A Framework to Integrate User Feedback for Rapid Conflict Resolution

Romila Pradhan
Department of Computer Science
Purdue University
West Lafayette, IN, USA
rpradhan@cs.purdue.edu

Siarhei Bykau
Bloomberg L.P.
New York, NY, USA
sbykau@bloomberg.net

Sunil Prabhakar
Department of Computer Science
Purdue University
West Lafayette, IN, USA
sunil@purdue.edu

***Abstract*—Data fusion addresses the problem of consolidating data from disparate information providers into a single unified interface. The different data sources often provide conflicting information for the same data item. Recently, several automated data fusion models have been proposed to resolve conflicts and identify correct data. Although quite effective, these data fusion models do not achieve a close-to-perfect accuracy.**

We present the demonstration of a system that leverages users as first-class citizens to confirm data conflicts and rapidly improve the effectiveness of fusion. This demonstration is built on solutions proposed in our previous work [1]. To utilize the user judiciously, our system presents claims in an order that is the *most beneficial* to effectiveness of fusion across data items. We describe ranking algorithms that are built on concepts from information theory and decision theory, and do not need access to ground truth. We describe the user input framework and demonstrate how conflict resolution can be expedited with minimal feedback from the user. We show that: (a) the framework can be easily adopted to existing data fusion models without any internal changes to the models, and (b) the framework can integrate both perfect and imperfect feedback from users.

I. Introduction

The collaborative web has proliferated the amount of data furnished by heterogeneous information providers. While the volume and variety of data have been growing at an unprecedented pace over the years, often there is little to no restraint over the quality of data available on the Internet. Instance heterogeneity highlights part of the data quality issue where distinct sources provide conflicting information for the same data item, e.g., financial firms publish different stock prices for the same company, sensors report conflicting measurements, on-line bookstores list different authors for identical books, etc. Data conflicts usually occur due to reasons such as inaccuracy of sources, copying between providers and failure to update the recent values. Resolving such conflicts is important since inaccurate information may result in unfavorable consequences such as a missed flight or financial losses.

Data integration systems consolidate multiple instances of the same real-world data item to present the user a single consistent record. Recently, a number of data fusion systems have been proposed (see [2] for a survey) that propose conflict resolution, i.e., distinguishing correct information from incorrect claims, as a way to integrate inconsistent data from multiple providers. Most of the existing data fusion systems propose fully automated solutions; as is evident from their low accuracies, these systems cannot be trusted to correctly identify true claims for all data items. Particularly for crucial data where it is imperative to distinguish correct from incorrect claims, we cannot solely rely on automated data fusion.

Users, on the other hand, are usually endowed with knowledge that, if leveraged properly, may benefit the conflict resolution process at different stages, e.g., knowing the correct claim for a data item may improve the accuracy of data fusion while knowing two claims are alternative representations of each other may remove disagreement among claims of a data item. In particular, knowledge of the true claim for a data item will correct our belief in sources that vote for the data item, which in turn will impact the correctness probabilities of other claims and may immediately enhance the accuracy of the data fusion model. This highlights the increasing need for a system that can combine the best of both: automatically resolving conflicts while efficiently involving the user to guide the conflict resolution process.

We propose to demonstrate a prototype implementation of our solution [1] to the problem of integrating user feedback toward rapid conflict resolution in data fusion. We present a system that solicits feedback from users in the form of validation of a data item, i.e., we ask the user to provide the correct claim for the data item. Validation of claims per se is an expensive task – to guarantee effective conflict resolution, it assumes access to highly accurate feedback (e.g., from domain experts). Moreover, users often have a limited budget of questions that they can answer. Since data fusion models typically deal with a large number of claims, it is of utmost importance to validate data items in an order that is the *most beneficial* to the performance of fusion across all items.

The key novelty of our system is that it supports algorithms to rank data items such that maximum benefit is achieved with minimal user efforts. Furthermore, the ranking algorithms do not depend on ground truth. To this end, we leverage the distribution of source votes toward claims of data items, the output of data fusion models, and the concept of *value of perfect information* (VPI) from decision theory to estimate the benefit of validating a data item. The objective is to rapidly improve the effectiveness of conflict resolution techniques and minimize user efforts. We demonstrate that our system: (a)

ID	Data Item	S_1	S_2	S_3	S_4
O_1	Zootopia		Howard*	Spencer	Spencer
O_2	Kung Fu Panda	Stevenson*		Nelson	
O_3	Inside Out		leFauve	Docter*	
O_4	Finding Dory				Stanton*
O_5	Minions	Coffin*	Renaud		
O_6	Rio	Jones		Saldanha*	

(a) **Database instance.**

ID	Probabilities of Claims
O_1	Howard (0), Spencer (1)
O_2	Stevenson (0.015), Nelson (0.985)
O_3	Docter (0.999), leFauve (0.001)
O_4	Stanton (1)
O_5	Coffin (0.921), Renaud (0.079)
O_6	Saldanha (0.985), Jones (0.015)

(b) **Output of fusion.**

TABLE I: A motivating example. Table Ia shows four sources providing information about directors of six movies. Correct claims are marked with a (*). Table Ib shows the correctness probabilities of claims as output by data fusion model ACCU [3]. For each data item, claim that has the highest probability of being true is considered correct.

can be easily adopted to existing data fusion models since the ranking algorithms are applicable to the generic output of data fusion, and (b) is fairly tolerant to user feedback errors. The system is demonstrated on a real-world dataset that is most commonly used in recent data fusion models. The prime motivation for this work has been the possibility of automated data fusion systems incorrectly judging false claims to be true – a reasonable concern in the present date where false claims are rampant on the Internet.

II. Motivating Example

Consider an example of websites providing information on directors of animation movies (Table I). Data fusion systems take the conflicting claims as input, and output the correctness of each claim (as an example, see Table Ib).

Source S_2 provides Howard as director for the movie Zootopia whereas sources S_3 and S_4 claim it to be Spencer. A data fusion system (e.g., ACCU [3]) that outputs Spencer to be the true claim of Zootopia can benefit from the validation that Howard is instead correct. With this knowledge, the fusion system can reconsider the claims provided by sources S_2, S_3 and S_4, and improve its output on other data items.

Validation of claims by users (domain experts, more so) is expensive because users usually have a limited budget. To utilize users judiciously, claims should be presented for validation in an order that is most beneficial to the effectiveness of fusion. The key question we are answering is: *Assuming we can validate any data item, and know which of its claims is correct, which item should we select for validation?*

The task of identifying the *best* data item for validation is challenging because we have to deal with a number of issues. Data fusion typically deals with a large number of claims (hundreds of thousands), thus limiting the ability to ask questions on a *very* small fraction of all claims. We, therefore, need a way to determine the data item best suited for validation. To that end, we need to quantify the definition of 'best', i.e., what is the basis for deciding whether or not one data item is more suitable for validation than another? Since we do not posses ground truth, to answer this question, we need to design heuristics that are independent of ground truth. Furthermore, the validation of each claim may potentially influence that of any other claim; the exhaustive computation of estimating the impact of validating each data item by re-running fusion is prohibitively expensive.

We make the following two observations as a step toward solutions. First, data items have different *levels of uncertainty* because of the consensus (or the lack thereof) among sources on claims. One may expect that validating Minions would be more advantageous than Zootopia because S_1 and S_2 disagree on the former whereas two of the three sources that vote for the latter agree on a common claim. This is because we expect to learn more from the validation of data items with *disagreement*. Second, although a data item may have conflict over its values, validating it may not be beneficial. For instance, validating Finding Dory would influence S_4 and that would have an effect only on Zootopia whereas validating Zootopia would potentially impact all other items.

III. System Description

Figure 1 represents the overall design of and steps involved in the user feedback integration framework. In the following, we describe the input and components of the system.

A. System Input

Below are the primary inputs to our system:

- **Database ($\mathcal{D}$)**: The first input to the system is a database instance $\mathcal{D}$, which is a collection of data items $\mathcal{O}$, set of data sources $\mathcal{S}$, set of distinct claims per data item $V = \{V_1, \ldots, V_{|\mathcal{O}|}\}$ and set of observations ψ provided by sources in $\mathcal{S}$ toward data items in $\mathcal{O}$. In other words, $\mathcal{D}$ can be represented as a tuple of its constituents, i.e.,
$$\mathcal{D} = \langle \mathcal{O}, \mathcal{S}, \psi, V \rangle$$
- **Data Fusion Model ($\mathcal{F}$)**: The second input to the system is a data fusion model $\mathcal{F}$ that takes a database instance $\mathcal{D}$ and outputs a set of probability assignments P for claims of data items in $\mathcal{D}$, i.e.,
$$\mathcal{F} : \mathcal{D} \rightarrow \langle P \rangle$$
where for each data item $O_i \in \mathcal{O}$, $P(v_i^k) \in [0, 1]$ is the correctness probability of claim $v_i^k \in V_i$.

B. Overview of System Components

As shown in Figure 1, the feedback integration framework is composed of three primary units: a data fusion model, a module for ranking algorithms, and a user input module. Arrows in the figure represent direction of the process flow.

Given a database instance, the data fusion model first generates correctness probabilities of claims even before the process of validation begins. Based on these probabilities and

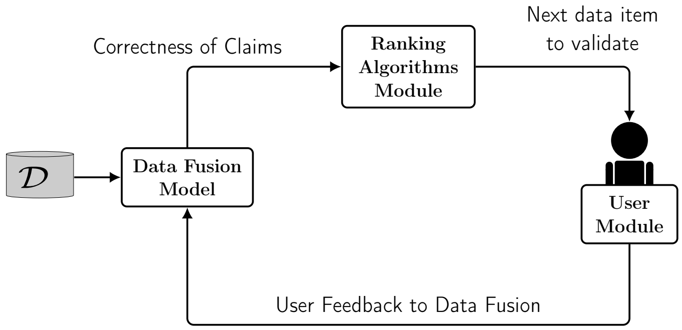


Fig. 1: The proposed user feedback framework.

the database instance, the ranking algorithms module presents the next data item for the user to validate. The user, depending upon their skill level, submits the correct claim for the selected data item. The feedback is then fed into the data fusion system in the form of labeled truth. Using the acquired ground truth labels, the data fusion model updates correctness probabilities of claims. This is a continual feedback process that continues as long as there are unvalidated data items and the user is available to provide feedback.

Data Fusion Model. The first step in the process is selection of a data fusion model. Most of the existing data fusion models generate correctness probabilities of claims and source quality metrics as an intermediate step. Subsequently, the fusion models identify correct claims of data items based on these probabilities. We support data fusion models that produce such intermediate correctness probabilities. Although the choice of data fusion model in subsequent rounds is independent of the data item selected for validation, (i.e., once an item is validated depending on the output from a data fusion model, we can fuse the data using another model), for the sake of this demonstration, we stick to being consistent in the data fusion model in further validations.

Ranking Algorithms. Once the correctness probabilities of claims are output, the role of ranking module is to determine the *single best* data item, validating which would prove the *most beneficial* to performance of the data fusion model over all data items. To that end, the ranking module utilizes the distribution of votes by data sources and the output of data fusion. Algorithms in the ranking module are built on concepts of entropy (from information theory) and value of perfect information (from decision theory). Since ground truth is not always available for all data items, the ranking module is comprised of techniques that do not depend on ground truth. Our ranking algorithms can be broadly defined in terms of two categories: (a) item-level, (b) holistic. While the item-level algorithms evaluate data items based on distribution of votes or output of fusion specific to each data item, the holistic approach assesses the global impact of a validation, i.e., the effect of validating an item on the output of other data items. The latter ranking algorithms may involve an all-pairs impact computation between data items and, therefore, may not always be scalable. To address this issue, our ranking module supports approximation algorithms (specific to the data fusion model) that can be scaled up to large datasets. For more details on the ranking algorithms, please refer to [1].

User Input Module. We solicit feedback from users in the form of validation of a data item, i.e., we ask them to provide us the correct label (true/false) for claims of a data item. Feedback labels from the user are input as ground truth to the data fusion model and allow updating its estimates on source quality metrics and correctness probabilities of claims. However, we realize that domain experts are expensive and may not be always available for providing feedback. Our user input module, therefore, also supports imperfect feedback from users in various forms, such as from a single erroneous user or from a crowd of workers. Given a user's profile, i.e., our system can leverage their historical success/error rate during integration of feedback received from them into the data fusion model. We are also able to incorporate a user's explicit confidence in the labels for claims of data items. Besides, our system is also designed to integrate crowdsourced input from a crowd of workers – we consider as feedback the correctness probabilities of claims as generated by a crowdsourcing platform. Through the incorporation of crowd input, we show that our system can also be integrated with existing crowdsourcing platforms without making any internal changes to their models. To incorporate the different kinds of users, our system assumes knowledge of the type of user interacting with the system.

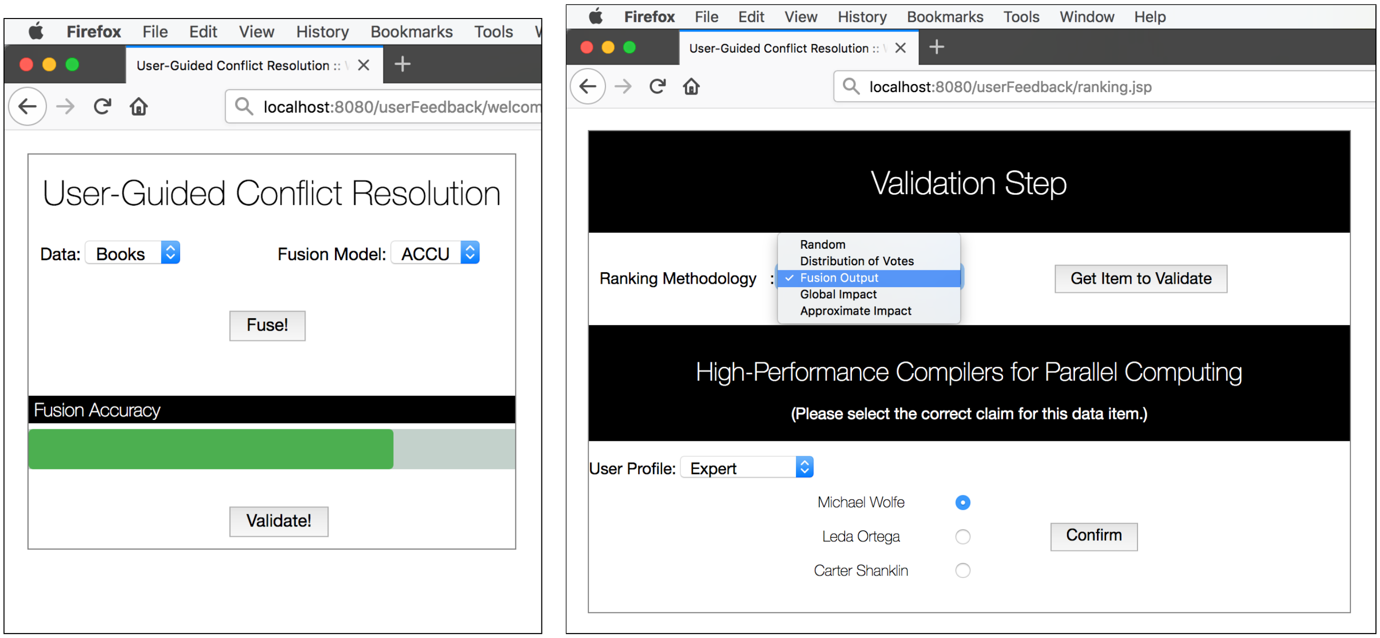
IV. Demonstration Overview

In the demonstration we will use the Books dataset [4] where we simulate user input for data items by treating the silver standard provided in [4] as ground truth.

We will show a working prototype of our solutions and show the applicability of our algorithms to existing data fusion models. Our prototype provides an interactive interface for users to continually interact with the data fusion model. Once the system is launched, the user will have the option of specifying a data fusion model from a list of available models, and can specify their budget, i.e., how many data items are they willing to validate. For the sake of this demonstration, we would deploy the solutions on top of ACCU [3], a data fusion model that uses Bayesian analysis to iteratively estimate source accuracies and claim correctness probabilities.

The ranking algorithms module, by default, presents to users the data item selected by the holistic ranking approaches, which have been shown in [1] to exhibit the best effectiveness among all ranking algorithms. Validation of the selected data item can be done by users in the role of: (a) a domain expert (who knows the correct label), (b) an imperfect user (who is only $x\%$ certain of a label or has a known historical error-rate of $y\%$), or (c) collation of crowdsourced labels in the form of correctness probabilities of claims. The acquired feedback is then input back to the data fusion model (as shown in Figure 1) for updating its estimates of correctness probabilities of claims.

Throughout the feedback process, the user will be able to monitor the benefit of a validation through a progress bar that displays the percentage improvement in accuracy of the



(a) User has the option of selecting data fusion model. Green bar at the bottom shows current accuracy of the fusion model.

(b) Users can specify the method to determine the next data item for validation and also let the system know whether input is from an expert, an average user or from a crowd of workers.

Fig. 2: GUI for integrating user feedback with the data fusion model.

data fusion model. After validation of the selected data item, the user will be able to immediately see the impact of their feedback on the accuracy of fusion.

At the end of feedback solicitation, we display to users the individual improvement in accuracies of fusion as achieved by the different ranking algorithms, thus enabling the user to compare and contrast the algorithms at run-time.

V. Related Work

The problem of conflict resolution has been extensively studied in recent years (see [2] for a survey) where data fusion models jointly estimate correctness of claims and source reliabilities. It has been shown in [5] that even a very small amount of ground truth can help improve the accuracy of data fusion by identifying reliable data sources. To the best of our knowledge, none of these efforts are directed toward actively acquiring and learning from ground truth data to expedite the process of conflict resolution.

Concepts from decision theory and active learning have previously been used for pay-as-you-go user feedback in various data management problems [6] where the foremost concern lies in determining the sequence in which human input is received. Our system is built upon an entropy-based utility function that uses the value of perfect information [7] to narrow down user preferences under uncertainty.

We present a framework that can be integrated with existing data fusion algorithms that generate correctness probabilities of claims. In the presence of noisy labels, our framework builds upon ongoing research in crowdsourcing [8] where instead of a domain expert, feedback is sought from a crowd of workers. Our framework, however, is orthogonal to crowdsourcing since we are not interested in modeling crowd workers, any of the existing crowdsourcing approaches can be used to obtain the most accurate label for data items and plugged into the feedback framework.

VI. Conclusion

We presented the demonstration of a system that addresses the problem of effectively soliciting feedback from users to resolve conflicts, and improve the performance of existing data fusion techniques. To the best of our knowledge, this problem has not been identified in earlier work. Our solution makes it possible to leverage user feedback in existing fusion models without making any changes to the internals of the models. We demonstrate that our solutions to determine the best data item for validation do not depend on ground truth and are applicable to a wide range of fusion models. We further demonstrate that our feedback framework is able to integrate imperfect user input to effectively improve the performance of fusion.

References

[1] R. Pradhan, S. Bykau, and S. Prabhakar, "Staging user feedback toward rapid conflict resolution in data fusion," in *SIGMOD*, 2017.
[2] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *SIGKDD Explorations Newsletter*, 2016.
[3] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Data fusion: Resolving conflicts from multiple sources," in *WAIM*, 2013.
[4] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava, "Truth finding on the deep web: Is the problem solved?" *PVLDB*, 2012.
[5] X. Yin and W. Tan, "Semi-supervised truth discovery," in *WWW*, 2011.
[6] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, "Pay-as-you-go user feedback for dataspace systems," in *SIGMOD*, 2008.
[7] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Prentice Hall, 2003.
[8] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin, "Crowddb: Answering queries with crowdsourcing," in *SIGMOD*, 2011.