Introducing the Factor *Importance* to Trust of Sources and Certainty of Data in Knowledge Processing Systems - A new Approach for Incorporation and Processing

Markus Jäger, Josef Küng
Institute for Application Oriented Knowledge Processing (FAW)
Faculty of Engineering and Natural Sciences (TNF)
Johannes Kepler University (JKU)
Linz, Austria
{*markus.jaeger, josef.kueng*}*@jku.at*

***Abstract*—In knowledge processing systems data is gathered from several sources. After some calculating and processing steps are taken in the system, a result is finally computed and may be used for further steps or by other systems. Most of the time the origin and provenance of input data is not verified. Using unverified data can cause inconsistencies in processing and generating output, and could lead to corrupting threats for the system and the environment as a whole.**

We propose an approach where several characterizing values in a given environment – trust of source, certainty of data, and importance (of data) in the current processing step – are used to compute new output characteristics of a knowledge processing system. These values represent the trustworthiness and the certainty of the output in multi-step processing systems based on all used sources and input data. We demonstrate the application of our approach on simple and advanced fictitious scenarios as well as on a real world scenario from the agricultural domain.

***Keywords*-Certainty; Importance; Knowledge; Knowledge Processing Systems; Provenance; Security; Trust;**

I. Introduction

When gathering and processing data in a system, the quality of data, especially from external sources, cannot be always ensured. It is always useful to know how trustworthy the source and how accurate the provided data is. We are currently developing an approach where values of trust (of sources) and certainty (of data) are used and processed by taking into account different importances of various inputs in the current processing step. The approach does not address the issue of how to compute trust or certainty of the source, we assume this step has been addressed before and will be covered in future work.

In our work we concentrate on knowledge processing in general, as it usually requires more complex calculations and processing of the available data and information than conventional data processing. Our aim is to finally face the challenges of knowledge processing.

As the reader will notice further below, our approach is at an early development stage, with many questions still unanswered.

This paper is structured as follows: Section II offers definitions for trust, provenance, security, and risk in the common area, and gives some insights into related work. Section III presents our approach for incorporating trust and certainty into knowledge processing systems – covering several questions concerning this topic and giving a current overview of our research – and contains the topic of importance. We put the approach into practice by defining the scopes of trust, certainty, and importance, and trying to give some options for further calculation to use these values in knowledge processing systems. In section IV we show a short and simple scenario as well as an advanced scenario by applying the approach on fictitious situations. Furthermore, we describe a disease pressure model (DPM) and demonstrate the application of our approach on a real world example in a current project.

Each scenario is evaluated directly after a calculation. Summarizing in section V, we provide a complete view on our developed approach and future development work.

II. Related Work

In this section we provide some insights into important terms which are relevant to our work. These are trust, provenance, security, and risk of data and information. Afterwards we show related research publications which are interesting for our approach.

A. Trust

"In a social context, trust has several connotations. Definitions of trust typically refer to a situation characterized by the following aspects: One party (trustor) is willing to rely on the actions of another party (trustee); the situation is directed to the future. In addition, the trustor (voluntarily or forcedly) abandons control over the actions performed by the trustee. As a consequence, the trustor is uncertain about the outcome of the other's actions; they can only develop and evaluate expectations. The uncertainty involves the risk of failure or harm to the trustor if the trustee will not behave as desired." [2]

The meaning of the term "Trust" always depends on the specific environment and field of research and application. In a recent publication regarding some researches

URI: http://hdl.handle.net/10125/41682
ISBN: 978-0-9981331-0-2

on trust, we state: *"The question of 'How can we trust anything/anybody?' is discussed since the beginning of mankind, but what does this topic mean in context to today's technology age and especially for the information technology?"* [15].

Recently, we also raised the issue of trusting in technology, especially in smart home systems, where everybody's personal security and safety can be touched in a very sensitive way.

"Usually, we have a high trust in man-made technology - from cars to airplanes, from computers and buildings to space shuttles. As long as they work properly, we most of the time dont even think about (not) trusting them. Only in case they stop working in their usual behaviours, the question of trust comes up. The trust in IT systems is becoming even more important, as today people rely on IT more than ever before. Beside the usage of IT in every part of our lives, special treatment has to be done with the Internet in this domain. Everybody is online, (most of the time) every time, and the trust into content from the Internet has to be handled crucial. When we talk about this, there isnt meant the information you retrieve or read on websites, much more the download of files must be taken into account: everybody trusts into a "Download Button" by clicking it, but nobody knows what is really behind this mechanism, when you download a file on your computer. You make yourself highly vulnerable, when downloading content from the Internet to your computer, because you never know, what is really inside a file (just one example: malware)." [15]

The three main types of applicable trust by Rousseau et al. [19] are (1) trusting beliefs, (2) trusting intentions, and (3) trusting behaviours, where these three types are connected to each other: *"1. Trusting beliefs means a secure conviction that the other party has favorable attributes (such as benevolence, integrity, and competence), strong enough to create trusting intentions. 2. Trusting intentions means a secure, committed willingness to depend upon, or to become vulnerable to, the other party in specific ways, strong enough to create trusting behaviors. 3. Trusting behaviors means assured actions that demonstrate that one does in fact depend or rely upon the other party instead of on oneself or on controls. Trusting behavior is the action manifestation of willingness to depend. Each of these generic trust types can be applied to trust in IT. Trusting behavior-IT means that one securely depends or relies on the technology instead of trying to control the technology."*

Another point of view is the similarity of trusting people and trusting technology, especially information technology, where the main difference is within the application of trust in the specific area. *"The major difference between trust in people and trust in IT lies in the applicability of specific trusting beliefs. People and technologies have both similar and different attributes, and those similarities and differences define which trusting beliefs apply. [...] With trust in people, one trusts a morally capable and volitional human; with trust in IT, one trusts a human-created artifact with a limited range of behaviors that lacks both will and moral agency. [...] Because technology lacks moral agency, trust in technology necessarily reflects beliefs about a technologys capability rather than its will or its motives. [...] Trust in information technology has several interesting implications. [...] Trust in technology is built the same way as trust in people."* [16]. We highly recommend reading the paper "Trust in Information Technology" from D. Harrison McKnight [16].

In their work "Not so different after all: a cross-discipline view of trust" [19] the authors state that trust is the willingness to be vulnerable, willingness to rely on confident and positive expectations. *"However, the compositions of trust are comparable across research and theory confusing on parties both inside and outside firms and investigate trust relations from different disciplinary vantage points."* The authors of [19] sum up that *"Trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another."*

Another very interesting publication about trust in information sources was written by Hertzum et al. [12] in "Trust in information sources: seeking information from people, documents, and virtual agents". They compare the concept of trust between people and virtual agents, based on two empirical studies. The testimonials were software engineers and users of e-commerce systems. Some relational aspects concerning trust in the industrial marketing and management sector can be found in "Concerning trust and information" from Denize et al. [11].

B. Provenance

When it comes to trust concerning trusting in data and trusting the sources of data, the term "Data Provenance" must be taken into account. It describes the origin and complete processing history of any kind of data. A good introduction and overview can be found in "Data provenance – the foundation of data quality" [4] and in "Data Provenance: Some Basic Issues" [5]: *"We use the term data provenance to refer to the process of tracing and recording the origins of data and its movement between databases."* and *"It is an issue that is certainly broader than computer science, with legal and ethical aspects."*

Several problems concerning data provenance are covered in "Research Problems in Data Provenance" [7].

Trusting the services used and established in a particular information processing and knowledge management (IPKM) system is highly related to the question of data provenance (where does "any" Data/Information/Knowledge come from?). In particular such a system has to be aware of the cumulated data of complex communication between services. If there is any communication between services inside

the system, a security system ensures the trustworthiness of data. However, the trustworthiness of data from outside the system can never be fully guaranteed. Since many systems require external data, minimizing the risk of uncertainty is key. E.g. weather data should come from external (and multiple) sensors to ensure correctness of the values and also legislation information or data from e.g. chemical-databases will also come "from the outside". Trustworthiness of sources or the provenance of data differs from source to source (e.g. values from governmental institutions can usually be given a higher trust value than from other third party providers).

Recent research work on provenance can be found in the following literature: In [21] ("Trust Evaluation Scheme of Web Data Based on Provenance in Social Semantic Web Environments"), the authors show a trust evaluation scheme of web data based on provenance in social semantic web environments. Castro et al. provide an application for tracking provenance information in distributed data system in [6] ("Transparently tracking provenance information in distributed data systems" - patent).

Zhao et al. [22] ("Research of Data Resource Description Method oriented Provenance") and Ram et al. [18] ("A semantic Foundation for Provenance Management") provide more theoretical and conceptual foundations for the usage and management of provenance.

C. Security

In our context security mainly refers to computer security (protection of IT systems, information systems, protection of hard- and software, prevention of undesired intruders, etc.) and information security. "Security is the degree of protection against danger, damage, loss, and crime." [2]

Concerning the threats when leaving data in the Cloud, we refer to our past research projects [20] and a related thesis [13]. The conclusion of this and further papers (like [8]) is that cloud security cannot be established in the way as it should be or as we wish, because the responsibility of security and safety is always the responsibility of the owner and provider of the cloud services. These are environment constraints which are unchangeable.

D. Risk

Risk in general addresses the potential of losing something with a special personal value. It is also seen as an intentional interaction with uncertainty, where the outcome is hard to predict [2].

Rousseau et al. [19] say that *"Risk is the perceived probability of loss, as interpreted by a decision maker [...]. The path-dependent connection between trust and risk taking arises from a reciprocal relationship: risk creates an opportunity for trust, which leads to risk taking."*

Relating to information technology or information processing systems, risk can also be categorized as IT risk. This area of risk is a wide area of possible incidents, where a loss of values can occur in many different ways.

E. Trust and Certainty in Knowledge Processing

To the best knowledge of the authors, there is no related work dealing with this topic directly – neither for processing trust and certainty, nor for the aggregation of (un)certainty.

A good approach for measuring trust is given in "An Approach to Evaluate Data Trustworthiness Based on Data Provenance" [10].

Recent research on modeling uncertainty is given by [14] and the usage of uncertainty in complex event processing can be found in [9].

III. Specification of Approach

A. Idea

A convenient approach for incorporating trust and certainty values into knowledge processing systems is currently being developed. In the next few paragraphs we will explain the principles of the development. The main subjects in our approach are:

- any Source (S), which provides information in the environment; there can be multiple sources in an environment.
- any Data (D)[1], which is provided by one Source; for our model, every source usually provides one or more data (elements).
- any Knowledge Processing System (KPS), which processes data from one or more sources; each KPS itself produces new data as output; in our model, every KPS produces only one output.

The source provides data in an abstract manner: it is not important which type of data it is – in our approach it can be a whole database as well as a single text file or a single data value. A knowledge processing system is any system using the provided data from the existing sources, processing it, and providing new data as output. To have computable and usable values in our approach, computation of these different values from existing input data is needed. The main values in our approach are:

- Trust value (T) of source (S), which defines how trustable the source is. The system (sources / data / knowledge processing systems) has to be seen as a whole environment, hence the trust level for one source should always be the same.
- Certainty value (C) of data (D), which describes how reliable, confident or steady the provided data is. In literature and research work many definitions of believability and certainty in knowledge based systems exist.

[1]In our work we combine the data and information layer referring to the Data-Information-Knowledge-Wisdom (DIKW) architecture in [3] from Russell Lincoln Ackoff, i.e. data has the role of information and belongs to the information layer.

- Importance value (I) of one input data (D), decided by the current knowledge processing system (KPS) for the current step of computation. We focus on importance in subsection III-B.

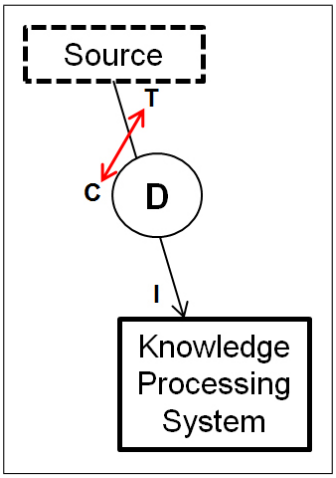

Figure 1: Introduction to our approach.

We introduced T, C and I and their relevance. Questions such as ”How to determine a value for trust T for a specific external source?” or ”How to determine a value for certainty C for specific input?”, as mentioned before, will not be answered in this paper, as the values are assumed to be given. This is due to the early development stage of our approach as already stated in the introduction. Related research on measuring trust and certainty are referenced in subsection II-E.

How the values for T, C, and I are defined and what their scopes will be, will be decided in a first attempt in subsection III-C. Answers for ”how the scope of trust T is determined for a specific source” and ”how the knowledge processing system decides on importance I for the data” will also be provided. Many approaches exist in literature to decide on certainty C, but it remains unclear how to define a value of certainty for the new approach. Another open question is the scope of values and if normalization is needed after calculation. For example, we propose values between 0 and 1 for certainty C and trust T, probably a three-step approximation for the importance I (e.g. 0.5 for unimportant, 1.0 for neutral, and 1.5 for important values). If the calculated new values (T_{new}, C_{new}) reach scopes above 1.0, the values have to be normalized for further usage (e.g. in a multi-step system, where several knowledge processing systems are calculating T, C, and I values multiple times). We also assume, that C and T can both be dependent and independent, which needs to be defined – see the arrow in figure 1.

Next, we will discuss the continuation of processing T, C, and I. If a model of application or calculation is provided, the new output value of the KPS has to be re-applied on trust – if the knowledge processing system generates an output which is used as an input for another knowledge processing system, a new trust value has to be considered for this output. This is a non-trivial problem, as it is not clear how the trustworthiness of your knowledge processing system should be measured or determined. Is it trustable because it is operated in a controlled environment? If it is an internal part of the overall system, it can be assumed to be a trustable source, but (as seen in figure 2) the initial values can come from an external source, where the trustworthiness is not guaranteed.

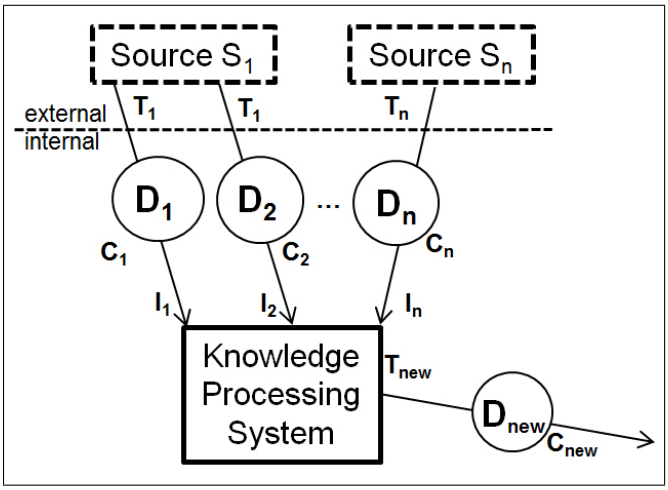

Figure 2: Details on our approach.

Another open question is how trust values should be handled in general for internal and external sources. It also has to be considered that there should always be the same trust values for the same external/internal source, however, the possibility that one source gains higher trust over time does exist (e.g. when the values are continuous and recognized as stable and certain).

A similar approach for data provenance and measuring believablity of data can be found in [17]. This, however, does not cover the usage in knowledge processing or knowledge based systems.

In subsection III-C the approach will be put into practice by defining scopes and calculation and answering several of the open questions.

B. *Introducing Importance*

In our approach we propose the use of an ”Importance (I)” factor, which is necessary in order to have different weights in the current step of calculation in a knowledge processing system. In this context the question of ”How does the knowledge processing system decide how important which input is?” arises. At this time we propose a staggered allocation of the importance values to avoid the necessity of normalizing calculated values as will be shown in the next subsection.

The importance belongs to one data but is defined by the KPS for each data separately - see figure 1 - therefore, it is possible that another KPS decides another importance for the same data, which will be shown in the provided scenarios further below. The explanation is the follows: in the KPS some computations are made with all the input data and the system itself has to decide about the importance of every single data in this specific context (assumption: there is only one main usage of all the input data and one computation which produces one output: another data D, respectively). In our approach every KPS has some degree of freedom to decide how important which data is for the current step of calculation.

In this context the term "Importance" could be substituted with e.g. "Influence" or something similar. The novelty in our approach is the usage of this factor to increase the representativeness of the calculated output. It is not our intention to feed data through the KPS without getting representative values.

As the formal definitions and boundaries are not yet specified, the current meaning of importance can be translated to influence, as the values for trust and certainty might heavily influence the results in every KPS.

C. Scopes & Calculation

We are now going to make a first attempt of concretizing the model described above by answering some of the questions and fixing the scopes of possible values as follows:

- Trust T of source S, for each S, has to be greater than 0 and less or equal than 1, where each value of T for each S has to be the same (if used multiple times) – a higher value represents higher trust:

$$0 < T \leq 1 \tag{1}$$

- Certainty C of data D, for each D, has to be greater than 0 and less or equal than 1, where each value of C for each D has to be the same (if used multiple times) – a higher value represents higher certainty:

$$0 < C \leq 1 \tag{2}$$

- Importance I of data D, decided by the KPS, is staggered:
  - 0.5 for values which are not very important
  - 1.0 for regular values, where no special impact on importance is given
  - 1.5 for very important values, concerning the current data processing

$$I = 0.5 \mid 1 \mid 1.5 \tag{3}$$

  Note: Regarding the current step of processing in the KPS for example: if data D_i is given the importance 1.5, there also has to be another data D_j with an importance of 0.5. There always has to be the same number of importance weighted D with 0.5 and with 1.5. The importance of 1.0, in fact, does not affect the current step of processing. This constraint guarantees avoiding an overestimation of grading input data too often as "very important", which would result in a deferral of representation of the output values. It also guarantees, that all calculated output values (T_{new} and C_{new}) stay in the scope between 0 and 1, resulting in the effect that no normalization is needed within the current scope of calculations.

Note: For this model it is necessary, that the input values of T and C are initialized as defined in (1) and (2). We do not investigate the calculation of T and C, as it is assumed that this step has been taken before!

We now define the formulas for processing new T and C values as outcome of a KPS. This is the arithmetical average of the input T or C weighted with the current I for each D.

$$T_{new} = \frac{1}{n}\sum_{i=1}^{n}(T_i * I_i) \tag{4}$$

Formula 4: Calculating T_{new} over all $T_{1\text{-}n}$ related to $I_{1\text{-}n}$.

$$C_{new} = \frac{1}{n}\sum_{i=1}^{n}(C_i * I_i) \tag{5}$$

Formula 5: Calculating C_{new} over all $C_{1\text{-}n}$ related to $I_{1\text{-}n}$.

D. Alternative Aggregation Functions

We have discussed several other methods to calculate T_{new} and C_{new} in our research. For example the following functions are also possible for calculation: (Note: in the equations 6-11, T_{new} and T_i can always be substituted with C_{new} and C_i to get the corresponding formulas, as our current intention is to compute T_{new} and C_{new} in the same manner).

$$T_{new} = \frac{1}{n}\sum_{i=1}^{n}T_i \tag{6}$$

$$T_{new} = \sum_{i=1}^{n}(T_i * I_i) \tag{7}$$

$$T_{new} = \frac{1}{n}\prod_{i=1}^{n}\frac{(T_i + I_i)}{2} \tag{8}$$

$$T_{new} = \frac{1}{n}\prod_{i=1}^{n}T_i \tag{9}$$

$$T_{new} = \prod_{i=1}^{n}\frac{(T_i + I_i)}{2} \tag{10}$$

$$T_{new} = \min_{n}(T_i) \tag{11}$$

For now, we use the aggregation functions 4 and 5 in the following sections, for our first approach.

IV. Scenarios

A. Simple Scenario

The following example relies on the provided model and formulas in section III-C. We introduce four sources (S_1 to S_4) with different trust values (T_1 to T_4), each providing one data (D_1 to D_4) with different certainty values (C_1 to C_4) for one knowledge processing system (KPS_A), which emphasizes the different importances (I_1 to I_4). The values are listed in table I as follows:

Table I: Initial values of T, C & I for a simple scenario.

S_1: T_1=0.8	D_1: C_1=0.9	KPS_A: I_1=1.0
S_2: T_2=0.4	D_2: C_2=0.2	KPS_A: I_2=1.5
S_3: T_3=0.9	D_3: C_3=0.2	KPS_A: I_3=0.5
S_4: T_4=0.2	D_4: C_4=0.7	KPS_A: I_4=1.0

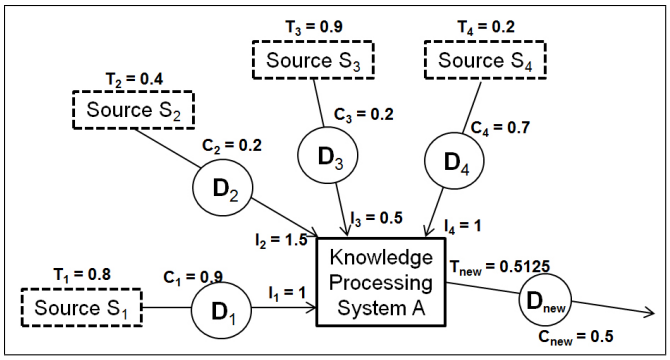

Figure 3: Simple Test Scenario.

The application of the formulas 4 and 5 are leading to the following results:

$$T_{new} = \frac{0.8 * 1 + 0.4 * 1.5 + 0.9 * 0.5 + 0.2 * 1}{4} = \\ = \frac{2.05}{4} = \underline{\underline{0.5125}} \tag{12}$$

$$C_{new} = \frac{0.9 * 1 + 0.2 * 1.5 + 0.2 * 0.5 + 0.7 * 1}{4} = \\ = \frac{2}{4} = \underline{\underline{0.5}} \tag{13}$$

As seen in figure 3 and in the calculation above, the outcome of T_{new} and C_{new} are very representative, regarding the input values, which were well balanced on purpose (the mixture of T and C values was chosen in a way, where all possible situations are represented with four inputs: high/low T with high/low C and each reversed).

Evaluation of Simple Scenario: The most impact on the final score was caused by S_2 due to its weighting by I_2=1.5 – both T and C values from this source are very low (0.4 and 0.2), which affects the final score in a meaningful way. The highest trust is provided by S_3, but because of its low importance in the current processing step, it does not affect the result that much (in fact, only 1/3 as much as S_2 does). The remaining trust values from S_1 and S_3 have a medium impact, as their weighted value is 0.5. The same argumentation is valid for the certainty values C in this scenario. This is the reason for the very average outcome of T_{new}= 0.5125 and C_{new}= 0.5. If you transfer the final results into a range of [0..100] (or percentage values), you can interpret them with T_{new}= 51.25% and C_{new}= 50.0% which can be seen as a good representative view on the whole systems' trust and certainty outcome.

We know that an application of our model in such a small use case shown here is only the representation of a simple reality. Most of the time, various processing steps occur in multiple knowledge processing systems. We demonstrate a more realistic example in the next subsection.

B. Advanced Scenario

For our advanced scenario, we are introducing six sources (S_1 to S_6) with different trust values (T_1 to T_6), each providing one or two data (D_{11}, D_{12}, D_2, D_{31}, D_{32}, D_4, D_{51}, D_{52}, and D_6) with different certainty values (C_{11}, C_{12}, C_2, C_{31}, C_{32}, C_4, C_{51}, C_{52}, and C_6) for multiple knowledge processing systems (KPS_A to KPS_E), which weight the different importances.

$KPS_{A\text{-}C}$ are working only with data from sources $S_{1\text{-}6}$, so the T and C values are given. KPS_D is working with data from source S_1 and $KPS_{A,B}$ and KPS_E is processing only output values from $KPS_{B,C,D}$ (no initial sources) – therefore, the calculation of T and C of KPS_D and KPS_E depend on the calculations of $KPS_{A,B,C}$, because they receive (most of) their input T and C values from previous processing steps. An overview of the advanced scenario including the calculated values is shown in figure 4 at the end of this section. The values for the first calculation step are listed in table II as follows:

An overview of the advanced scenario is shown in figure 4.

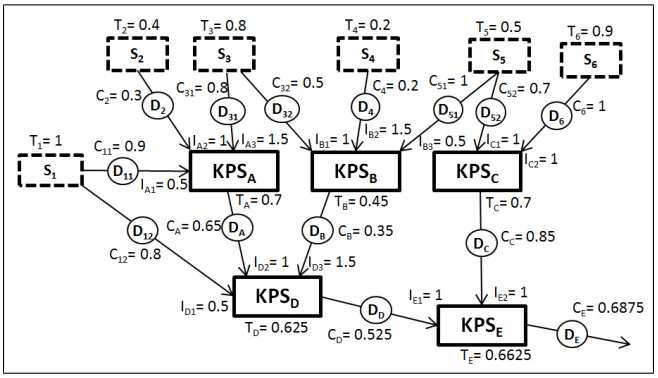

Figure 4: Advanced Test Scenario.

Table II: Initial values of T, C & I for an advanced scenario.

S_1: T_1=1.0	D_{11}: C_{11}=0.9	KPS_A: I_{A1}=0.5
S_2: T_2=0.4	D_2 : C_2 =0.3	KPS_A: I_{A2}=1.0
S_3: T_3=0.8	D_{31}: C_{31}=0.8	KPS_A: I_{A3}=1.5
	D_{32}: C_{32}=0.5	KPS_B: I_{B1}=1.0
S_4: T_4=0.2	D_4 : C_4 =0.2	KPS_B: I_{B2}=1.5
S_5: T_5=0.5	D_{51}: C_{51}=1.0	KPS_B: I_{B3}=0.5
	D_{52}: C_{52}=0.7	KPS_C: I_{C1}=1.0
S_6: T_6=0.9	D_6 : C_6 =1.0	KPS_C: I_{C2}=1.0

Due to the fictional character of this scenario (with no detailed information about the involved knowledge processing systems), all values are chosen freely. Particularly the importance values are picked in a way to demonstrate the practicality as much as possible.

With these input values, we are able to compute trust and certainty values for the output data of $KPS_{A\text{-}C}$ in a first step with the formulas 4 and 5, similar to the simple scenario in the previous section:

$$T_A = \frac{1*0.5 + 0.4*1 + 0.8*1.5}{3} = \frac{2.1}{3} = 0.7 \quad (14)$$

$$C_A = \frac{0.9*0.5 + 0.3*1 + 0.8*1.5}{3} = \frac{1.95}{3} = 0.65 \quad (15)$$

$$T_B = \frac{0.8*1 + 0.2*1.5 + 0.5*0.5}{3} = \frac{1.35}{3} = 0.45 \quad (16)$$

$$C_B = \frac{0.5*1 + 0.2*1.5 + 1*0.5}{3} = \frac{1.05}{3} = 0.35 \quad (17)$$

$$T_C = \frac{0.5*1 + 0.9*1}{2} = \frac{1.4}{2} = 0.7 \quad (18)$$

$$C_C = \frac{0.7*1 + 1*1}{2} = \frac{1.7}{2} = 0.85 \quad (19)$$

To provide a structured way of progress, we accumulate the calculated values in table III, for the ongoing process of calculating the output of KPS_D and KPS_E. Note, that KPS_A, KPS_B, and KPS_C act as new/additional sources for the ongoing calculations.

Table III: Calculated values and initial values for processing output of KPS_D.

S_1: T_1=1.0	D_{12}: C_{12}=0.80	KPS_D: I_{D1}=0.5
KPS_A: T_A=0.7	D_A : C_A =0.65	KPS_D: I_{D2}=1.0
KPS_B: T_B=0.45	D_B : C_B =0.35	KPS_D: I_{D3}=1.5

With these calculations, we can now continue finding the values for KPS_D.

$$T_D = \frac{1*0.5 + 0.7*1 + 0.45*1.5}{3} = \frac{1.875}{3} = 0.625 \quad (20)$$

$$C_D = \frac{0.8*0.5 + 0.65*1 + 0.35*1.5}{3} = \frac{1.575}{3} = 0.525 \quad (21)$$

The important values for calculating the output of KPS_E are shown in table IV:

Table IV: Values for final processing step.

KPS_C: T_C=0.70	D_C: C_C=0.85	KPS_E: I_{E1}=1.0
KPS_D: T_D=0.625	D_D: C_D=0.525	KPS_E: I_{E2}=1.0

With these calculated values, we can now proceed finishing the scenario by computing the values for KPS_D and KPS_E, where KPS_E generates the final output values of this scenario.

$$T_E = \frac{0.625*1 + 0.7*1}{2} = \frac{1.325}{2} = \underline{\underline{0.6625}} \quad (22)$$

$$C_E = \frac{0.525*1 + 0.85*1}{2} = \frac{1.375}{2} = \underline{\underline{0.6875}} \quad (23)$$

Evaluation of Advanced Scenario: The results of this advanced scenario are:

- Trust T_E of KPS_E is computed with 0.6625
- Certainty C_E of D_E is computed with 0.6875

Here the sources S_3 and S_4 have the highest impact in the whole calculation concerning trust, because of their high importance in KPS_A and KPS_B as well as in the further calculation step in KPS_D. Concerning certainty, the low value of C_4=0.2 has high influence in this model, as its importance is rated with 1.5.

If you transfer the final results into a range of [0..100] (or percentage values), you can interpret them with T_E= 66.25% and C_E= 68.75% which can be seen as a good representation of the whole systems' trust and certainty outcome, similar to the outcome in the simple scenario.

C. Scenario in the Agricultural Domain

1) Disease Pressure Model: We are now referring to the DPM (Disease Pressure Model, used in the Project CLAFIS [1]) for calculating an accurate daily risk value. This shows how certain a specific disease outbreak for a specific agricultural field can be. The DPM is provided in figure 5, together with an explanation of the single parts and a description of the used functions.

The DPM uses input values from a FMIS (farm management information system), which stores information such as this year's and last year's crop as well as the used tillage method. The needed weather data comes from several weather stations, which gathers information such as temperature, relative humidity, amount of rainfall, and wind speed is gathered.

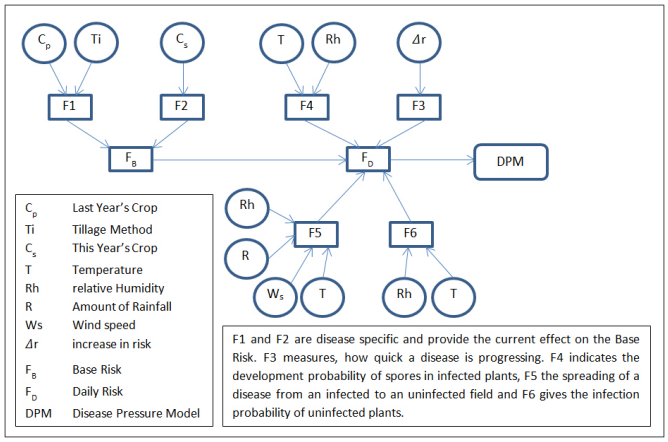

Figure 5: Disease Pressure Model (DPM).

2) Application of our approach on DPM: We now apply our approach on the model of DPM. The whole process of calculation can be seen in figure 6. The initial values for the first steps of calculation are listed in table V.

Table V: Initial values of T, C and I for calculating F1, F2, F4, F5, and F6.

T_{FMIS}=1	C_{Cp}=0.8	F1: $I_{F1\text{-}Cp}$=1
	C_{Ti}=0.9	F1: $I_{F1\text{-}Ti}$=1
	C_{Cs}=0.8	F2: $I_{F2\text{-}Cs}$=1
$T_{WeatherStation1}$=0.9	C_{T}=0.9	F4: $I_{F4\text{-}T}$=0.5
		F6: $I_{F6\text{-}T}$=0.5
		F5: $I_{F5\text{-}T}$=1
	C_{Rh}=0.8	F4: $I_{F4\text{-}Rh}$=1.5
		F6: $I_{F6\text{-}Rh}$=0.5
		F5: $I_{F5\text{-}Rh}$=1
$T_{WeatherStation2}$=0.7	C_{R}=0.7	F5: $I_{F5\text{-}R}$=0.5
	C_{Ws}=0.8	F4: $I_{F5\text{-}Ws}$=1.5

The application of the formulas 4 and 5 are leading to the following results of trust (T_{F1}, T_{F2}, T_{F4}, T_{F5}, T_{F6}) and certainty (C_{DF1}, C_{DF2}, C_{DF4}, C_{DF5}, C_{DF6}):

$$T_{F1} = 1 \quad T_{F2} = 1 \tag{24}$$

$$C_{DF1} = 0.85 \quad C_{DF2} = 0.8 \tag{25}$$

$$T_{F4} = 0.9 \quad T_{F6} = 0.9 \quad T_{F5} = 0.8 \tag{26}$$

$$C_{DF4} = 0.825 \quad C_{DF6} = 0.875 \quad C_{DF5} = 0.8125 \tag{27}$$

In order to process the final calculations we have to compute the outcome of FBase first. The needed input values for FBase are listed in table VI.

Table VI: Values for calculating the outcome of FBase.

T_{F1}=1	C_{DF1}=0.85	FB: $I_{FB\text{-}DF1}$=1
T_{F2}=1	C_{DF2}=0.8	FB: $I_{FB\text{-}DF2}$=1

$$T_{FB} = 1 \qquad C_{DFB} = \frac{0.85 + 0.8}{2} = 0.825 \tag{28}$$

Having finished the the calculation of trust and certainty of FBase, F4, F6, and F5, we are now able to continue the scenario.

Table VII: Values for calculating the outcome of FDaily.

T_{FB}=1	C_{DFB}=0.825	FD: $I_{FD\text{-}DFB}$=1.5
T_{F4}=0.9	C_{DF4}=0.825	FD: $I_{FD\text{-}DF4}$=1
T_{F6}=0.9	C_{DF6}=0.875	FD: $I_{FD\text{-}DF6}$=1
T_{F5}=0.8	C_{DF5}=0.8125	FD: $I_{FD\text{-}DF5}$=0.5

With the calculated values, we can now finish the scenario by measuring the outcome of trust T_{FD} and certainty C_{DFD} of the function FDaily. The needed values for calculating FDaily are listed in table VII.

$$T_{FD} = \frac{1.5 * 1 + 0.9 * 1 + 0.9 * 1 + 0.5 * 0.8}{4} = \\ = \frac{3.7}{4} = \underline{\underline{0.925}} \tag{29}$$

$$C_{DFD} = \frac{1.5 * 0.825 + 0.825 * 1 + 0.875 * 1 + 0.5 * 0.8}{4} = \\ = \frac{3.34375}{4} = 0.8359375 = \underline{\underline{0.836}} \tag{30}$$

Evaluation of DPM Scenario: The DPM model is a good example for our approach, as the sources (FMIS and Weather Stations) are "under our control". This means, that these sources are highly trustable (T_{FMIS}=1, $T_{WeatherStation1}$=0.9, $T_{WeatherStation2}$=0.7 – the chosen trust value of $T_{WeatherStation2}$ causes a better variation in our model – in fact, it would be similarly high as $T_{WeatherStation1}$ in reality). The provided data is very likely to be accurate as it is entered into the FMIS by the person who controls the system (often the farmer himself). The data provided by the weather stations can also be assumed to be accurate.

The most important step in the calculation of the DPM is the provision of the base risk where the outcome of FBase is very useful for further calculations: T_{FB} stays on the highest possible value 1 and the certainty of C_{DFB}=0.825 is also a very good indicator for FDaily. Particularly because of the importance of FBase in FDaily $I_{FD\text{-}DFB}$=1.5, its affection is quite high on the last step of calculation.

With the outcome of FDaily with T_{FD}=0.925 and the certainty C_{DFD}=0.836 of FDaily's produced output D_{FD}, we can sum up, that the DPM is producing quite trustable and certain values. If you transfer the final results into a range of [0..100] (or percentage values), you can interpret them with T_{FD}=92.5% and C_{DFD}=83.6%. These can be considered to be good values for trust and certainty.

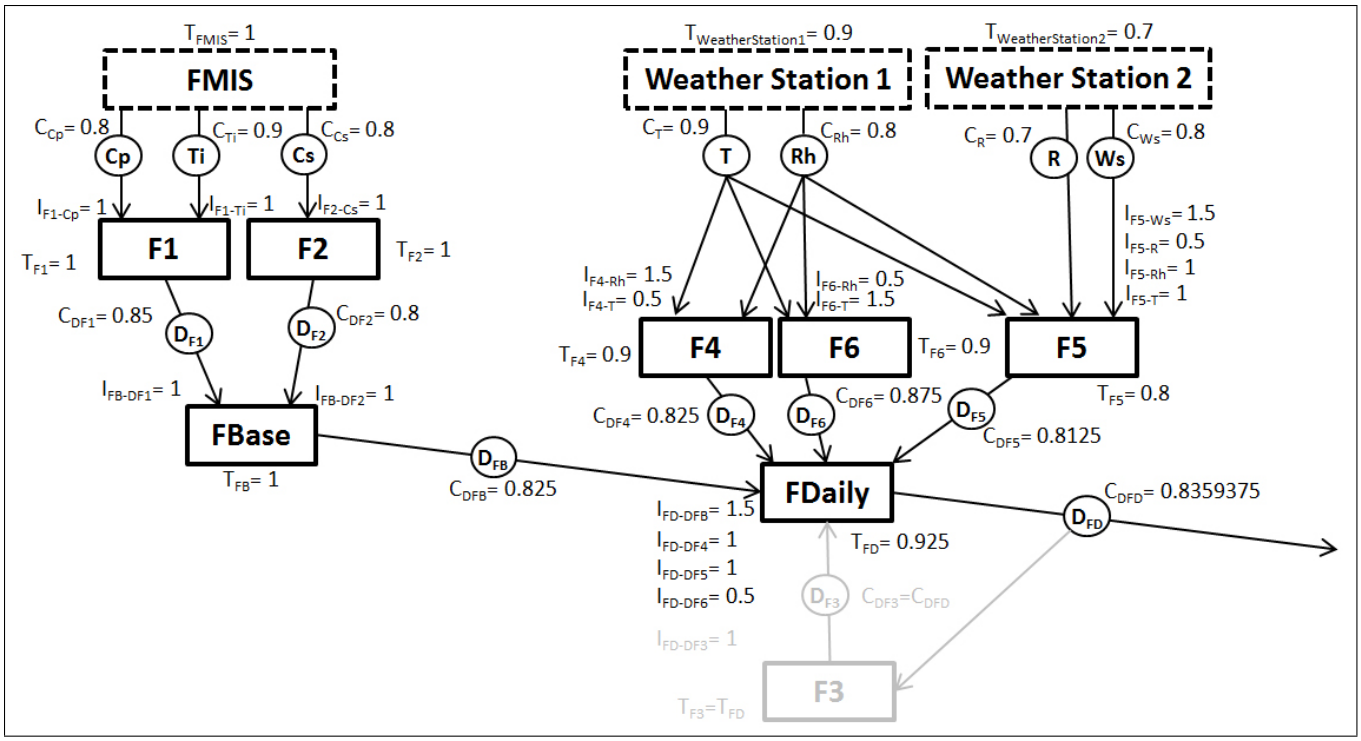

Figure 6: Approach applied on disease pressure model (DPM).

Some parts of the graphic above (figure 6) are marked in light grey: the values of function F3, are the same as the output of FDaily, would be included as an input for the calculation of FDaily itself. As recursion is not yet covered in our approach, we do not consider the function F3 in this work and in the calculation of trust and certainty values in the DPM. However, we will consider recursion in further research work, therefore improving our approach.

V. Summary & Outlook

We addressed the question of how to determine trust- and certainty-values of a KPS output, when different trust- and certainty-values are given for the input data.

We provided a first solution on a simple and advanced example as well as on a real world scenario, the disease pressure model (DPM). The results are realistic and the computed values are promising.

Further steps such as analyzing runtime-complexity, proof of non-converging, evaluation of usage of the approach, experiments and testing the approach on several more realistic multi-step scenarios, and their evaluation will be done in further work. Additionally, we will evaluate of more complex aggregation functions, hereby incorporating statistical distributions of trust and certainty values. Moreover, we will consider recursion in our approach and dealing with questions like ”Is staggering of Importance (I) needed?” and ”Are T and C (in)dependent?”.

A philosophical element has to be discussed too: ”Are we allowed to alter a trust value according to its importance?”. Interpreting and calling it an influence would probably be less controversial. However, it does not eliminate the underlying aspect and the much needed discussion.

The implications for research can be expected as stated in section II-E. There are no other developed approaches concerning the processing of trust and certainty, neither for their aggregation. As soon as the development of the approach has been completed, it will cover the entire life-management of trust and certainty in a system. Therefore, the system output can be assessed much better. Its novelty and innovation will have a profound impact on further research in this area.

Our aim is to develop a complete model for calculating representative values in KPS by incorporating trust, certainty and importance values. This approach can then be applied to all other processing systems as well. Such a system, which can be applied to a variety of applications, would be incredibly useful in practice.

In further research stages, we will take the semantic foundation of provenance management (related to Ram et al. [18], (W7 - What, How, Where, When, Who, Which, and Why of data and its history)) into account and will cover the origin of data and the measurement, definition, determination and grading of trust and certainty.

ACKNOWLEDGMENT

The research leading to these results has received funding partly from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no.604659 and partly from the federal county of Upper Austria.

The authors would like to thank all partners in CLAFIS [1] for the excellent cooperation within the project. The collaboration between experts from different fields made it possible to write a paper with this content. Special thanks to our colleagues from LUKE in Finland for the development of the DPM (disease pressure model). This allowed us to apply the approach to a real world scenario in our current project.

Thanks to Lara Aigmüller und Lukas Stieger for final proofreading, correcting some of our special german-english wordings, and improving the quality of this publication.

REFERENCES

[1] CLAFIS: Crop, livestock and forests integrated system for intelligent automation, 2013-2016. EU Seventh Framework Programme NMP.2013.3.0-2.

[2] Wikipedia, the free Encyclopedia, 2016.

[3] R. L. Ackoff. From data to wisdom. *Journal of Applied System Analysis*, 16:3–9, 1989.

[4] Peter Buneman and Susan B Davidson. Data provenance–the foundation of data quality, 2010.

[5] Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Data provenance: Some basic issues. In Sanjiv Kapoor and Sanjiva Prasad, editors, *FST TCS 2000*, volume 1974 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2000.

[6] P.C. Castro, M. Pistoia, and J. Ponzo. Transparently tracking provenance information in distributed data systems, August 7 2014. US Patent App. 13/761,916.

[7] Wang chiew Tan. Research problems in data provenance. *IEEE Data Engineering Bulletin*, 27:45–52, 2004.

[8] Mihai Christodorescu, Reiner Sailer, Douglas Lee Schales, Daniele Sgandurra, and Diego Zamboni. Cloud security is not (just) virtualization security: A short paper. In *Proceedings of the 2009 ACM Workshop on Cloud Computing Security*, 2009.

[9] Gianpaolo Cugola, Alessandro Margara, Matteo Matteucci, and Giordano Tamburrelli. Introducing uncertainty in complex event processing: model, implementation, and validation. *Computing*, 97(2):103–144, 2015.

[10] Chenyun Dai, Dan Lin, Elisa Bertino, and Murat Kantarcioglu. An approach to evaluate data trustworthiness based on data provenance. In *Proceedings of the 5th VLDB Workshop on Secure Data Management*, SDM '08, pages 82–98, Berlin, Heidelberg, 2008. Springer-Verlag.

[11] Sara Denize and Louise Young. Concerning trust and information. *Industrial Marketing Management*, 36(7):968 – 982, 2007. Opening the network - Bridging the {IMP} tradition and other research perspectives2006 {IMP} Conference Special Issue22nd Industrial Marketing and Purchasing Group Conference.

[12] Morten Hertzum, Hans H.K Andersen, Verner Andersen, and Camilla B Hansen. Trust in information sources: seeking information from people, documents, and virtual agents. *Interacting with Computers*, 14(5):575 – 599, 2002.

[13] Markus Jäger. Sicherheitsaspekte bei Virtualisierungen. Master's Thesis, 7 2014.

[14] Alexander Karlsson, Björn Hammarfelt, H. Joe Steinhauer, Göran Falkman, Nasrine Olson, Gustaf Nelhans, and Jan Nolin. Modeling uncertainty in bibliometrics and information retrieval: an information fusion approach. *Scientometrics*, 102(3):2255–2274, 2015.

[15] Markus Jäger, Stefan Nadschläger and Trong Nhan Phan. Towards the trustworthiness of data, information, knowledge and knowledge processing systems in smart homes. IDIMT, 2015.

[16] D. Harrison McKnight. *Trust in Information Technology*. The Blackwell Encyclopedia of Management: Operations management. Blackwell Pub., 2005.

[17] Nicolas Prat and Stuart Madnick. Measuring data believability: A provenance approach. *Hawaii International Conference on System Sciences*, 0:393, 2008.

[18] Sudha Ram and Jun Liu. A semantic foundation for provenance management. *Journal on Data Semantics*, 1(1):11–17, 2012.

[19] Denise Rousseau, Sim Sitkin, Ronald Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 1998.

[20] Rudolf Hörmanseder and Markus Jäger. Cloud security problems caused by virtualization technology vulnerabilities and their prevention. In *IDIMT-2014 Networking Societies - Cooperation and Conflict*, 2014.

[21] Sangwon Yoon, Kitae Choi, Jaeyeol Park, Jongtae Lim, Kyoungsoo Bok, and Jaesoo Yoo. Trust evaluation scheme of web data based on provenance in social semantic web environments. *Journal of KIISE*, 43(1):106–118, 2016.

[22] Yan-peng Zhao, Chao-fan Dai, and Xiao-yu Zhang. Research of data resource description method oriented provenance. In *Proceedings of the 22nd International Conference on Industrial Engineering and Engineering Management 2015*, pages 215–224. Springer, 2016.