Walid Cherifi, Bolesław Szafrański, Grzegorz Bliźniuk

Faculty of Cybernetics
Military Unversity of Technology, Warsaw

Towards evidence-based data conflict resolution in data integration process

1. Introduction

The public sector in Poland and other countries is made up of many different organizations, ranging from large government departments to universities, health care facilities and libraries.[1] Moreover, it generally comprises of different segments, i.e. defence, finance, education, health, environment etc. They each face different challenges, but the common theme for these diverse segments is the need for efficiency, visibility, and transparency.[2] This decentralized structure of public administration suggests that in certain cases public agencies at different administration levels and different functional areas produce, gather, and disseminate similar data i.e. data about the same real-world objects. This situation results in a number of challenges regarding the quality of data, as it is possible that the disseminated data is incomplete, controversial and/or obsolete.[3] Therefore, finding ways to integrate and bring diverse data sets together has the potential to increase the government's transparency, improve the functioning of public administration, contribute to economic growth and provide social value to citizens.[4] However, to reach this goal, a difficult technical problem has to be solved first: the integration of typically distributed, inherently heterogeneous, and possibly inconsistent data sources.

Data integration systems harmonize data from different independent sources into a single coherent representation. They aims to provide a unified access to a set of data sources in a specific application domain, such as business, technology, government,

[1] E. Ziemba, I. Obłąk, *The survey of information systems in public administration in Poland*, "Interdisciplinary Journal of Information, Knowledge and Management" 2011, vol. 9, pp. 31–56.

[2] E. Kalampokis, E. Tambouris, K. Tarabanis, *Open government data: A stage model*, "Electronic Government" 2011, Lecture Notes in Computer Science 6846, pp. 235–246.

[3] M. Fatehali, *Building the business case for Master Data Management in the Public Sector*, "Oracle White Paper" 2011.

[4] Ibid.

healthcare, sports and tourism, where users can put their queries to the system and wait to receive a correct, concise and complete answers collected from distinct sources. This can be done by resolving the heterogeneities and offering an integrated view to the disparate sources. Then, users are able to submit queries over this uniform view without having to spend a lot of time to access all data sources separately.

The most important challenge for data integration is to provide the users with data of high quality. This means that the collected data must be as complete and accurate as possible. Whereas high completeness can be achieved by adding more data sources to the integration system, reaching the accuracy is not an easy task. Indeed, various facts about the same real-world object can be gathered from diverse sources. For instance, a patient's medical records can be obtained from several hospitals; a customer's information may get collected from multiple databases in the company; and finally, the observation and registration of natural events is carried out by different laboratories. Unfortunately, these diverse sources are generally of various quality and often provide unreliable and conflicting information. Moreover, decisions based on inaccurate information usually lead to severe harm. For example, wrong diagnosis based on incorrect measurements of a patient will absolutely lead to serious consequences; erroneous account information in a company's database may cause financial losses; and scientific discoveries may be guided in the wrong direction, if they are derived from incorrect data.[5] Therefore, resolving those conflicts is a crucial step before providing data to the requester.

In this paper, we propose a new approach that resolves the conflict between contradictory duplicate records. Our proposal is based on the evidence theory, which provides a powerful framework for representing uncertain and imprecise information better than probability functions do. Indeed, unlike the probability theory, the evidence theory is able to express in a more faithful manner a whole continuum of information availability: from complete or partial ignorance to total knowledge. Besides, it offers a mathematical way to combine evidence from different experts without the need to know about a priori or conditional probabilities. Therefore, this theory seems to provide an excellent tool for the issue of conflict resolution.

The remainder of this paper is structured as follows. Section 2 describes related previous work regarding data fusion, known also as truth finding or conflict resolution. Section 3 briefly presents some basic concepts of the evidence theory. Section 4 details our proposed evidence-based conflict resolution model. Finally, Section 5 concludes the paper and discusses some future directions for our work.

[5] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, J. Han, *Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation*, Proceedings of the 2014 SIGMOD Conference.

2. Related work

Data fusion is the problem of resolving conflicting values from multiple sources, and combining different representations of the same real world object into one single representation. The importance of this problem in data integration systems made it an active research topic.

First approaches to data fusion methods were typically baseline, such as, considering the value which has the highest number of occurrences in the case of categorical data, or taking the average/maximum/minimum for numerical values, where the focus was only on improving efficiency with the use of database queries. Bleiholder and Naumann[6] summarized the most commonly used baseline function and classified the conflict resolution into three main strategies based on the way of handling conflicting data: ignorance, avoidance, and resolutions.

Afterwards, more advanced solutions were proposed that apply probabilistic Bayesian reasoning to resolve the conflicts.[7] In fact, Yin, Han, and Yu[8] were the first to formally address the conflict resolution problem. This probabilistic method uses an iterative mechanism to jointly infer the truth by exploiting the mutual dependency between source accuracy and fact trustworthiness. After that, Dong, Berti-Equille and Srivastava[9] modified the aforementioned method in the way that different values provided on the same data item are disjoint and their probability must sum to 1.

Earlier studies also focused on other aspects such as the relationship between sources and more complex data types. Dong, Berti-Equille and Srivastava[10] analysed the copying relationships between the sources by discounting the vote count of the copier sources. Blanco et al.[11] also specified that is worthwhile to consider complex data instead of atomic values. Li et al.[12] integrated the conflict resolution process for diverse data

[6] J. Bleiholder, F. Naumann, *Data fusion*, "ACM Computing Surveys" 2008, vol. 41, no. 1, pp. 1–41.

[7] For a recent survey see: Li X., Dong X.L., Lyons K.B., Meng W., D. Srivastava, *Truth finding on the deep web: Is the problem solved?*, Proceedings of the VLDB 2013 vol. 6, no. 2.

[8] X. Yin, J. Han, P.S. Yu, *Truth discovery with multiple conflicting information providers on the web* "SIGKDD" 2007.

[9] X.L. Dong, L. Berti-Equille, D. Srivastava, *Integrating conflicting data: The role of source dependence,* Proceedings of the VLDB 2009, vol. 2, no. 1, pp. 550–561.

[10] Ibid. and X.L. Dong, L. Berti-Equille, D. Srivastava, *Truth discovery and copying detection in a dynamic world*, Proceedings of the VLDB 2009, vol. 2, no. 1, p. 573.

[11] L. Blanco, et al., *Probabilistic models to reconcile complex data from inaccurate data sources*, Conference on Advanced Information Systems Engineering 2010, pp. 83–97.

[12] Q. Li, et al., *Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation*, Proceedings of the 2014 SIGMOD Conference.

types seamlessly and modelled it as an optimization problem. Yin and Tan[13] studied the problem of data fusion with semi-supervised graph learning by using a small set of known truth data to help distinguish true facts from false ones and recognize accurate data sources.

In this work, we apply a new evidential approach based on Dampster Shafer theory. In fact, we are not aware of any work that exploited the belief theory in the conflict resolution problem.

3. Review of the evidence theory

The evidence theory, also called Dempster-Shafer theory or theory of belief functions, was first introduced by Dempster[14] in order to represent some imprecise information with upper and lower probabilities. Then, Shafer rebuilt the mathematical theory around the Dempster concept by introducing degrees of belief rather than lower probabilities.[15] This theory is well-known for its usefulness to express uncertain judgement of experts and its efficiency to represent imperfect (uncertain, imprecise and/or incomplete) information. This section presents some of its basic concepts.

3.1. Frame of discernment

In the evidence theory, the frame of discernment, also known as universe of discourse $\Theta = \{H_0, H_1, ..., H_N\}$, is a set of N mutually exclusive and exhaustive hypotheses. These hypotheses are all the possible and eventual solutions of the studied problem. The set of all subsets of Θ is its power set 2^Θ. A subset of those 2^Θ sets may consist of a single hypothesis or a conjunction of hypotheses.

3.2. Basic belief assignment

The main element of this theory is the basic belief assignment (*bba*), known also as mass function. A *bba* represents the degree of belief and is defined as a mapping $m : 2^\Theta \longrightarrow [0,1]$ satisfying the properties of equation (1).

[13] X. Yin, W. Tan, *Semi-supervised truth discovery*, Proceedings from the WWW Conference 2011, pp. 217–226.

[14] A.P. Dempster, *Upper and Lower probabilities induced by a multivalued mapping*, “Annals of Mathematical Statistics” 1967, vol. 38, pp. 325–339.

[15] G. Shafer, *A mathematical theory of evidence*, Princeton University Press 1976.

One or many subsets $H \in 2^{\Theta}$ may have a non-null mass and are considered *focal elements*. This mass is the source's degree of belief that the solution of the problem under study is in that subset. A situation of total ignorance is given by $m(\Theta) = 1$ and of total certainty by $m(H_i) = 1$ where H_i represents a singleton proposition.

$$\begin{aligned} & m(\emptyset) = 0 \\ & m(H) \geq 0, \forall H \in 2^{\Theta} \\ & \sum_{H \in 2^{\Theta}} m(H) = 1 \end{aligned} \tag{1}$$

3.3. Belief functions

In the framework of the evidence theory, several functions (we call them belief functions) are in one to one correspondence with the *bba*:

- The *belief function* (*bel*) is computed from a *bba* m . $bel(A)$ is the minimal belief allocated to *A* justified by available information on *B* $(B \subseteq A)$:

$$\begin{aligned} bel : 2^{\Theta} & \rightarrow [0,1] \\ A & \mapsto \sum_{B \subset A, B \neq \emptyset} m(B) \end{aligned} \tag{2}$$

- The *plausibility function* (*pl*) is also derived from a *bbam*. $pl(A)$ is the maximal belief affected to *A* justified by information on *B* that are not contradictory with *A* $(A \cap B \neq 0)$:

$$\begin{aligned} pl : 2^{\Theta} & \rightarrow [0,1] \\ A & \mapsto \sum_{B, \mathrm{A} \cap B \neq \emptyset} m(B) \end{aligned} \tag{3}$$

The above bel and pl measures can be viewed as the lower and upper bound of probability. From the definition, we have $bel(A) \leq pl(\mathrm{A})$. Their difference, $pl(\mathrm{A}) - \mathrm{bel}(\mathrm{A})$ indicates the degree to which the evidence set is uncertain whether to support *A* or $\overline{A}$.

3.4. Combining evidence sets

A basic belief assignment is treated as some belief assignment on domain of values. It is possible to have multiple mass functions on the same domain Θ that correspond to different experts' opinions. A great number of combination rules are proposed, such as the Dempster's Rule of combination,[16] which can be used to combine several independent sources. Given two *bbas* m_1 and m_2 associated to two independent evidence sources, the combined mass, denoted $m_{1\oplus 2}(H) = m_1 \oplus m_2(H)$, is defined as follows:

$$m_{1\oplus 2}(H) = m_1 \oplus m_2(H) = \begin{cases} \dfrac{\sum\limits_{H_1 \cap H_2 = H} m_1(H_1) \times m_2(H_2)}{1 - \sum\limits_{H_1 \cap H_2 = \emptyset} m_1(H_1) \times m_2(H_2)} \forall H \subseteq \Theta, H \neq \emptyset \\ 0 \qquad\qquad if\, H = \emptyset \end{cases} \tag{4}$$

The denominator is interpreted as a measure of conflict between the pieces of evidence and evaluating the quality of combination.

3.5. Discounting of information

In practice, sources of evidence may not be completely reliable, to reflect this, we can weaken the *bba* by introducing a discount rate α between 0 and 1[17] by which the mass function may be discounted in order to reflect the accuracy of a source. The discounted mass function using α is represented as:

$$\begin{aligned} m^{\alpha}(H) &= (1-\alpha)m(H) \quad for\, H \in \Theta \\ m^{\alpha}(\Theta) &= \alpha + (1-\alpha)\mathrm{m}(\Theta) \end{aligned} \tag{5}$$

When $\alpha = 0$ the source is absolutely accurate and when $\alpha = 1$ the source is completely inaccurate. After discounting, the source is treated as totally reliable.

3.6. Decision-making

In order to make the best decision, it is usually preferable to use a well-defined probability function. Smets[18] proposed the pignistic transformation which is constructed

[16] K. Sentz, S. Ferson, *Combination of evidence in Dempster-Shafer theory*, SANDIA Technical Report 2002, SAND2002–0835.

[17] Z. Elouedi, K. Mellouli, P. Smets, *Assessing sensor reliability for multisensor data fusion within the transferable belief model,* "IEEE Transactions on Systems, Man, and Cybernetics" 2004, vol. 34, no. 1, pp. 782–787.

[18] P. Smets, *Decision making in the TBM: the necessity of the pignistic transformation*, "International Journal of Approximate Reasoning" 2005, vol. 38, pp. 133–147.

from the basic belief assignments. This pignistic transformation aims to take the optimal decision, i.e., the one that maximizes the expected utility. It is defined by:

$$BetP(H) = \sum_{H_2 \subseteq \Theta} \frac{|H_1 \cap H_2|}{|H_2|} m(H_2), \forall H_1 \subseteq \Theta \tag{6}$$

The pignistic transformation can be useful if we want to compare different uncertain measures. The pignistic probability is used in the decision phase to select the most likely singleton hypothesis as a solution for the problem under study.

4. The proposed Evidence-Based Conflict Resolution method

We start with defining how we model data for the method proposed here. Then, we describe the proposed model.

4.1. Data model

To make the presentation clear and to facilitate the later discussions, we will start by explaining some concepts that are important to understand our proposal:

- *Data Source:* It is the source which provides information (facts) that may be conflicting, such as databases, web sites, etc. In our case, we assume we have $|S|$ data sources. A set of data sources can be represented as $S = \left\{ s_1, s_2, \ldots, s_{|S|} \right\}$, where $s_i \ (1 \le i \le |S|)$ is the ith data source.
- *Object:* An object is a real world entity which is recognized as being capable of an independent existence and which can be uniquely identified, such as a country, a patient, a natural event etc. We assume we have $|O|$ objects. The set of all objects can be presented as follows: $O = \left\{ o_1, o_2, \ldots, o_{|O|} \right\}$.
- *Attribute:* Obviously, an attribute represents a particular aspect of a real world object, such as the capital city of a country, the name of a patient, the duration of a natural event. We assume we have $|At|$ entity attributes. The set of all entity attributes can be expressed as $At = \left\{ a_1, a_2, \ldots, a_{|At|} \right\}$.
- *Fact:* For each attribute, the value provided by a data source can be called fact. For example, for an entity attribute At_i (the capital city of “Poland”), the data source s_i provides the fact f_j (“Warsaw”). A set of facts can be expressed as $F = \left\{ f_1, f_2, \ldots, f_{|F|} \right\}$.

- *Similarity between facts*: Two facts f_1 and f_2 on the same subject may be consistent or in conflict with each other. A function $sim\left(f_1, f_2\right)$ is provided to show the degree of consistency or conflict between them (0≤sim(f_1, f_2)≤1). $sim\left(f_1, f_2\right)=0$ means that the facts are in total conflict, whereas $sim\left(f_1, f_2\right)=1$ shows that they are totally similar. The similarity function is domain-specific and is generally provided by experts of the domain. The similarity function should be symmetric ($sim\left(f_1, f_2\right)=sim\left(f_2, f_1\right)$), and $sim\left(f_i, f_i\right)=1$ for any fact f_i $\left(1 \leq i \leq \left|F\right|\right)$.
- *Data Conflict:* Data conflict arises when different data sources provide different facts for the same attribute. For instance: f_1 ("Gdansk") versus f_2 ("Warsaw").

To illustrate our model and facilitate the understanding of our proposal, we can use the following example (see: Figure 1).

We suppose we have three data sources s_1, s_2 and s_3 that provides facts $f_1 = "Gdansk"$ and $f_2 = "Warsaw"$ about the attribute $a_1 = "capital"$ describing the capital city of the object $o_1 = $"Poland" which represents the country. In this example, it is clear that f_1 and f_2 are conflicting facts since $sim(f_1, f_2) = 0$. Our aim here is to resolve the conflict and choose the correct fact.

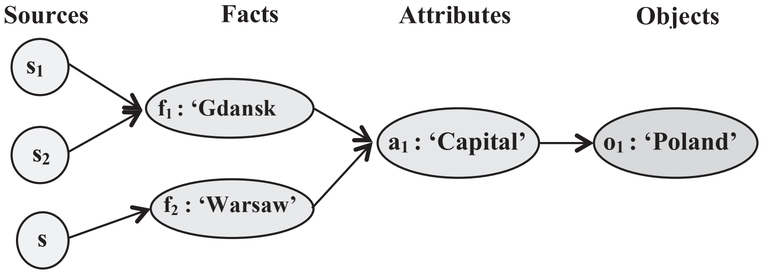

Figure 1. Example of data conflict between two ways of the same data writing

4.2. Frame of discernment

Our aim in this study is to resolve the conflict between conflicting facts, and select the most appropriate one. To do so, we define the following frames of discernment.

Let $\Theta_i = \{A_i, \overline{A}_i\}$ be the frame of discernment for each data source s_i. A_i means that the source s_i is accurate, while $\overline{A}_i$ expresses that the source is inaccurate. The hypothesis $A_i \cup \overline{A}_i$ represents total ignorance.

$\Omega_j = \{T_j, \overline{T}_j\}$ is the frame of discernment for each fact f_i. T$_j$ expresses that the fact is trustworthy, $\overline{T}_j$ shows that it is untrustworthy, and $T_j \cup \overline{T}_j$ means the total ignorance. Here, the ignorance arises because of the lack of knowledge.

In our example we have three frames of discernment Θ_1, Θ_2 and Θ_3 for each data source s_1, s_2 and s_3 respectively. And two frames of discernment Ω_1 and Ω_2 for each fact f_1 and f_2 respectively.

4.3. Evidence construction

In this subsection, we define an evidence-based conflict resolution model, which is a generalization of the probabilistic model proposed in the literature.[19] We wish to emphasize that the proposed model is to be considered as a proposal and that other models are possible.

Let m^{Θ_i} and $m_i^{\Omega_j}$ be the *bbas* corresponding to the frames of discernment Θ_i and Ω_j respectively. The m^{Θ_i} represent the degree of belief with regard to the accuracy of each data source s_i. In the present paper – to simplify the study – we suppose that the m^{Θ_i} are given, such that each m^{Θ_i} verifies the condition presented in equation (1).

In the previous example, we suppose we have the following *bbas*:

$$(0,0.1,0.2,0.7)^{\Theta_1},\ (0,0.2,0.3,0.5)^{\Theta_2} \text{ and } (0,0.5,0.1,0.4)^{\Theta_3},$$

where the quadruplet $(\mathrm{a,b,c,d}) = (m^{\Theta_i}(\emptyset), m^{\Theta_i}(A_i), m^{\Theta_i}(\overline{A}_i), m^{\Theta_i}(A_i \cup \overline{A}_i))$.

On the other hand, the $m_i^{\Omega_j}$ describe the trustworthiness of the facts f_j. Here, the index *i* means that the source s_i is considered an expert which provides opinions – 'degrees of belief' – for each fact f_j. Thus, each f_j has $|S|$ *bbas*.

We propose the following definition to quantify the $m_i^{\Omega_j}$:

$$\begin{cases} m_i^{\Omega_j}(T_j) = Sim(f_j, Fact(s_i))m^{\Theta_i}(A_i) \\ m_i^{\Omega_j}(\overline{T}_j) = (1 - Sim(f_j, Fact(s_i)))m^{\Theta_i}(A_i) + Sim(f_j, Fact(s_i))m^{\Theta_i}(\overline{A}_i) \\ m_i^{\Omega_j}(T_j \cup \overline{T}_j) = (1 - Sim(f_j, Fact(s_i)))m^{\Theta_i}(\overline{A}_i) + m^{\Theta_i}(A_i \cup \overline{A}_i) \end{cases} \tag{7}$$

where the function $Fact(s_i)$ returns the fact f_k that is provided by the source s_i.

[19] L. Blanco et al., op.cit.; A.P. Dempster, op.cit.; X.L. Dong, L. Berti-Equille, D. Srivastava, *Integrating…*, op.cit.; X.L. Dong, L. Berti-Equille, D. Srivastava, *Truth discovery…* op.cit.; X. Li et al.; X.L. Dong, B. Saha, D. Srivastava, *Less is more: Selecting sources wisely for integration*, Proceedings of the VLDB 2013 vol. 6, no. 2; Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, J. Han, *Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation*, Proceedings of the 2014 SIGMOD Conference; X. Yin, J. Han, P.S. Yu, *Truth discovery with multiple conflicting information providers on the web* "SIGKDD" 2007; X. Yin, W. Tan, *Semi-supervised truth discovery*, Proceedings from the WWW Conference 2011, pp. 217–226.

Our proposed *bba* has the following basic principles:

- The proposed *bba* exploits the accuracy of the source to assess the trustworthiness of the facts. Thus, if a fact is provided by accurate sources, then its trustworthiness will be higher.
- The use of similarity function allows two similar facts to support each other.
- The similarity function is type-specific i.e. it can handle different data types.
- If two facts f_1 and f_2 are conflicting, then the inaccuracy of the source s_1 that provides f_1 does not support the trustworthiness f_2, but it supports the total ignorance. In fact, this is the most important property of our model.

  In our example, we have: $sim(f_1, f_2) = 0$. Then we obtain the following result:
  - $(0,0.1,0.2,0.7)_1^{\Omega_1}$ $(0,0.2,0.3,0.5)_2^{\Omega_1}$ $(0,0,0.5,0.5)_3^{\Omega_1}$.
  - $(0,0,0.1,0.9)_1^{\Omega_2}$ $(0,0,0.2,0.8)_2^{\Omega_2}$ $(0,0.5,0.1,0.4)_3^{\Omega_2}$.

4.4. Evidence Combination

By using the Dempster's Rule of combination over the same frame of discernment Ω_j, we build new evidence representing the consensus of the evidence obtained from the disparate opinions of data sources.

For $|S|$ data sources, the combination of the $|S|$ *bbas* $m_1^{\Omega_j}, m_2^{\Omega_j}, ..., m_{|S|}^{\Omega_j}$ using equation (4) generates a new *bba* m^{Ω_j}. Since we have $|F|$ facts, then we obtain $|F|$ new *bbas* $m^{\Omega_1}, m^{\Omega_2}, ..., m^{\Omega_{|F|}}$:

By applying the combination rule in our example we obtain:

- $(0,\ 0.125,\ 0.66,\ 0.215)^{\Omega_1}$
- $(0,\ 0.42,\ 0.246,\ 0.334)^{\Omega_2}$

4.5. Decision-Making

With regard to each new *bba* m^{Ω_j}, we use equation (5) to calculate the pignistic transformation. This transformation allows us to generate the probabilities needed to make decisions, i.e. resolving the conflict and selecting the most accurate fact. Our decision-making procedure consists of the following steps:

- Firstly, we select for each $\Omega_j = \{T_j, \overline{T}_j\}$ the hypothesis $\hat{T}_j$ that has the highest pignistic probability.

$$\hat{T}_j = \arg\max_{H \in \Omega_j} BetP_j(\{\mathrm{H}\})$$

- Secondly, we reject all $\hat{T}_j$ where the $\overline{T}_j$ is selected, i.e. $\hat{T}_j = \overline{T}_j$. And we keep the ones where $\hat{T}_j = T_j$. We get then a set of N $(0 \le N \le |F|)$ facts.

- Thirdly, for some critical data integration systems, one must avoid the risk of making wrong decisions. Thus, a safe probability threshold p_{th} is established for the decision-making system. We use this threshold to re-select another set from the N filtered facts by removing all facts that have a pignistic probability less than the threshold ($BetP_j(T_j) < \mathrm{p}_{th}$). We obtain another set of M $(0 \leq M \leq N)$ trustworthy facts.
- Finally, we chose the appropriate reliable fact that has the highest pignistic probability.

  If we apply the decision-making step to our example, we obtain:
- $BetP_1(T_1) = 0.23 \quad BetP_1(\overline{T}_1) = 0.77$

  $\hat{T}_1 = \arg\max_{H \in \Omega_1} BetP_1(\{\mathrm{H}\}) = \overline{T}_1$ Then the fact $f_1 = "Gdansk"$ is untrustworthy and must be rejected.
- $BetP_2(T_2) = 0.59 \quad BetP_1(\overline{T}_1) = 0.413$

  $\hat{T}_2 = \arg\max_{H \in \Omega_2} BetP_2(\{\mathrm{H}\}) = T_2$ And since we do not specify a threshold p_{th}, then the fact $f_2 = "Warsaw"$ is trustworthy and must be consider as the correct fact for the attribute $a_1 = "capital"$.

5. Conclusion and future works

We have proposed in this paper a new evidence-based conflict resolution model. Our proposed model is based on the Dempster-Shafer theory of evidence, which is considered a generalization of the probability theory. Our model exploits the power of evidence theory in both the ability of handling uncertainty and imprecision and offering an adequate framework to combine multiple sources' opinions.

We believe that this work is a first step toward a generic and a flexible conflict resolution framework. In this regard, in our future work we will carry out the validation of our proposal with real-world data which will allow us to quantify the real benefit of the proposed methodology. Moreover, we intend to investigate other evidence-based conflict resolution models. Furthermore, we also plan to propose new possible extensions, such as an evidential estimation of the sources accuracy, and an evidential selection of the k-most relevant sources. This later extension aims to reduce the cost and maximize the accuracy of the provided data, especially in the context of big data integration.[20]

[20] X.L. Dong, B. Saha, D. Srivastava, *Less is more: Selecting sources wisely for integration*, Proceedings of the VLDB 2013 vol. 6, no. 2.

References

Bleiholder, J., Naumann, F., *Data fusion*, "ACM Computing Surveys" 2008, vol. 41, no. 1, pp. 1–41.

Blanco L., Crescenzi V., Merialdo P., Papotti P., *Probabilistic models to reconcile complex data from inaccurate data sources*, Conference on Advanced Information Systems Engineering 2010, pp. 83–97.

Dempster A.P., *Upper and Lower probabilities induced by a multivalued mapping*, "Annals of Mathematical Statistics" 1967, vol. 38, pp. 325–339.

Dong X.L., Berti-Equille L., Srivastava D., *Integrating conflicting data: The role of source dependence,* Proceedings of the VLDB 2009, vol. 2, no. 1, pp. 550–561.

Dong X.L., Berti-Equille L., Srivastava D., *Truth discovery and copying detection in a dynamic world*, Proceedings of the VLDB 2009, vol. 2, no. 1, p. 573.

Dong X.L., Saha B., Srivastava D., *Less is more: Selecting sources wisely for integration*, Proceedings of the VLDB 2013 vol. 6, no. 2.

Elouedi Z., Mellouli K., Smets P., *Assessing sensor reliability for multisensor data fusion within the transferable belief model,* "IEEE Transactions on Systems, Man, and Cybernetics" 2004, vol. 34, no. 1, pp. 782–787.

Fatehali M., *Building the business case for Master Data Management in the Public Sector*, "Oracle White Paper" 2011.

Kalampokis E., Tambouris E., Tarabanis K., *Open government data: A stage model*, "Electronic Government" 2011, Lecture Notes in Computer Science 6846, pp. 235–246.

Li Q., Li Y., Gao J., Zhao B., Fan W., Han J., *Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation*, Proceedings of the 2014 SIGMOD Conference.

Li X., Dong X.L., Lyons K.B., Meng W., Srivastava D., *Truth finding on the deep web: Is the problem solved?*, Proceedings of the VLDB 2013, vol. 6, no. 2.

Shafer G., *A mathematical theory of evidence*, Princeton University Press 1976.

Sentz K., Ferson S., *Combination of evidence in Dempster-Shafer theory*, SANDIA Technical Report 2002, SAND2002–0835.

Smets P., *Decision making in the TBM: the necessity of the pignistic transformation*, "International Journal of Approximate Reasoning" 2005, vol. 38, pp. 133–147.

Yin X., Han J., Yu P.S., *Truth discovery with multiple conflicting information providers on the web*, Knowledge and Data Engineering, IEEE Transactions on 20.6(2008), pp. 796–808.

Yin X., Tan W., *Semi-supervised truth discovery*, Proceedings from the WWW Conference 2011, pp. 217–226.

Ziemba E., Obłąk I., *The survey of information systems in public administration in Poland*, "Interdisciplinary Journal of Information, Knowledge and Management" 2011, vol. 9, pp. 31–56.

* * *

Nowa metoda rozwiązywania konfliktów danych w procesie integracji informacji bazująca na dowodach

Streszczenie

W dzisiejszych czasach, wraz ze wzrostem użycia danych w Internecie oraz publicznych rejestrach, dane tworzone są w coraz większej ilości zarówno przez maszyny, jak i przez ludzi. Z powodu tej eksplozji danych pozyskiwanie dokładnych informacji z wielu rozproszonych źródeł jest skomplikowane. Fuzja danych, zwana również rozwiązywaniem konfliktów (ang. *conflict resolution*), jest istotnym etapem w procesie integracji danych. Jej celem jest rozwiązywanie konfliktów pomiędzy sprzecznymi informacjami dotyczącymi tego samego rzeczywistego obiektu. W tym artykule przedstawiamy nową metodologię rozwiązywania tego problem, która wykorzystuje siłę teorii Dempstera–Shafera.

Słowa kluczowe: integracja danych, fuzja danych, rozwiązywanie konfliktów, teoria Dempstera–Shafera