension by another stakeholder, for different reasons [16].
- **dqsn:considerimportant**, is the property aimed at capturing the level of importance that a *dqmo:stakeholder* gives to a given data quality dimension. As previously said, for the assessment of the data quality of a document, some authors like [8, 17] propose to make a weighted average of normalized vectors. Since it may be very difficult to give a precise value for each weight, we propose using foundations of computing with words [4]. This set limits the possibility of giving values for each level of importance to only a few linguistic labels representing the perception of data quality dimension importance. After assigning the corresponding linguistic label, it is necessary to synthesize the global perception of the data quality dimension importance. This task can be done using the operator Majority guided Linguistic Induced Ordered Weighted Averaging (MIOWA) provided by [18]. MIOWA operators will be discussed later.
- **dqsn:uses** establishes a relationship between the social network and the data quality model used to assess a document. This relationship is necessary because information consumers assess quality within specific business contexts [8].
- **sno:evaluates** is used to connect a specific data quality model with data quality dimensions that are really important for the data quality social network.

2.2 The MIOWA and the Influence-biased MIOWA Operators

In the previous subsection, we have introduced the need for the MIOWA operators to synthesize opinions of the majority of stakeholders (*decision makers*) in order to calculate the overall level of importance of a data quality dimension. In order to best understand the paper, in this subsection we are going to briefly introduce such operators, and then, we are going to present our contribution for synthesizing the opinion of the majority, but taking into account the possibility that the opinion of some stakeholders can be influenced by that of another.

An Ordered Weighted Average (OWA) operator is an aggregation operator taking a collection of argument values a_i and returning a single value. Yager and Filev in [19] define an OWA operator of dimension n as a function $\Phi: \mathfrak{R}^n \rightarrow \mathfrak{R}$, which has an associated weighting vector $W = \{w_1, w_2, \dots, w_n\}$ such that $w_i \in [0, 1]$ with $\sum_i w_i = 1$ for any arguments $a_1, a_2, \dots, a_n \in [0, 1]$: $OWA(a_1, a_2, \dots, a_n) = \sum_i b_i w_i$, with b_i being the i^{th} largest element of the a_j . Let B be a vector that contains the a_i ordered according to a certain criteria such as b_j corresponding to the value a_i before being ordered:

$$\begin{array}{c} [a_1, a_2, \dots, a_i, \dots, a_j, \dots, a_{n-1}, a_n] \\ \searrow \\ [b_1, b_2, \dots, b_i, \dots, b_j, \dots, b_{n-1}, b_n] \end{array}$$

For this example, i is not necessarily greater than j . It is possible to define the function $a\text{-index}(i) = i$ which allows $b_i = a_{a\text{-index}(i)}$. So it is possible to define an OWA operator as expressed in formula (1):

$$OWA(a_1, a_2, \dots, a_n) = W^T B \quad (1)$$

Sometimes, a means to induce the order of the arguments to be aggregated is provided. This means is represented by a function $a_{v\text{-index}(i)} = v_j$, which is known as *order inducing value*. This implies that for any a_i to aggregate, there is another value v_j associated to it. The way to order a_i is done with respect to v_j . Let B_v be a vector where $v\text{-index}(i)$ is the index of the i^{th} largest v_i , it is possible to define the **Induced Ordered Weighted Averaging Operator (IOWA)** [19] as shown in (2):

$$OWA(a_1, a_2, \dots, a_n) = W^T B_v \quad (2)$$

The way to calculate the weights w_i must be such that it allows us to represent the increase of satisfaction in getting S_i instead of S_{i-1} (see Fig. 2). Let Q be: $[0, 1] \rightarrow [0, 1]$ a function so that $Q(0) = 0$ and $Q(1) = 1$, with $Q(x) \geq Q(y) \forall x > y$, with $x, y \in \mathfrak{R}$. Each w_i can be calculated as follows:

$$w_i = Q(i/n) - Q((i-1)/n) \quad (3)$$

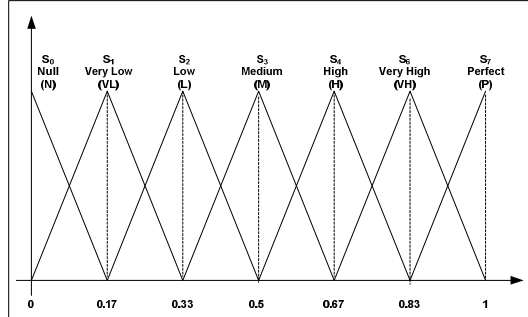


Fig. 2. A set of seven linguistic terms ([20])

The main strength of this weighting vector is that it can have an associated semantic meaning that determines the behavior of the OWA operator, since they have the effect of emphasizing or deemphasizing different components in the aggregation. Therefore, a weighting vector allows better control over the aggregation stage developed in the resolution processes of the Group Decision Making problems [21]. Thus, by introducing corresponding changes in the way of calculating it, it is possible to adapt it to any required situation of computing. It is possible to use the operator to synthesize the opinion of **most** of the decision makers, or some of them, etc. This can be satisfied by defining the function Q as a fuzzy membership function μ . In our case, as we want to get the opinion of the most, the function represented in Fig. 3 can be used.

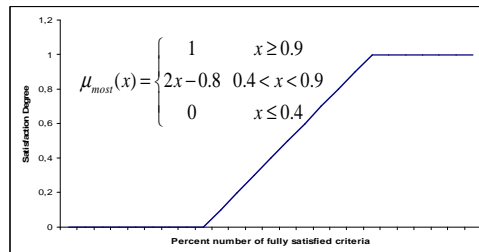


Fig. 3. Definition of the linguistic Quantifier *most* [18]

Pasi and Yager in [18] introduce the Majority IOWA operator by (1) suggesting the idea that the most similar values have close positions in the induced ordering value so they can be aggregated and (2) suggesting a new strategy for constructing the weighting vector so as to model the “majority-based” of the aggregation better. For the first suggestion, they use the support function to compute similarities between pairs of opinion values. Let $\text{Sup}(\mathbf{a}, \mathbf{b})$ be the binary function that expresses the support from \mathbf{b} for \mathbf{a} (see Fig. 4); then the more similar two values are, the more closely they support each other.

$$\text{Sup}(a, b) = \begin{cases} 1 & \text{if } |a - b| < \alpha \\ 0 & \text{otherwise} \end{cases}$$

Fig. 4. Support Function

For the weighting vector, they suggest a procedure for its construction with non-decreasing weights. So, they introduce a modification for the overall support s_i (the sum of the corresponding $\text{sup}(a_i, a_j)$) by making $t_i = s_i + 1$, having $t_i \leq t_j \forall i < j$ and defining w_i (see formula 4). In formula 4, Q is a function μ for linguistic quantifier “most” (see Fig. 3).

$$w_i = Q(t_i/n) / \sum_i Q(t_i/n) \quad (4)$$

Summing up, to use the MIOWA operators, the steps presented in Table 1 can be followed.

Table 1. Steps to apply a MIOWA operator

| |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ol style="list-style-type: none"> 1. Calculate the Order Inducing Values by means of the Support Function (see Fig. 4). 2. Calculate the Weighting Vector W taking into account the quantifiers (see formula 4) 3. Compute the Weighted Average (see formula 2) |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

This calculus must be repeated for each one of the data quality dimensions identified. At the end of all the calculus required, we will have obtained a synthesizing value for the level of importance of each data quality dimension in the data quality model, in the form **LevelOfImportance(NumericalValue)**. With the set of all levels of importance, it is possible to compute the weighted average, together with the values of measures corresponding to each data quality dimension for the assessment of an entity containing data of interest for the stakeholders’ task in hand (e.g. a document of Web Semantic, an XML file...). As a result of the measurement method, one could have a quantitative (or numerical) or a qualitative (like a linguistic label) value. Depending on the nature of the measures, results must be converted so as to get adapted to the suitable operator. Hence, there are two alternatives:

- If the measures obtained are in a numerical format, we can (1) normalize the numerical values associated to the results of the MIOWA operator applied in the calculus described previously (see example in section 3) and then calculate the weighted average, or (2) get the fuzzy values for the results of the measurement and apply a corresponding soft-computing operator based on OWA in order to get a fuzzy value, which must be interpreted according to organizational policies.
- If the obtained measures have qualitative values (e.g. linguistic labels corresponding to fuzzy sets), we can (1) apply the corresponding soft-computing operator based on OWA directly, and then interpret the result according to the aforementioned organizational policies, or (2) defuzzy all the measures corresponding to values of the measures and those obtained for the level of importance of each data quality measure, and then compute a numerical weighted average.

However, any combination of the previous alternatives must be aligned to only one of the four, if we are to get an assessment for the data quality level of the entities being assessed. Nevertheless, we must take into account the nature of data quality measures for the best computational performance.

Now that MIOWA operators have been presented, we are going to introduce the Influence-biased MIOWA, since contextual factors such as personal characteristics, decision tasks, and organizational settings strongly influence the perception of data quality [8]. We have noted that the importance of the influence of decision-makers amongst themselves is not currently observed by the MIOWA. Indeed, MIOWA operators consider decision-makers' opinions as independent. Our idea is that in a social network, one stakeholder (decision-maker) can influence the opinion of another (e.g., group leaders can suggest their vision to the remainder of the group) with the result that the overall opinion of the group is not actually the real opinion of the world, since it could be biased. Indeed, people tend to assume that a thing is of quality simply because another more experienced person has a more positive opinion about the thing being assessed. Since our main goal is to approach the actual level of importance of each data quality dimension, we are conscious that in order to compute the overall perception, all of the opinions must have the same value.

In this first approach, we are going to represent the state of the level of influence of a stakeholder U_i to another U_j with a function $\mathbf{LInfluence}(U_i, U_j)$, which, only for demonstration purposes, can have one of the three values represented in Fig 5. The values provided are hypotheses for demonstration and they must be adapted to any other specific scenario. Indeed, as the levels and their corresponding values do not interfere the researching, they must be calibrated as properly required for each scenario.

$$LInfluence(U_i, U_j) = \begin{cases} 0.2 & U_i \text{ influences } U_j \\ 0.1 & U_i \text{ semi influences } U_j \\ 0 & U_i \text{ not influences } U_j \end{cases}$$

Fig. 5. Function for computing the Level of Influence

The Influenced-biased MIOWA operator is a MIOWA one, in which the similarity between pairs of values is going to be modified by considering the effect of the level of influence. To achieve this goal, we have defined a modified function **support Sup***(**a,b**) (see Fig. 6) based on the previous one shown in Fig. 4 (being a and b the level of importance of two data quality dimensions, and α a threshold value of support):

$$Sup^*(a, b) = \begin{cases} 1 - LInfluence(U_i, U_j) & \text{if } |a - b| < \alpha \\ 0 & \text{otherwise} \end{cases}$$

Fig. 6. Support Function for Influenced-Biased MIOWA Operator

The remainder of the necessary calculus is the same as suggested by Pasi and Yager in [18] for calculating the weighting vectors and the order influence value.

The main purpose of this Influenced-Biased MIOWA is to help DQSN analyzers to find patterns of influence in the perception of data quality in a social network, by comparing results obtained by the MIOWA operator with the ones obtained by applying the Influenced-biased ones.

3 An Example of Application

In [6], authors presented an example that could be extended here to illustrate the use of the proposed framework. In this work, the authors provide a situation in which the owners of a newspaper want to know whether the data quality of the news published on their Web satisfies data quality consumers or not. There, the data quality model used for assessing the soundness of the provided data consisted of two data quality dimensions, namely reliability and completeness. Let's expand now this data quality model by adding a third dimension: timeliness.

To make the example easier to understand, let us suppose that users of the newspaper website (namely *stakeholders*) are interested in performing three tasks: T01, T02 and T03. Let us suppose that the nature of T01 and T02 is so similar that, from the point of view of data quality, stakeholders performing them can be allocated within the same group. This group is the basis for depicting the corresponding DQSN-1. They are going to use a document D, which is intended for assessment.

There are five users forming the DQSN-1. We assume that influence may not be a symmetric relationship, so if U_i influences the data quality perception of U_j , U_j does not necessarily influence the perception of U_i at the same level. We consider only the three possible levels provided in Fig. 5. Table 2 shows in rows the influence level of each user for each workmate in the network (addressed by columns). These values are used to compute the overall perception importance of each data quality dimension by applying the Influence-biased MOWA operator.

Table 2. Influences on users' perception of importance of the data quality dimensions

| | U01 | U02 | U03 | U04 | U05 |
|-----|------|------|------|------|-----|
| U01 | - | Yes | Semi | Yes | Non |
| U02 | Non | - | Non | Yes | Non |
| U03 | Semi | Semi | - | Semi | Non |
| U04 | Semi | Semi | Non | - | Yes |
| U05 | Non | Non | Semi | Not | - |

Let us use a linguistic set $S = \{S_5=VH, S_4=H, S_3=M, S_2=L, S_1=VL\}$ (see Fig. 2), corresponding to the perceptions of level of importance of data quality dimensions for the assessment: VH= "Very High", H= "High", N= "Normal", L= "Low", VL= "Very Low". The perceptions provided by the different users are gathered in Table 3.

Table 3. Perception of the importance level of Data Quality Dimensions for the working example (a_i to aggregate)

| | RELIABILITY | COMPLETENESS | TIMELINESS |
|-----|-------------|--------------|------------|
| U01 | H | VH | L |
| U02 | VH | VL | H |
| U03 | L | M | M |
| U04 | M | M | L |
| U05 | VL | L | L |

Let us now go into the calculations by following the summarized steps provided in Table 1. First, we calculate the corresponding values t_i and t_i^* by means of the support functions (Fig. 4 and Fig. 6) for the cases of non-influenced and influenced-biased described in section 2.2. In Table 4, we show the results for Reliability.

Table 4. Calculating the overall modified Support t_i and Support* t_i^* for Reliability

| | H | VH | L | M | VL | t_i | H | VH | L | M | VL | t_i^* |
|----|---|----|---|---|----|-------|------|------|------|------|------|---------|
| H | 1 | 1 | 0 | 1 | 0 | 3 | 1 | 0.8 | -0.1 | 0.8 | 0 | 2.7 |
| VH | 1 | 1 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | -0.2 | 0 | 1.8 |
| L | 0 | 0 | 1 | 1 | 1 | 3 | -0.1 | -0.1 | 1 | 0.9 | 1 | 2.7 |
| M | 1 | 0 | 1 | 1 | 0 | 3 | 0.9 | -0.1 | 1 | 1 | -0.2 | 2.6 |
| VL | 0 | 0 | 1 | 0 | 1 | 2 | 0 | 0 | 0.9 | 0 | 1 | 1.9 |

The corresponding order inducing values v_i and v_i^* and the consequent induced order are shown in Table 5.

Table 5. Order Inducing values v_i and v_i^* for Reliability

(a) Not Considering Influenced-Biased

| t_i | 2 | 2 | 3 | 3 | 3 |
|-------|-----|-----|-----|-----|-----|
| v_i | U02 | U05 | U01 | U03 | U04 |

(b) Considering Influenced-Biased

| t_i^* | 1.8 | 1.9 | 2.6 | 2.7 | 2.7 |
|---------|-----|-----|-----|-----|-----|
| v_i^* | U02 | U05 | U04 | U01 | U03 |

The induced order for the values a_i to aggregate is represented in the values in Table 6 for both cases: not considering and considering the Influenced-Biased.

Table 6. Calculating the Vectors Bv_i and Bv_i^*

(a) Not Considering Influenced-Biased

| a_i | H | VH | L | M | VL |
|-----------|-----|-----|-----|-----|-----|
| v_i | U02 | U05 | U01 | U03 | U04 |
| $io-a_i$ | VH | VL | H | L | M |
| B_{v_i} | 5 | 1 | 4 | 2 | 3 |

(b) Considering Influenced-Biased

| a_i | H | VH | L | M | VL |
|-------------|-----|-----|-----|-----|-----|
| v_i^* | U02 | U05 | U04 | U01 | U03 |
| $io-a_i^*$ | VH | VL | M | H | L |
| $B_{v_i}^*$ | 5 | 1 | 3 | 4 | 2 |

The second step in Table 1 is used to calculate the Weighting Vector. Table 7 shows how to calculate w_i and w_i^* according to the formula 4, by using a linguistic quantifier *most* Q_2 (Fig. 3). Note that both of them are non-decreasing functions.

Table 7. Calculating the Weighting Vector W by Using a Q2

| t_i | t_i/n | $Q_2(t_i/n)$ | w_i | t_i^* | t_i/n | $Q_2(t_i/n)$ | w_i^* |
|-------|---------|--------------|-------------|---------|---------|--------------|-------------|
| 2 | 0,4 | 0,2 | 0,09 | 1.8 | 0.36 | 0.12 | 0.07 |
| 2 | 0,4 | 0,2 | 0,09 | 1.9 | 0.38 | 0.16 | 0.10 |
| 3 | 0,6 | 0,6 | 0,27 | 2.6 | 0.52 | 0.44 | 0.26 |
| 3 | 0,6 | 0,6 | 0,27 | 2.7 | 0.54 | 0.48 | 0.29 |
| 3 | 0,6 | 0,6 | 0,27 | 2.8 | 0.54 | 0.48 | 0.29 |

Finally, we can now calculate (step 3 in Table 1) the definition of the OWA operator, by applying the formula (2).

$k = \text{round}(\sum_i w_i B_{vi}) = 0.09*5 + 0.09*1 + 0.27*4 + 0.27*2 + 0.27*3 = \text{round}(3) = 3$. So $S_k = S_3 = \text{“Medium”}$

And repeating the calculus for k^* , we obtain a value of $k^* = \text{round}(2,95) = 3$, obtaining $S_{k^*} = \text{“Medium”}$.

By repeating the same calculus for the remainder of the data quality dimensions of the data quality model, the values (and normalized ones for each case) shown in Table 8 are obtained. The data quality model to be used for the tasks T01 and T02 is thereby now fully described.

Table 8. Data Quality Model for DQSN-1

| Data Quality Dimensions | Overall Importance Perceived without considering influence | | Overall Importance Perceived considering influence | |
|-------------------------|------------------------------------------------------------|--------------|----------------------------------------------------|------------------|
| | Reliability | Medium (k=3) | 0.387 | Medium (k*=2.95) |
| Completeness | Low (k=2.41) | 0.313 | Low (k*=2.42) | 0.311 |
| Timeliness | Medium (k=2.33) | 0.301 | Medium (k*=2.40) | 0.309 |

Let us now use the data quality model to assess the data quality level of the data contained in the aforementioned document D. For the sake of simplicity, let us suppose that we have obtained numerical values when measuring the three data quality dimensions on D. Let us also suppose that all of the values are in the same scale, so that it is possible to operate them. The values for the data quality dimensions, ranging between 0 and 100, are shown in Table 9.

Table 9. Measured values obtained for the data quality dimensions

| Reliability | Completeness | Timeliness |
|-------------|--------------|------------|
| 52 | 80 | 90 |

Using a weighted average [17], it is easy to see that without considering influence between members of the DQSN, the resulting value of the assessment is $0.387*52 + 0.313*80 + 0.301*90 = 72.098$, whereas for the other case is $0.380*52 + 0.311*80 + 0.309*90 = 72.450$. This difference (72.098 vs 72.450), although little, demonstrates that influence in the perception of the level of importance between

members of the DQSN affects to the global perception of the level of data quality of a document D, then it deserves to take it into account.

Finally, the results can be expressed by using the DQSN ontology provided in this paper and the final result is the RDF file shown in Fig. 7.

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF [... ]
  <dqsn:DQSN rdf:ID="DQSN-1">
    <foaf:member>
      <dqmo:stakeholder rdf:ID= "U01">
        <dqsn:hasbeeninfluencedby rdf:about= "#U03">
          <dqsn:LevelOfInfluence rdf:about= "values:semiInfluences" />
        </dqsn:hasbeeninfluencedby>
        <dqsn:considerImportant rdf:about= "dqmo:Reliability">
          <dqmo:LevelOfImportance rdf:about= "values:High" />
        </dqsn:considerImportant>
        [... ]
      </dqmo:stakeholder>
      <dqmo:stakeholder rdf:ID= "U02">
        <dqsn:hasbeeninfluencedby rdf:about= "#U01">
          <dqsn:LevelOfInfluence rdf:about="values:Influences" />
        </dqsn:hasbeeninfluencedby>
      </dqmo:stakeholder>
      [... ]
    </foaf:member>
    <foaf:interest>
      <foaf:Document rdf:about= "http://www.itee.uq.edu.au/~dasfaa/MCIS" />
      <foaf:Document rdf:about="http://www.uclm.es"/>
    </foaf:interest>
    <dqsn:uses rdf:ID= "DQModelExampleForMCIS2009">
      <smo:evaluates>
        <dqmo:DataQualityDimension rdf:about= "dqmo:Reliability">
          <dqsn:overallLevelOfImportance rdf:about= "values:Medium" />
        </dqmo:DataQualityDimension>
        [... ]
      </smo:evaluates>
    </dqsn:uses>
  </dqsn:DQSN>
</rdf:RDF>
```

Fig. 7. Extract of a RDF file for the DQSN-1

4 Conclusions and future work

In this paper, a framework for managing data quality models has been presented. This framework is based on the idea that not all data quality dimensions contained in the data quality model, are equally important in the assessment of the data quality of a document. The data quality model is composed of data quality dimensions and an associated weight for each one.

These weights are obtained as a synthesizing value that reflects the perception of the importance level of the stakeholders and other relevant issues, like the effect of the influence on other stakeholders. Since it would be very difficult to assess this level of importance numerically, we have proposed the use of the computing with words theory to model, represent and manage the corresponding levels of perception by means of the MIOWA operators. The more stakeholders participate in the calculus of the overall importance of data quality dimensions, the more representative and reliable the results are. In this sense, a first operator, the Influence-based MIOWA, has been proposed and its existence has been justified by means of an example.

It has also been proposed to manage the necessary information about stakeholders, their relationships and their perception of level of importance of the data quality dimensions by using a data quality social network. This enables to use, on one hand, Social Network Analysis [22] and, on the other hand, Semantic Web Technologies to make the framework operative.

An interesting usage of the DQSN consists of studying the dynamics of data quality perceptions over a period of time, as [23] propose, or even studying the effect on stakeholders by choosing some data quality dimensions instead of others, as studied by [24]. With these conclusions, some rules to infer and predict the future behavior of network members can be arranged in order to avoid data quality problems related to the data quality dimensions identified.

As a main conclusion, we can argue that the work presented is relevant as a promising step in the direction of improving the effectiveness of applications. This may include ways in which it is possible to “filter” or “rank”, according to an established data quality model, the great amount of documents available from different sources like the Internet. Being a DQSN built on semantic web technologies, semantically interoperability becomes possible between different applications.

In the future, we want to refine both the data model for the DQSN and the operators by applying them to several study cases in practice. This will be done by using a tool that automates the calculus and the management of the DQSN. We want to study the different types of influences identified by [8], that condition the election of the weights, in order to extend the operators and manage these issues properly. In addition, we also want to introduce different techniques of Social Network Analysis to the study of the behavior of stakeholders in the assessment of data quality.

Acknowledgments. This research is part of the projects DQNet (TIN2008-04951-E) supported by the Spanish Ministerio of Educación y Ciencia, IVISCUS (PAC08-0024-5991) supported by the JCCM, and IQMF-Tool supported by the University of Castilla-La Mancha.

References

1. Wang, R. Y., A Product Perspective on Total Data Quality Management. *Communications of the ACM* 41(2), 58-65 (1998)
2. Cappelletto, C., Francalanci, C., and Pernici, B. Data quality assessment from the user's perspective. In *International Workshop on Information Quality in Information Systems, (IQIS2004)*, pp. 68-73: ACM Paris, Francia (2004).
3. Wang, R. and Strong, D., Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems; Armonk; Spring 1996* 12(4), 5-33 (1996)
4. Zadeh, L., From Computing with Numbers to Computing with Words-From Manipulation of Measurements to Manipulation of Perceptions. *Fuzzy Control: Theory and Practice* (2000)
5. Ding, L., Zhou, L., Finin, T., and Joshi, A., How the Semantic Web is Being Used: An Analysis of FOAF Documents, in *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04(eds)*. p. 113.3. IEEE Computer Society (2005).

6. Caballero, I., Verbo, E.M., Calero, C., and Piattini, M. A Data Quality Measurement Information Model based on ISO/IEC 15939. In 12th International Conference on Information Quality, pp. MIT, Cambridge, MA (2007).
7. Batini, C. and Scannapieco, M., Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications. Berlin: Springer-Verlag Berlin Heidelberg (2006).
8. Even, A. and Shankaranarayanan, G., Utility-driven assessment of data quality. SIGMIS Database 38(2), 75-93 (2007)
9. Caballero, I., Verbo, E.M., Calero, C., and Piattini, M. DQRDFS:Towards a Semantic Web Enhanced with Data Quality. In Web Information Systems and Technologies, pp. 178-183 Funchal, Madeira, Portugal (2008).
10. Cappiello, C., Francalanci, C., Pernici, B., and Martini, F. Representation and Certification of Data Quality on the Web. In 9th International Conference on Information Quality, pp. 402-417 MIT, Cambridge, MA (2004).
11. Lassila, O. and Hendler, J., Embracing "Web 3.0". IEEE Internet Computing 11(3), 90-93 (2007)
12. DCMI, Dublin Core Metadata Element Set, Version 1.1.
[\(http://dublincore.org/documents/dces/#DCTERMS\)](http://dublincore.org/documents/dces/#DCTERMS).(2008)
13. Brickley, D. and Miller, L., FOAF Vocabulary Specification 0.91. Available at [\(http://xmlns.com/foaf/spec\)](http://xmlns.com/foaf/spec).(2007)
14. García, F., Bertoa, M.F., Calero, C., Vallecillo, A., Ruiz, F., and Genero, M., Towards a consistent terminology for software measurement. Information and Software Technology 48(8), 631-644 (2006)
15. Strong, D.M., Lee, Y.W., and Wang, R.Y., Data Quality in Context. Communications of the ACM 40(5), 103-110 (1997)
16. Zhang, D., Lowry, P.B., Zhou, L., and Fu, X., The impact of Individualism-Collectivism, Social Presence, and Group Diversity on Group Decision Making under majority influence. Journal on Management Information System 23(4), 53-80 (2007)
17. Pipino, L., Lee, Y., and Wang, R., Data Quality Assessment. Communications of the ACM 45(4), 211-218 (2002)
18. Pasi, G. and Yager, R., Modeling the concept of majority opinion in group decision making. Information Sciences 176(4), 390-414 (2006)
19. Yager, R. and Filev, D., Induced ordered weighted averaging operators. Systems, Man and Cybernetics, Part B, IEEE Transactions on 29(2), 141-150 (1999)
20. Herrera-Viedma, E., Herrera, F., Martínez, L., Herrera, J.C., and López, A.G., Incorporating filtering techniques in a fuzzy linguistic multi-agent model for information gathering on the web. Fuzzy Sets and Systems. Web Mining Using Soft Computing 148(1), 61-83 (2004)
21. Chiclana, F., Herrera-Viedma, E., Herrera, F., and Alonso, S., Some induced ordered weighted averaging operators and their use for solving group decision-making problems based on fuzzy preference relations. European Journal of Operational Research 182(1), 383-399 (2007)
22. Jamali, M. and Abolhassani, H., Different Aspects of Social Network Analysis, in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence(eds). p. 66-72. IEEE Computer Society (2006).
23. Stivilia, B. A Model for Information Quality Change. In 12th International Conference on Information Quality, pp. 39-49 MIT, Cambridge, USA (2007).
24. DeAmicis, F., Barone, D., and Batini, C. An Analytical Framework to analyze Dependencies among data Quality Dimensions. In ICIQ'06, pp. MIT, Cambridge, MA, USA (2006).