Aggregation of data quality metrics using the Choquet integral

Soumaya Ben Hassine-Guetari
Université de Lyon (ERIC Lyon 2) & A.I.D. company
4, rue Henri Le Sidaner,
78000 Versailles, France.
+33 643 187 344
sbenhassine@aid.fr

Jérôme Darmont
Université de Lyon (ERIC Lyon 2)
5, avenue Pierre Mendès-France,
69676 Bron Cedex, France.
+33 478 774 403
jerome.darmont@univ-lyon2.fr

Jean-Hugues Chauchat
Université de Lyon (ERIC Lyon 2)
5, avenue Pierre Mendès-France,
69676 Bron Cedex, France.
+33 478 772 379
jean-hugues.chauchat@univ-lyon2.fr

ABSTRACT

In the context of multi-source databases, data fusion is a tricky task, and resolving inconsistency problems when merging duplicate information is one of the most intricate issues as it is generally resolved through subjective approaches.

Using data quality dimensions may help sort out such a question impartially. Quality metrics are the objective criteria that justify the preference of a value *v1* over a value *v2*; where *v1* and *v2* are both referring to the same real world entity but issue from different sources. However, this technique is fairly complicated when the *v1* quality criteria are not all better than the *v2* ones; when we have to choose, for instance, between a highly fresh but inconsistent data, and a consistent old one. Hence, we need a global qualifying score to facilitate the comparison.

In this perspective, aggregation of data quality metrics can be the solution for computing a global and objective data quality score.

In this paper, we introduce a solution that uses the Choquet integral as a means of aggregating data quality metrics.

1. INTRODUCTION

In the context of multi-source information systems, having a complete overview of the available information is not always feasible when it has to deal with separate and disparate data sources[4]. One potential solution is information integration into a federated table where a complete yet concise synopsis of all information sources is provided. Many integration systems are thus defined such as Carnot, Rainbow, Multibase, etc. [3]

The integration process in such systems requires especially inconsistencies resolution among duplicates using comparison criteria such as data quality dimensions.

In the mid-nineties, many data quality dimensions were indeed introduced to facilitate decision makers' and database administrators' consistency resolving tasks [17]. Nowadays, their number is continuously increasing making duplicate management intricate and ambiguous. Consequently, numerous studies have investigated methods that combine these criteria and therefore reduce the dimensionality of this assessment problem. For instance, aggregation techniques were explored in order to compute a global evaluation score with respect to a set of quality metrics.

In this context, many techniques were suggested varying from the classical weighted arithmetic mean to fuzzy measures[1].
We may distinguish between:

- additive aggregation functions that are dedicated to summarize commensurable measures, such as weighted sum;
- non additive aggregation functions that look for a representative statement of the underlying set of criteria by computing either a belief function or a utility function, such as minimum or maximum functions, ordered weighted average and the Choquet integral.

Our paper suggests the use of the Choquet integral as a mean of resolving consistency issues among duplicate alternatives with respect to their quality metrics.

The remainder of this paper is organized as follows. In Section 2, we describe our business context. In Section 3, we present multicriteria aggregation techniques and then explain our choice for the use of fuzzy measures and especially the Choquet integral. In Section 4, we experimentally illustrate our approach. We finally conclude this paper and discuss some perspectives in Section 5.

2. BUSINESS CONTEXT

We are working on a business-to-business prospecting tool where targeted prospects are gathered from various data vendor files.
A typical marketing campaign works as the following way:

1. Users submit a prospecting request to a multi-source database, in which they describe the targeted population by attributes (such as activity code, number of employees, etc.).
2. The expert (or broker) gives priority to well-reputed data sources

Such source-based selection is therefore subjective and does not take advantage from complementary information that may be provided by other sources as no data merging task is performed.
Our main purpose in this project focuses on the enhancement of such marketing campaigns, especially on the improvement of the prospect selection process from various data vendor files.
For this sake, data vendor files are integrated into a central database. This process, unless providing a unified view of

[1] The term fuzzy is used here to express degrees of satisfaction from the attainment of goals and from satisfaction of soft constraints.

information sources, increases data replication and data inconsistency rates in the federated database, where numerous duplicates of the same information with variable quality levels are found.
An example of telephone number inconsistencies detected after integrating external files is provided in Table1.

Table1. Telephone number inconsistencies

Contact ID	Data Source	Telephone number
0299	S1	0655555555
0299	S1	0639233923
0299	S2	0101010101
0299	S3	112342345
0299	S4	0176772227

Table1 features five different telephone numbers related to the same contact identifier, issued from four different sources. The business issue consists in identifying the right telephone number in order to perform a phoning marketing campaign. A first solution consists in choosing the information related to well-reputed source, which is straightforward when dealing with S2, S3 or S4, but more complicated if S1 is the best source due to the dissimilar duplicates provided by S1. A second, more objective solution consists in computing comparison criteria, i.e., quality.
We have defined two classes of quality concerns with the underlying dimensions and metrics: source quality and intrinsic data quality. We inspired from both intuitive (expert experience-based) and empirical (user need-based) approaches in order to define the suitable dimensions and metrics for data quality assessment and thus optimization. The experts' standpoint is indeed crucial for the good functioning of the brokering system: the selection of the most accurate, complete and reliable information. The users' point of view helps produce a successful data marketing campaign. For more details on this business purpose, please refer to [1].
Some of the selected dimensions follow.

1. Source quality: reputation, credibility, added value, price, files freshness…
2. Intrinsic data quality: syntactic and semantic accuracy, freshness, consistency, added value, cost (that translates, in addition to price, the usability of data)…

We focus, in a first step, on computing quality metrics related to some of the above dimensions:

- freshness is represented by the number of months separating current date from delivery date (or creation date);
- syntactic accuracy is represented by a Boolean value (0 if false, 1 otherwise);
- semantic accuracy is the probability that the underlying value is equal to the real world one;
- added value is a subjective metric defined by business experts. For instance, mobile phone numbers (which start with "06" or "07" in France) can be used in mobile phoning marketing campaigns. Thus, they are rated with the highest score;
- cost is the cost of data value expressed in Euros (€).

We obtained many atomic data of various semantics (a set of five metrics for each data value) to help business experts set their preference decisions. Table 2 gives an example of expert decision regarding the telephone number inconsistency problem illustrated in Table1.

Table2. Expert decision

Telephone number	Freshness	Syntactic accuracy	Semantic accuracy	Added Value	Cost (€)	**Expert's point of view**
0655555555	6	1	0.9	1	0.002	**OK++**
0639233923	500	1	0.9	1	0.002	**OK+**
0101010101	500	1	0.3	0,1	0.002	**NO**
112342345	5	0	0.7	0,5	0.012	**OK**
0176772227	1	1	0.8	0,5	0.012	**OK+++**

Indeed, from a marketing point of view, a correct and fresh telephone number is considered as the best option (telephone number "*0176772227*") whereas a correct but old telephone number having no added value is considered as a the worst one (telephone number "0101010101"). Also, we notice that telephone number "112342345" is considered as a valid option for the expert despite being syntactically inaccurate. In fact, it is obvious for the broker that the first digit"0" was inadvertently omitted when capturing the data value.

However, performing this task manually for a whole federated database is infeasible by a human decision maker, especially when dealing with millions of records and tens of attributes, each of which being described by five quality dimensions.
We are thus looking therefore for an automatic function that takes various metrics, as well as expert dimension preferences into consideration, in order to select the best data items.

Our purpose is then to use aggregation techniques to compute a global quality score for a data item. We intend to find the best combination function that summarizes a set of quality dimensions in order to make it simpler to objectively compare a group of data items. The aggregation function has to handle:

- the dependency between dimensions. Helfert devides indeed relationships of information quality dimensions into two categories: negative correlated and positive correlated dependencies[12]:
    - Negative correlation refers to the improvement of one information quality dimension that may lead to a decreasing value in another dimension. For example, by introducing new information to improve completeness, the new introduced information may be inconsistent with the existing information. In this way completeness and consistence are negatively correlated.
    - Positive correlation means two information quality dimensions are mutually contributing to a shared set of information quality problems. For example, when timeliness and accuracy are sharing outdated data as their mutual information quality problem, the improvement of timeliness may lead to an increasing value in accuracy. In this way,

timeliness and accuracy are positively correlated;

- the synergy among dimensions such as the interaction between syntactic and semantic accuracy as they complementary;
- the incommensurability of metrics.

Given these constraints, mutual preferential independence among criteria cannot be assumed, hence making the use of classical additive models such as weighted sum inappropriate. We are looking, then, for an additive function that takes into account:

- the interaction phenomena among criteria,
- the intrinsic importance of criteria as well as the importance of each subset of criteria.

Once a global data quality metric is computed, we intend to generalize the aggregation function at the record level, in order to be able to appraise the quality of merged records.

In the following section, we give an overview of the existing aggregation functions.

3. MULTICRITERIA AGGREGATION TECHNIQUES

Data aggregation refers to any process in which information is expressed as a summary of numerical or fuzzy values for purposes such as reporting, analysis, decision-making or even anonymization for information protection.

Aggregation techniques, also called consensus functions, are used to address many problems in databases. They were also used to address projection as they help reduce, by definition, the dimensionality of the underlying vector or record. Another application of data aggregation is the summarization that synthesizes data into reports. For any of the needs above, data aggregation is the basis of the analysis and the core of the decision making process.

It has thus to deal with a crucial task. An ineffective aggregation function indeed implies incongruous analysis, and, therefore, drastic decisions.

3.1 Additive vs. non-additive methods

Many methods from the literature define consensus functions. We distinguish between additive methods that summarize commensurable measures through a continuous crescent function such as weighted and simple means; and non-additive subjective and objective functions such as ordered weighted average (OWA) and fuzzy integrals.

Despite their simplicity, additive functions entail restrictions on the nature of aggregated measures. They indeed suppose there are neither conflict phenomena nor any synergy among indicators. Thus, they are independent, and consequently allow compensation among measures. In this context, Gustave Choquet proposed to substitute a monotone set function called capacity or fuzzy measure, to the weight vector involved in the classical additive models [5].

Michel Grabisch approves this purpose, declaring that “a natural extension of the weighted arithmetic mean in such a context is the Choquet integral with respect to the defined capacity” [7].

3.2 The Choquet integral

Gustave Choquet, a French statistician, was a pioneer in the theory of non-additive set functions with his theory of capacities [14]. He proposes to substitute a monotone set function called capacity or fuzzy measure, to the weight vector involved in the calculation of weighted sums.

According to Marichal, the Choquet integral may be viewed as an *n*-ary aggregation operator where we can adopt a connective-like notation instead of the usual integral form, the integrand being assimilated to a set of *n* values $x_1, \dots, x_n$ from $\mathbb{R}$. We state then the following definition [13]:

Let $v \epsilon F_N$. The Choquet integral of $x: N \to \mathbb{R}$ with respect to *v* is defined by:

$$C_v(x) = \sum_{i=1}^{n} x_{(i)}[v(A_{(i)}) - v(A_{(i+1)})]$$

where (.) indicates a permutation on N such that $x_{(1)} \leq \cdots \leq x_{(n)}$.and $A_{(i)} = \{(i), \dots, (n)\} and\ A_{(n+1)} = \emptyset$.

For instance, if $x_3 \leq x_1 \leq x_2$,

$$C_v(x_1, x_2, x_3) = x_3[v(3,1,2) - v(1,2)] + x_1[v(1,2) - v(2)] + x_2 v(2)$$

Thus the discrete Choquet integral is a linear expression up to a reordering of the elements. The use of the Choquet integral in a multi-attribute aggregation process then requires the prior identification of a capacity. Capacities, also called fuzzy measures [7,8,9,10], describe criterion importance and can be regarded as generalizations of weighting vectors involved in the calculation of weighted sum [7].

Given a set $N = \{1, \dots, n\}$ of criteria, a capacity $v: 2^N \to [0,1]$ represents the overall score of binary alternatives $(1_A, 0_{A^c})$, $A \subseteq N$, where the notation stands for an alternative having a score of 1 on criteria in A, and 0 else[8].

As stated by Sugeno in his Ph.D. thesis [16], capacity function satisfies the following axioms:

1. it is increasing,
2. *continuous on the right*
3. and *strongly subadditive*[2] when it handles disjoint subsets.

Such methods are least-square-based approaches, maximum split approaches, minimum variance and minimum distance approaches, and less constrained approach [7].

Once the capacity function chosen, another crucial step in the Choquet integral application consists in determining a utility function. The utility function is used to model expert (or decision maker) preferences. It is generally determined by means of an interactive and incremental process requiring from the expert that he expresses his preferences over a small subset of selected objects. It is also important to notice that utility measures and not

[2] In mathematics, subadditivity is a property of a function that states, roughly, that evaluating the function for the sum of two elements of the domain always returns something lesser or equal to the sum of the function's values for each element. There are numerous examples of subadditive functions in various areas of mathematics, particularly norms and square roots. Additive functions are special cases of subadditive functions.

data values are concerned by aggregation functions using capacity methods.

A utility function may be related to:

- the partial preorder between objects (or alternatives) such as the telephone number "*0176772227*" is better than the telephone number "*0101010101*",
- the partial preorder between criteria such as freshness is more important than cost,
- the quantitative importance of criteria such as weighting vectors.

As partial preorder between objects is equivalent to manual ranking that is unfeasible in the real-world case, we discard that option.
The quantitative importance of criteria is computed through the Shapley index that describes the importance or power of a single criterion into the aggregation problem. It acts as a weight vector in a weighted arithmetic mean.
Nonetheless, the Shapley importance index is not enough to have a good description of criteria behavior, as no relation among criteria is taken into account. An interaction index has, therefore, been defined. We may distinguish the following interactions among pairs of criteria:

- positive interaction or positive synergy between criteria when criteria are complimentary (although the importance of a single criterion for decision is almost zero, the importance of the pair is high);
- negative interaction or negative synergy between criteria when criteria are redundant (their union does not bring any information and the importance of the pair is almost the same as the importance of the criteria considered separately);
- independency between criteria (the importance of the pair is more or less the sum of the individual weights of the criteria).

3.3 Why using the Choquet integral?

As seen in the previous paragraph, the Choquet integral can model both user preferences and the synergy among criteria thanks to the utility and capacity functions. In our case of data quality metric aggregation, many interactions exist between the different quality dimensions and it is of great importance to take it into consideration. For instance, added value and accuracy are both complimentary. In fact, when a data value it provides no added value.

Moreover, the Choquet integral is a non-additive function that does not assume independency between criteria, which corresponds to our context, as data quality dimensions are dependent.

We briefly described in this section the basis of the Choquet integral, its use in the aggregation problems and its advantages comparing to multicriteria aggregation techniques. In the following section, we describe our experiment using the Kappalab package[3] for the GNU R statistical system [11].

[3] Kappalab contains high-level routines for capacity and non-additive integral manipulation on a finite setting.

4. EXPERIMENTATION

Let us consider the case described in Table2 and resumed in the Table2-bis below.

Table2-bis. Expert decision

Alternatives	Freshness **C1**	Syntactic accuracy **C2**	Semantic accuracy **C3**	Added Value **C4**	Cost (€) **C5**	**Expert's point of view**
A	6	1	0.9	1	0.002	**OK++**
B	500	1	0.9	1	0.002	**OK+**
C	500	1	0.3	0,1	0.002	**NO**
D	5	0	0.7	0,5	0.012	**OK**
E	1	1	0.8	0,5	0.012	**OK+++**

From a marketing point of view, correct, fresh and cheap data is indeed the best solution. However, correct, old and costly data is favored to inaccurate, fresh and cheap data. Actually, according to the preference table above, *E* is preferred to *A* which is preferred to *B*, and *B* is preferred to *D* and *C*.

Our aim is to find the finest function that models these user preferences in order to find the best *confidence measure*:

- taking the different interactions existing between criteria into account, alternatives or even constraints;
- summarizing the right quality of the underlying data values and representing the marketing expert's point of view.

Our approach is therefore unsupervised as initial user preferences are not considered as input parameters when computing the targeted model.

4.1 Interaction among criteria

We express in this paragraph the interaction existing among the underlying criteria. These interactions are expressed by human decision makers and basing to Helfert's survey [12]:

- Syntactic accuracy (C2) and semantic accuracy (C3) are complementary as they both describe accuracy. Thus, there is a positive synergy between C2 and C3.
- Accuracy (C2 and C3) and added value (C4) are complementary as incorrect values have no added values. Thus, there is a positive synergy between (C2, C3) and C4.
- Freshness (C1) and semantic accuracy (C3) are complementary.
- Freshness (C1) and cost (C5) are independent. Thus, there is no interaction between C1 and C5.
- Accuracy (C2 and C3) and cost (C5) are independent. Thus, there is no interaction between (C2, C3) and C5.
- Added value (C4) and cost (C5) are independent. Thus, there is no interaction between C4 and C5.
- Added value (C4) and freshness (C1) are independent. Thus, there is no interaction between C4 and C1.

We model these criteria using as utility function the Shapley preorder value describing the constraints relative to the preorder of the **criteria** using the *Shapley.preorder* R-package function [11].

We choose as a capacity function the minimum variance approach that is generally regarded as a maximum entropy approach. This

method leads to a unique solution. In the case of insufficient initial preferences involving a small number of constraints, this unique solution does also not exhibit too specific behaviors characterized, for instance, by very high positive or negative interaction indices or a very uneven Shapley value [7].

A generalization of this approach consists in finding, if it exits, the closest capacity to a capacity defined by the expert and compatible with his initial preferences.

When we first apply our constraints, we obtain the Choquet values depicted in Table 2.

Table3. Using the Choquet integral on row data quality metrics

Alternatives	Fresh-ness **C1**	Syntactic accuracy **C2**	Semantic accuracy **C3**	Added value **C4**	Cost (€) **C5**	Choquet Result
A	6	1	0.9	1	0.002	**1.89038**
B	500	1	0.9	1	0.002	**110.5704**
C	500	1	0.3	0,1	0.002	**110.2884**
D	5	0	0.7	0,5	0.012	**1.33228**
E	1	1	0.8	0,5	0.012	**0.68228**

As we can see, B and C have the highest values as their freshness values are important. We can, indeed, remark that freshness (C1) and cost (C5) follow a decreasing function, as higher is the value, lower is the appreciation; unlike C2, C3 and C4 that follow an increasing function.

This is assimilated to an incommensurability problem. In fact, to use the Choquet integral as an aggregation function, it is necessary that utility functions are commensurable, i.e., given a utility function *u*, and two criteria *i* and *j*, $u_i(x) = u_j(x)$ if and only if, from the expert's point of view, object *x* is satisfied to the same extent on criteria *i* and *j* [7].

To solve this issue, two options are possible:

- either we model the alternative preorders using the Choquet preorder function. This option can be biased if the learning examples are not representative. The Choquet preorder function does indeed not describe the monotony of criteria, but gives the global appraisal of the value;
- or we model C1 and C5 by a positive increasing function (that follows the monotonies of C2, C3 and C4) such as the inverse function.

As the Choquet integral is a generalization of classical additive sums, negative values will also appear at the aggregation results level. That could be a constraint when we will deal with the combination at the record level, when chosen criteria are the values global quality scores.

To avoid this constraint, we choose the second option, where freshness is modeled as follows:

freshness=$\frac{1}{months\_between(current\ date - creation\ date)}$; and cost is modeled as follow: cost=$\frac{1}{cost\ expressed\ in\ cents}$.

The results we obtain are depicted in Table 3.

Table4. Using the Choquet integral on monotonous criteria

Alternatives	Fresh-ness* **C1**	Syntactic accuracy **C2**	Semantic accuracy **C3**	Added value **C4**	Cost* (€) **C5**	Choquet Result
A	0.16	1	0.9	1	0.5	**0.7002**
B	0.002	1	1	1	0.5	**0.68544**
C	0.002	1	0.3	0,1	0.5	**0.38344**
D	0.2	0	0.7	0,5	0.083	**0.28977**
E	1	1	0.8	0,5	0.083	**0.69577**

This model almost suits the point of view of decision maker. An old but correct value is indeed preferred to a fresh incorrect one (B>C). Moreover, a fresh, more correct and expensive value is preferred to an old, inaccurate and cheap one (E>C). However, the Choquet score of alternative C is greater than D's, which is not the expert decision. This means that our Choquet model favors the cost and accuracy criteria over freshness.

5. RELATED WORK

Many studies have focused on data quality assessment methodologies. However, while the majority aims at finding the best data quality alternative, few take interest in computing a global quality score and aggregating quality metrics.

Laure Berti-Equille defines a recommendation strategy based on five aggregation techniques [2]:

- linear affectation: This method is assimilated to weighted sum aggregation where mutual preferential independence among criteria is assumed. This independency assumption has to be considered very carefully as it entails a total compensation between the criteria when aggregating them;
- maximax model: considers the best value among all quality metrics related to the underlying item;
- lexicographic order: extends lexicographic model to the overall criteria;
- elimination based on criteria importance: eliminates items having the worst score on the most important criterion;
- Anderson, Subramanian and Gershon methods: describe evaluation between pairs of data items through the use of concordance, discordance and preference matrices.

Beyond linear affectation, all the proposed techniques focus on comparing items without computing a numeric global score.

Naumann details a set of decision making techniques in the context of a data integration process by comparing four methods [14]:

- SAW (Simple Additive Weighting method) is based on the following steps: scaling quality criteria scores to make them comparable, weighting and summing up the values for each criterion
- TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) is based on scaling values and then computing the Euclidean distance to an ideal source.
- AHP (Analytical Hierarchy Process method) is based on the following steps:

    - development of a goal hierarchy,
    - pairwise comparison of goals,
    - consistency check of the comparisons,
    - aggregation of the comparisons.
- DEA (Data Envelopment Analytics method) determines the efficiency of each source separately by solving a linear program.

Naumann defined five comparison criteria to qualify these decision making methods:

- Interaction: the necessity of the user to state preferences or compare alternatives
- Weighting: setting the different importance of the criteria to the user
- Dominance: the ability of the method to discover the dominating alternative
- Scaling: making the different scores comparable
- Result type: a total ranking of the alternatives or a classification of the alternatives

Naumann's approach is indeed the closest to ours as an aggregated quality score is computed in order to rank alternatives when performing the integrating process; and, according to the comparison criteria above, the Choquet integral is comparable to the SAW, TOPSIS and AHP methods as it:

- requires user interaction to set criteria synergy,
- discovers dominating alternatives,
- requires scores scaling,
- generates a ranking score.

Finally, Davoli suggests the use of FQT4Web (Fuzzy Quality Tree for Web Inspection) as a quantitative inspector-based methodology in the assessment process of a set of cultural web sites. The FQT4Web methodology produces [6]:

- six measures of quality dimensions: basic functionality, advanced functionality, usability, accessibility, efficiency and maintainability and compliance, forming a hierarchical tree;
- an overall quality score for a web site setting aggregation criteria through the OWA fuzzy operator.

This work emphasizes a relevant feature of OWA operators that is their implementation of linguistic quantifiers (such as *many*, *most*, *at least*, *about*…), permitting to express, in a mathematically transparent way, sentences like "if at least some of the values to be aggregated are satisfactory, the aggregation score is satisfactory". Therefore, OWA operators allow managing the existing arbitrariness, highlighting the role of human researcher choices in a transparent way, so that their influence on the final quality judgment can be evidenced.

6. CONCLUSIONS AND PERSPECTIVES

The Choquet integral has first been studied and applied in decision making under uncertainty at the end of the eighties in the works of Schmeidler, and at the beginning of the nineties for multi-criteria decision aid (MCDA). Since, the application fields of the Choquet integral have incredibly grown.

In this paper, we described an approach that uses the Choquet integral for data quality metric aggregation in order to help merge multi-source items. This work is part of a research project that aims to optimize the selection of merged alternatives, in the context of business-to-business applications so as to enhance the return on investment of marketing campaigns.

As a next step, we aim to perform aggregation at record, and then at database levels and apply it to a real-world case.

7. REFERENCES

[1] Ben Hassine-Guetari, S. *Data quality evaluation in an e-business environment: a survey*. In 14th ICIQ international conference, Potsdam, Germany, 2009, pp. 189-201.

[2] Berti-Equille, L. *La qualité des données et leur recommandation: modèle conceptuel, formalisation et application à la veille technologique*. Ph.D. Thesis, Université de Toulon-Var, France, 1999.

[3] Bleiholder J. and Naumann F. *Data Fusion*. ACM Computing Surveys, Vol. 41, No.1, 2008.

[4] Bleiholder J. and Naumann F. *Conflict handling strategies in an integrated information system*. International workshop on Information Integration on the Web (IIWeb). 2006

[5] Choquet, G. *Theory of capacities*. Annales de l'Institut Fourier, 5, 1953, pp. 131–295.

[6] Davoli, P., Mazzoni, F. and Corradini, E. *Quality assessment of cultural websites with fuzzy operators*. Journal of Computer Information Systems, Part 1, Vol. 46, 2004, pp. 44-57.

[7] Grabisch, M., Kojadinovic, I. and Meyer, P. *A review of methods for capacity identification in Choquet integral based multi-attribute utility theory. Applications of the Kappalab R package*, European journal of operational research, no. 2, vol. 168, 2008, pp. 766-785.

[8] Grabisch, M. and Labreuche C. *Capacities on lattices and k-ary capacities*. 3rd International Conference on Fuzzy Logic and Technology (EUSFLAT 2003). 2003. Pp 473 – 490.

[9] Grabisch, M.and Labreuche, C. *A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid*. Annals of operations research no1, vol. 175, 2010, pp. 247-286.

[10] Grabisch, M. and Roubens, M. *Application of the Choquet integral in multicriteria descision making*. Fuzzy measures and integrals, Physica Verlag, 2000, pp. 348-374.

[11] Grabisch M., Kojadinovic, I.and Meyer P., *Package 'kappalab'*,version 04-4, http://cran.r-project.org/web/packages/kappalab/kappalab.pdf

[12] Helfert, M., Foley, O.Ge, M. and Cappiello, C. *Analysing the effect of security on information quality dimensions*. In 17th European conference on information systems (ECIS 2009), Verona, Italy, 2009, pp. 2785-2797.

[13] Marichal J.L. *Aggregation of interacting criteria by means of the discrete Choquet integral*. Aggregation operators: new trends and applications, Physica-Verlag GmbH, pp. 224 – 244.

[14] Naumann F. *Data Fusion and Data Quality*. In the New Techniques & Technologies for Statistics Seminar (NTTS), Sorrento, Italy, 1998.

[15] Pap, E. *Variations of non-additive measures*, Acta Polytechnica Hungarica, Vol. 2, 2005, pp. 5- 13.

[16] Sugeno, M. *Theory of fuzzy integrals and its applications*. Ph.D. thesis, Tokyo Institute of Technology, Tokyo, Japan, 1974.

[17] Wand Y., Wang R. Y. *Anchoring Data Quality Dimensions in Ontological Foundations*, no. 11, Vol. 39, 1996, pp. 86-95.