On the Need for Explicit Confidence Assessments of Flexible Query Answers

Guy De Tré(✉), Robin De Mol, and Antoon Bronselaer

Department of Telecommunications and Information Processing,
Ghent University, St.-Pietersnieuwstraat 41, 9000 Ghent, Belgium
{Guy.DeTre,Robin.DeMol,Antoon.Bronselaer}@UGent.be

Abstract. Flexible query answering systems aim to exploit data collections in a richer way than traditional systems can do. In approaches where flexible criteria are used to reflect user preferences, expressing query satisfaction becomes a matter of degree. Nowadays, it becomes more and more common that data originating from different sources and different data providers are involved in the processing of a single query. Also, data sets can be very large such that not all data within a database or data store can be trusted to the same extent and consequently the results in a query answer can neither be trusted to the same extent. For this reason, data quality assessment becomes an important aspect of query processing. In this paper we discuss the need for explicit data quality assessments of query results. Indeed, To correctly inform users, it is in our opinion essential to communicate not only the satisfaction degrees in a query answer, but also the confidence about these satisfaction degrees as can be derived from data quality assessment. As illustration, we propose a hierarchical approach for query processing and data quality assessment, supporting the computation of as well a satisfaction degree, as its associated confidence degree for each element of the query result. Providing confidence information adds an extra dimension to query processing and leads to more soundly query answers.

Keywords: Fuzzy criterion evaluation · Big data · Data quality handling

1 Introduction

With ever increasing data volumes, database systems face new challenges. An important characteristics of 'Big' data is veracity. Veracity refers to the trust one has in the data that are being used. Our aim with this paper is to contribute to the development of novel techniques for the proper handling of veracity problems. More specifically, we depart from the fact that not all data are of the same quality in large data collections. This is especially the case if data result from data integration, are provided by (volunteered) users, are collected from social media, do not serve the same purposes, or have a different precision.

H. Christiansen et al. (Eds.): FQAS 2017, LNAI 10333, pp. 51–58, 2017.
DOI: 10.1007/978-3-319-59692-1_5

As an illustrative example consider a database with geological data, describing sediment samples and built to establish the substrate composition of the seabed for the purpose of sustainable resource management. Such a database has been built in the project Transnational and Integrated Long-term Marine Exploitation Strategies (TILES) [7]. Seabed samples are taken by various parties for different purposes. For example, construction companies collect samples for stability studies, the government collects samples for the purpose environmental monitoring, while extracting companies might collect samples for resource-quality assessment. Each party can voluntary share data with the others. Different sample data are of different quality. The varying confidence in sample quality propagates to a varying confidence in query results: query satisfaction degrees computed during query processing can neither be trusted to the same extent.

In this paper we describe how confidence in computed satisfaction degrees can be estimated and properly handled. The proposed solution consists of a novel technique that assesses data quality and computes an additional confidence degree for each computed satisfaction degree. In this way, users are provided with extra information needed to end up with best solutions. The data quality of each data item used in query evaluation is characterized by a number of elementary aspects. For example, elementary quality aspects of sample descriptions include the sampling method and sampling date. These elementary data quality aspects are evaluated and their evaluation results are aggregated to an overall confidence degree. This aggregation takes into account how the query results are computed, such that an overall confidence degree reflects the confidence in the query result.

The paper is organized as follows. In Sect. 2 we give some preliminaries on relational databases and ‘fuzzy’ querying of regular databases. In Sect. 3 we respectively deal with the specification of elementary aspects of data quality assessment, the evaluation of elementary quality aspects and the aggregation of quality aspects. An illustrative example is presented in Sect. 4. Finally, in Sect. 5 we provide some conclusions of this work.

2 Preliminaries

In this paper, conventional relational databases are considered. A relational database consists of a collection of relations comprising of attributes (columns) and tuples (rows) [1]. Each relation R can be represented by a table and is defined by a relation schema

$$R(A_1 : T_1, \ldots, A_n : T_n)$$

where the $A_i : T_i$'s are the attributes of R, each consisting of a name A_i and an associated data type T_i. This data type, among others, determines the domain dom_{T_i} consisting of the allowed values for the attribute. Each tuple

$$t_i(A_1 : v_1, \ldots, A_n : v_n)$$

with $v_i \in dom_{T_i}$, $1 \leq i \leq n$ represents a particular entity of the (real) world modelled by the given relation.

Relational database systems support the SQL query language, which among others, offers users facilities to formulate Boolean selection criteria that express what they are looking for. However, adequately translating the user's needs and preferences into a representative Boolean expression is often considered to be too restrictive because Boolean conditions either evaluate to true or false and do not allow for any flexibility regarding partial criterion satisfaction. Soft computing techniques help developing fuzzy approaches for flexible querying that solve these limitations [5]. An overview of basic works can be found in [8].

The essence of 'fuzzy' querying techniques is that they allow to express user preferences with respect to query conditions using linguistic terms which are modelled by fuzzy sets. The basic kind of preferences considered are those which are expressed *inside* an elementary query condition that is defined on a single attribute $A : T$. Hereby, fuzzy sets are used to express in a gradual way that some values of the domain dom_T are more desirable to the user than others. During query processing, basically all relevant database tuples t are evaluated to determine whether they satisfy the user's preferences (to a certain extent) or not. Hereby, each elementary query criterion c_i, $i = 1, \ldots, m$ of the query is evaluated, resulting in an elementary satisfaction degree $\gamma_{c_i}(t)$ which is usually modelled by a real number of the unit interval $[0, 1]$ (where $\gamma_{c_i}(t) = 1$ represents that the tuple t fully satisfies the criterion and $\gamma_{c_i}(t) = 0$ denotes no satisfaction).

Next, the elementary satisfaction degrees are aggregated to compute the overall satisfaction degree $\gamma(t)$ of the tuple. In its simplest form, the aggregation of satisfaction degrees is determined by the fuzzy logical connectives conjunction, disjunction and negation which are respectively defined as follows:

$$\gamma_{c_1 \wedge c_2}(t) = i(\gamma_{c_1}(t), \gamma_{c_2}(t)) \tag{1}$$

$$\gamma_{c_1 \vee c_2}(t) = u(\gamma_{c_1}(t), \gamma_{c_2}(t)) \tag{2}$$

$$\gamma_{\neg c}(t) = 1 - \gamma_c(t) \tag{3}$$

where i and u resp. denote a t-norm and its corresponding t-conorm.

In a more complex approach, users are allowed to express their preferences related to the relative importance of the elementary conditions in a query, hereby indicating that the satisfaction of some query conditions is more desirable than the satisfaction of others. Such preferences are usually denoted by associating a relative weight w_i ($\in [0, 1]$) to each elementary criterion c_i, $i = 1, \ldots, m$ of the query.

The impact of a weight can be computed by first matching the condition as if there is no weight and then second modifying the resulting matching degree in accordance with the weight. A modification function that strengthens the match of more important conditions and weakens the match of less important conditions is used for this purpose. As described in [5], some of the most practical interpretations of weights can be formalised in a universal scheme. Namely, let us assume that query condition c is a conjunction of weighted elementary query conditions c_i (for a disjunction a similar scheme has been offered). Then the matching degree $\gamma_{c_i^*}(t)$ of an elementary condition c_i with associated implicative importance weight w_i is computed by

$$\gamma_{c_i^*}(t) = (w_i \Rightarrow \gamma_{c_i}(t)) \tag{4}$$

where ⇒ denotes a fuzzy implication connective. The overall matching degree of the whole query composed of the conjunction of conditions c_i is calculated using a standard t-norm operator. Other uses and interpretations of weights have been presented [8].

3 Data Quality Handling

The illustrative example in the introduction and many other applications reveal that there is a need for facilities to properly handle data quality in databases.

Most research on data quality assessment has been done in the area of Semantic Web as trust management is an important part of its architecture [6]. Pioneering work in the area of data warehouses has been presented in [3,4]. Herewith, data quality is assessed by means of a linguistic scale and evidence theory is used to estimate the overall reliability of the used data. The approach handles conflicting information by using a merging strategy, which is based on maximal coherent subsets (MCS). Overall reliability scores can be used to order the data and MCS gives insight on how an overall reliability score has been obtained. Query answers can also be enriched with reliability scores, which provides the users with extra information.

Data quality assessment has also been addressed in conventional relational databases [2]. Basically, a database is enriched with quality relations which contain data for data quality assessment and regular relations are extended with foreign key attributes to refer to related data quality assessment data. Selection criteria on quality relations can be included in a query to put extra constraints on data quality characteristics. As a consequence, data quality evaluations and evaluations of other user preferences are mixed and no extra information on data quality is provided to the user.

In this paper we propose to extend 'fuzzy' querying on conventional relational databases in such a way that a separate confidence assessment for each tuple in a query result is computed based on the quality assessments of the data that are used to produce the query result. Such an approach is relevant for many applications like TILES where one cannot afford it to discard data that are of lower quality because else there will not be enough data left. Instead of putting extra quality constraints on the data, a separate assessment of the confidence in each result is computed an provided to the user. The user can use this extra information to better interpret the results.

In the remainder of this section, we introduce an approach where satisfaction degrees of 'fuzzy' queries are enriched with confidence degrees. For a given database tuple, these degrees respectively express to what extent the tuple satisfies a 'fuzzy' query and to what extent one can be confident about this satisfaction.

3.1 Data Quality Assessment

In this stage of our research it is assumed that the database schema contains extra attributes and/or relations that are used to denote data quality. We call these attributes elementary data quality attributes. For every conventional attribute in the database, one or more elementary data quality attributes can be provided. For example consider a relation that contains information about sediment composition and water depth at a location (x, y, z), as given in Table 1. The attributes P_{clay}, P_{c_sand}, P_{m_sand} and P_{f_sand} respectively denote the probability (percentage) for clay, coarse sand, medium sand and fine sand at that location. The attributes *s_method* and *s_year* are elementary quality attributes that respectively denote which sampling method has been used and on which year the sample was taken. The attribute *w_depth* contains information about the water depth at the location, whereas *m_year* is an elementary quality attribute indicating on which year the water depth was measured.

Table 1. An example of a relation 'geology' that contains elementary quality attributes

Location	P_{clay}	P_{c_sand}	P_{m_sand}	P_{f_sand}	*s_method*	*s_year*	*w_depth*	*m_year*
$P1$	0%	50%	50%	0%	m_1	1993	54 m	2016
$P2$	0%	45%	25%	30%	m_1	1980	32 m	2012
$P4$	50%	45%	5%	0%	m_3	2016	41 m	2016

Data for elementary data quality attributes can originate from meta data, e.g. sampling method and sampling date in Table 1, or can be the result of a data audit process.

3.2 Evaluation of Elementary Quality Aspects

Elementary data quality attributes can be queried like conventional attributes. This implies that the 'fuzzy' querying techniques described in the preliminary section can be applied to them.

The innovative aspect proposed in this paper is that we advocate to make an explicit distinction between criteria on conventional attributes and criteria on data quality attributes. The elementary criteria c_i, $i = 1, \ldots, m$ on conventional attributes are evaluated and their evaluation with the data that are related to a given tuple t results in an elementary satisfaction degree $\gamma_{c_i}(t)$. The criteria c_i^Q, $i = 1, \ldots, p$ on elementary data quality attributes are also evaluated and their evaluation with the data that are related to a given tuple t results in an elementary confidence degree $\gamma^Q_{c_i^Q}(t)$.

Elementary satisfaction degrees will be aggregated to an overall satisfaction degree $\gamma(t)$ and elementary confidence degrees will be aggregated independently to an overall confidence degree $\gamma^Q(t)$. Considered together, $\gamma(t)$ and $\gamma^Q(t)$ respectively express to what extent tuple t satisfies the criteria imposed by the query and to what extent one can be confident about this overall satisfaction degree.

By making an explicit distinction between criteria on conventional attributes and criteria on data quality attributes, one can keep control over and be more adequately informed about the quality of the data involved in the query processing and hence also about the confidence each tuple of a query result.

3.3 Aggregation of Quality Aspects

Consider a database tuple t and a 'fuzzy' database query with weighted elementary selection criteria c_i, $i = 1, \ldots, m$. The impact of each weight is modelled by a 'fuzzy' implication connective as presented in Eq. (4). Furthermore, consider that the associated criteria on data quality attributes for the attributes in these elementary selection criteria are c_i^Q, $i = 1, \ldots, p$. For each elementary selection criterion c_i, $i = 1, \ldots, m$ we then have one of the following situations:

1. **One elementary data quality criterion c^Q is specified for c_i.** In this case c^Q is evaluated with t and $\gamma^Q_{c^Q}(t)$ becomes the confidence score for c_i, i.e. $\gamma^Q_{c_i}(t) = \gamma^Q_{c^Q}(t)$.
2. **Multiple elementary data quality criteria c_j^Q, $j = 1, \ldots, k$ are specified for c_i.** In such a case, all c_j^Q, $j = 1, \ldots, k$ are evaluated with t and the resulting confidence scores $\gamma^Q_{c_j^Q}(t)$, $j = 1, \ldots, k$ are aggregated. Up to now we use a simple t-norm operator i as aggregator. Using the minimum t-norm, the confidence score for c_i is $\gamma^Q_{c_i}(t) = \min_j(\gamma^Q_{c_j^Q}(t))$.
3. **No elementary data quality criteria are specified for c_i.** In this case, the user has to assign an ad hoc confidence score to c_i. If the data to evaluate c_i are considered to be adequate enough, an ad hoc confidence score of $\gamma^Q_{c_i}(t) = 1$ can be assigned to c_i. Otherwise another value $0 \leq v < 1$ can be chosen.

The overall confidence degree $\gamma^Q(t)$ for the database query is then computed by aggregating the confidence scores $\gamma^Q_{c_i}(t)$ of its elementary selection criteria c_i, $i = 1, \ldots, m$. As aggregator, a weighted sum can be used as follows.

$$\gamma^Q(t) = \frac{w'_1}{\sum_{i=1}^m w'_i}\gamma^Q_{c_1}(t) + \cdots + \frac{w'_m}{\sum_{i=1}^m w'_i}\gamma^Q_{c_m}(t) \tag{5}$$

where $w'_i = d_{max}(t) - |\gamma(t) - \gamma_{c_i}(t)|$ and $d_{max}(t) = \max_i(\gamma_{c_i}(t)) - \min_i(\gamma_{c_i}(t))$.

With Eq. (5) it is reflected that the impact of a satisfaction degree $\gamma_{c_i}(t)$ on the computation of the overall satisfaction degree $\gamma(t)$ determines the impact of its associated confidence degree $\gamma^Q_{c_i}(t)$ on the computation of the overall confidence degree $\gamma^Q(t)$. This impact is estimated by the difference $|\gamma(t) - \gamma_{c_i}(t)|$. The smaller this difference, the larger the impact. Hence a weight $w'_i = d_{max}(t) - |\gamma(t) - \gamma_{c_i}(t)|$ can be considered for each $\gamma^Q_{c_i}(t)$, where $d_{max}(t)$ is the largest possible difference between an input and the outcome of the aggregation of the satisfaction degrees. The property of internality, which states that the output of an aggregator is bound by the minimum and maximum of its inputs, holds for standard aggregation in 'fuzzy' weighted querying that is based on the operators given in Eqs. (1)–(4). Hence, $d_{max} = \max_i(\gamma_{c_i}(t)) - \min_i(\gamma_{c_i}(t))$. A weighted average aggregator is used to normalize the results.

4 Illustrative Example

Consider the 'fuzzy' query: 'Find locations with a *reasonable* probability for coarse sand which are at a *workable* water depth for ships of type A', imposed on the relation presented in Table 1. Assume that we have 'fuzzy' criteria for the probability of coarse sand (i.e., a fuzzy set with membership function $c_1 = \mu_{reasonable}$) and water depth (i.e., a fuzzy set with membership function $c_2 = \mu_{workable}$). Moreover we have elementary quality criteria for the sampling method, sampling date and depth measurement date. These criteria are respectively defined by fuzzy sets with membership functions $c_1^Q = \mu_{trusted\_method}$, $c_2^Q = \mu_{recent\_sampling\_year}$ and $c_3^Q = \mu_{recent\_measurement\_year}$.

Assume that evaluating these criteria with the data in Table 1 yields the elementary satisfaction degrees and confidence degrees presented in Table 2.

Table 2. Evaluation of elementary query criteria and elementary data quality criteria

Location	$\gamma(c_1)(t)$	$\gamma(c_2)(t)$	$\gamma^Q(c_1^Q)(t)$	$\gamma^Q(c_2^Q)(t)$	$\gamma^Q(c_3^Q)(t)$
$P1$	$c_1(50\%) = 0.8$	$c_2(54\,\text{m}) = 0.5$	$c_1^Q(m_1) = 0.5$	$c_2^Q(1993) = 0.5$	$c_3^Q(2016) = 1$
$P2$	$c_1(45\%) = 0.6$	$c_2(32\,\text{m}) = 0.9$	$c_1^Q(m_1) = 0.5$	$c_2^Q(1980) = 0.2$	$c_3^Q(2012) = 0.6$
$P4$	$c_1(45\%) = 0.6$	$c_2(41\,\text{m}) = 0.7$	$c_1^Q(m_3) = 1$	$c_2^Q(2016) = 1$	$c_3^Q(2016) = 1$

Aggregation yields the query results presented in Table 3. The interpretation of these results is as follows. Location $P1$ satisfies the query to an extent 0.5, whereas locations $P2$ and $P4$ satisfy it to an extent 0.6. The satisfaction degree for location $P1$ is due to the criterion on the water depth, for which the confidence in data quality is 1, hence the full confidence in the result $P1$. The satisfaction degrees for locations $P2$ and $P4$ are both due to the criterion on coarse sand. The confidence in the data for this criterion is 0.2 for $P2$ and 1 for $P4$, hence the confidence of 0.2 and 1 for the results $P2$ and $P4$.

Table 3. Aggregation of elementary query criteria and elementary data quality criteria

Location	$\gamma(t) = \min(\gamma(c_1)(t), \gamma(c_2)(t))$	$\gamma^Q(t)$
$P1$	0.5	1
$P2$	0.6	0.2
$P4$	0.6	1

Both locations $P2$ and $P4$ equally satisfy the query, but lower confidence in the query results makes location $P2$ less attractive. Location $P1$ satisfies the query slightly less, but this satisfaction is obtained with data with higher confidence.

5 Conclusions

In this paper, we discussed and advocated the need for explicit data quality assessment in database and information management systems. Giving the users explicit feedback on the confidence in query results is useful, especially in case of very large data sets with varying data quality. The presented research is also relevant in view of studying the veracity problem in 'big' data. A novel, initial technique for data quality assessment in 'fuzzy' database querying has been presented. At the core of this technique is the explicit distinction between query criteria on conventional attributes and criteria on data quality attributes.

More research is definitely required. Among the research topics we identify are: the development of a better data quality assessment framework, the handling of uncertain data, advanced aggregation techniques and the incorporation in a query language like SQL.

References

1. Codd, E.F.: A relational model of data for large shared data banks. Commun. ACM **13**(6), 377–387 (1970)
2. de F. Mendes Sampaio, S., Dong, C., Sampaio, P.: DQ^{2}S - a framework for data quality-aware information management. Expert Syst. Appl. **42**, 8304–8326 (2015)
3. Destercke, S., Buche, P., Charnomordic, B.: Data reliability assessment in a data warehouse opened on the web. In: Christiansen, H., Tré, G., Yazici, A., Zadrozny, S., Andreasen, T., Larsen, H.L. (eds.) FQAS 2011. LNCS, vol. 7022, pp. 174–185. Springer, Heidelberg (2011). doi:10.1007/978-3-642-24764-4_16
4. Destercke, S., Buche, P., Charnomordic, B.: Evaluating data reliability: an evidential answer with application to a web-enabled data warehouse. IEEE Trans. Knowl. Data Eng. **25**(1), 92–105 (2013)
5. Dubois, D., Prade, H.: Using Fuzzy Sets in Flexible Querying: Why and How? Flexible Query Answering Systems. Kluwer Academic Publishers, Dordrecht (1997)
6. Richardson, M., Agrawal, R., Domingos, P.: Trust management for the semantic web. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 351–368. Springer, Heidelberg (2003). doi:10.1007/978-3-540-39718-2_23
7. Van Lancker, V., Francken, F., Kint, L., Terseleer, N., Van den Eynde, D., De Mol, L., De Tré, G., De Mol, R., Missiaen, T., Chademenos, V., Bakker, M., Maljers, D., Stafleu, J., van Heteren, S.: Building a 4D voxel-based decision support system for a sustainable management of marine geological resources. In: Diviacco, P., Leadbetter, A., Glaves, H. (eds.) Oceanographic and Marine Cross-Domain Data Management for Sustainable Development, pp. 224–252. IGI Global, Hershey (2017)
8. Zadrozny, S., Tré, G., Caluwe, R., Kacprzyk, J.: An overview of fuzzy approaches to flexible database querying. In: Handbook of Research on Fuzzy Information Processing in Databases, pp. 34–54. IGI Global, Hershey (2008)