Tobias Meuser, Patrick Lieser, The An Binh Nguyen, Doreen Böhnstedt and Ralf Steinmetz. *Adaptive Information Aggregation for Application-specific Demands*. In Proc. BalkanCom 2017

Adaptive Information Aggregation for Application-specific Demands

Tobias Meuser, Patrick Lieser, The An Binh Nguyen, Doreen Böhnstedt and Ralf Steinmetz
Multimedia Communications Lab (KOM), Technische Universität Darmstadt, Darmstadt, Germany
Email: {firstName.lastName}@KOM.tu-darmstadt.de

***Abstract*—In distributed sensing systems information is measured by various types of sensors. Due to measuring inaccuracy, each measurement only provides a noisy view on the measured variable. This inaccuracy can further increase if other variables are derived from those measurements. Nodes make wrong decisions based on the inaccurate information which they are provided with. Consequently, the correctness of measurements and derived information is crucial.**

In this paper, we focus on the accuracy of information w. r. t. the correctness of provided information. We introduce an accuracy metric based on application requirements. This metric can be used to determine if the provided information satisfies the requirements of an application. It requires a data representation that assumes each sensor measurement can be depicted with a distribution vector. This representation contains all information available and can be used to track accuracy while aggregating and fusing information. We propose an approach to model the spatiotemporal changes of the measured probability vector. As a result past and geographically distant information can be used to enhance the accuracy of information. The evaluation results confirm that using the past and present data can increase the accuracy by up to 200% than when these data are absent.

I. Introduction

In distributed sensing systems various information is provided by various types of sensors. Measured information, however, is generally a noisy observation of the measured variable. This noise is dependent on the quality of the available sensors.

For some applications, the accuracy of the provided information might not be sufficient. Common examples like Advanced Driver Assistance Systems (ADAS) in vehicular networks require information of high accuracy.

The accuracy of a measurement can be increased by aggregating multiple independent sensor measurements. Information aggregation and fusion have been heavily investigated in the field of Wireless Sensor Networks (WSN). According to Durrant-Whyte [1], information fusion can be divided into three categories: complementary, redundant and cooperative fusion. In this work, we focus on redundant aggregation. The idea of redundant aggregation is to aggregate similar or almost similar messages to reduce network load while increasing accuracy and reliability [2]. The processing and aggregation of messages can be performed by intermediate nodes as in [3].

Although our developed approach can be used in most distributed sensing systems, we utilize the vehicular scenario to elaborate our concept. Accuracy and reliability are important in vehicular networks, as wrong or inaccurate information may reduce driver safety and comfort.

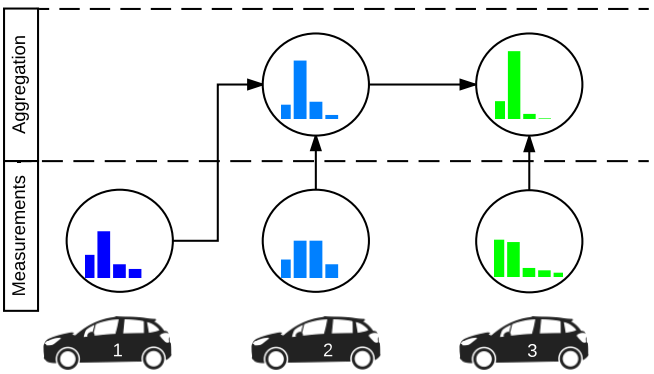

Figure 1: In-Network aggregation process

In a vehicular network, for example, each vehicle senses the temperature. In most cases calculating the average of the sensed values leads to more reliable results. However, as vehicles are equipped with sensors of different quality, the sensor accuracy is not the same for all retrieved values. Taking the sensor accuracy into account, this result can be enhanced further. Therefore each sensor does not only measure the exact value but provides a distribution vector based on the sensor properties. This distribution vector can be produced by performing multiple measurements in a certain time interval or by using knowledge of the sensor accuracy. In the case of temperature sensors, the distribution vector will be based on a Gaussian distribution. Therefore a probability vector can be created using a Gaussian distribution with the measured value and the sensor deviation. The resulting vector can be aggregate with other measurements to increase accuracy.

An exemplary scenario is shown in figure 1. In this example, the observed variable has five possible states. In our example, vehicle 1 is deployed with a sensor of moderate accuracy. To allow the succeeding vehicles to enhance their view on the variable, the measured probability vector is shared. The probability vector is dependent on several factors, one of which is sensor quality. Using only its own measurements, vehicle 2 would not be able to decide in which state the observed variable is, as the probabilities for state 2 and 3 are equal. However, using the information provided by vehicle 1, vehicle 2 is able to decide about the state of the observed variable in a cooperative manner. Simultaneously, it increases the accuracy compared to

the measurement of vehicle 1. The resulting probability vector is shared with vehicle 3. Vehicle 3 by itself would even have made the wrong decision, as the probability for state 1 is higher than the probability for state 2, but state 1 has been barred from vehicle 1 and 2. As vehicle 3 receives the information from vehicle 2, it can correct its inaccurate measurement.

Our contribution is the modeling of the measurement process including the spatiotemporal influences on measurements. Consequently past and geographically distant measurements can be used to increase the measurement quality.

II. Related Work

Data aggregation is a pivotal technique in distributed sensing systems. Simple aggregation algorithms are among others maximum and mean. Aggregating information using these algorithms can decrease the accuracy of information. For redundant aggregation, different approaches have been developed.

The Kalman filter proposed by Kalman in 1960 was one of the first approaches to increase accuracy by combining redundant information [4]. Since its proposal, different algorithms have been developed on this basis. The idea of Kalman filter is to reduce Gaussian noise on low-level sensor data. Some extensions like the Extended Kalman Filter [5] and the Unscented Kalman Filter [6] have been proposed to increase its performance in different scenarios. Moreover, Kalman filter has been extended for the usage in distributed scenarios [7].

To handle information that does not necessarily match the requirements for the usage of Kalman filter, other approaches have been presented. Xiao et al. [8] introduced an aggregation scheme for constant variables. The proposed distributed aggregation scheme is based on maximum-likelihood parameter estimation. Each node in the network holds a local estimate of the global state. By communicating with neighbor nodes, the global estimate is synchronized between the nodes.

Boulis et al. introduced an aggregation algorithm based on a probability distribution, which provided a tradeoff between energy efficiency and accuracy [9]. The idea is to keep meta-information of the aggregation to improve future aggregations. That information is shared with neighbor nodes. Using that information they were able to merge information without a central node. The aggregation itself is dependent on the used aggregation function.

The authors only investigate the issues of "snapshot aggregation". Compared to that, we utilize past and geographically distant measurements to further increase the performance of the aggregation algorithm. We moreover do not focus on one specific aggregation function. The used data format contains all necessary information to extract information using an arbitrary aggregation function.

III. Data Modeling

We assume a variable has a set of possible states S. Without loss of generality, the number of possible states S is finite. At a given time and location, the measured variable is in a certain state $s \in S$. The state of the measured variable cannot be determined with certainty from a single measurement.

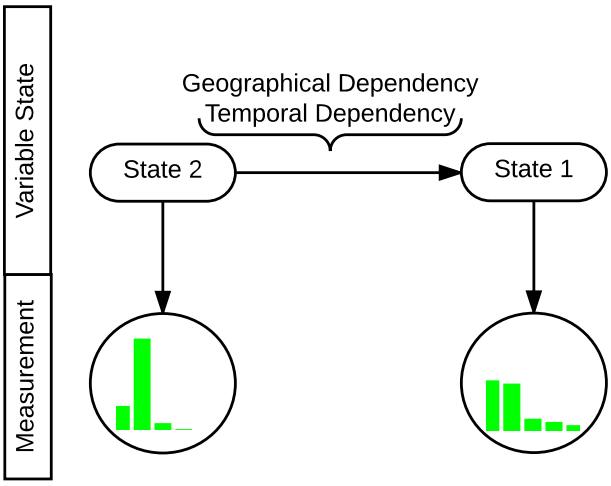

Figure 2: Modeling of the measurement process in the vehicular scenario

Data in information systems is often modeled as a tuple [10] [11]. This tuple contains besides the measured value the location and time of the measurement. However, an important characteristic of the data is missing in this definition, i. e. the accuracy of the measurements. While the actual data value is sufficient for most use cases, data fusion and other operations could utilize the distribution of the measured information.

We model this system as displayed in figure 2. The measured variable is a partially observable variable, whose state can be approximated only with measurements. Each measurement is only a noisy view on the measured variable. Due to sensor inaccuracy, after a certain amount of measurements, a probability vector can be determined for the current variable state. Each probability vector is the output of one measurement process. Using this vector, the state of the variable can be estimated.

To provide additional information about the distribution to other nodes, we do not store the measured value, but the probability vector $\vec{p}$. This data representation was initially proposed by Boulis et al. [9]. As the number of possible states in S is finite, the vector dimension is likewise. The vector is shown in equation 1.

$$\vec{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} \tag{1}$$

For each probability vector the following requirements must hold:

$$p_i \geq 0$$

$$\sum_{i=1}^{n} (p_i) = 1$$

The probability vector for each measurement is dependent on the sensor performing the measurement. Before sharing,

additional meta-information is added to the vector. The final packet consists of the following five entities:

1) *Data ID*: The data ID of the information. This is important for later processing based on this information.
2) *Vector-based Representation*: The measured or processed information value
3) *Location & Time*: The location and time for which the information has been created. Using this information the freshness of the information and the vehicle's distance to it can be determined.
4) *Contributing Vehicles*: The id of each vehicle that has contributed to this information. This meta-information is important for the aggregation process, as the aggregation only works correctly if all measurements are independent. Storing the vehicle ids is a simplification to prevent the same measurement being aggregated.

In the following, the influence of the sensor and spatiotemporal dependencies are elaborated on.

A. Sensor Accuracy

In this section, we focus on the probability vector. The accuracy of the measuring sensor has a high impact on the resulting probability vector. Sensors with low accuracy will produce vectors without high peaks. Therefore the standard deviation is high. On the contrary sensors with high accuracy will produce a vector with few high peaks.

This representation of data requires more information compared to transmitting only the value determined by the sensor. However, the applications on the vehicle are enabled to perform operations while having the insight of how accurate the provided information is.

B. Temporal Dependency

Due to the changes of the measured variable, the probability vector changes over time. The more time has elapsed since a measurement, the less accurate it becomes as the measured variable has its specific behavior. This behavior is modeled to utilize previous measurements in the aggregation process. Thus those can be used to increase the accuracy of current measurements. In most cases only information of one past period is available. Therefore a Markov chain is used to model the fluctuation of the measured variable. It does not describe the fluctuation of the actual sensor measurements, but the changes of the variable itself.

The transition matrix T_t of the measured variable over time has the same dimension as the probability vector. It is shown in equation 2.

$$T_t = \begin{pmatrix} t_{11} & \dots & t_{n1} \\ \vdots & \ddots & \vdots \\ t_{n1} & \dots & t_{nn} \end{pmatrix} \tag{2}$$

In order to predict the state of the measured variable in a future period, the transition matrix is multiplied with the measured probability vector. Assuming a probability vector is available in period t_0, the state of the variable in the future state t can be predicted as shown in equation 3.

$$\vec{p}_t = T_t^{t-t_0} * \vec{p}_{t_0} \tag{3}$$

With increasing amount of predicted periods, the prediction becomes less accurate. However, this inaccurate value can still be used to enhance the accuracy of currently measured information.

C. Geographical Dependency

The geographical dependency can be modeled similarly to the temporal dependency. Instead of using the temporal transition matrix T_t, the transition matrix T_d is used to model the geographical relation.

The geographical relation is dependent on the distance, which itself is dependent on the event type though.

For events that are not bound to streets like temperature, the geographical relation is only dependent on the linear distance. However, for events that are map-based, this relation is not valid.

In this case, the length of the shortest path between the current location and the information's location should be used. Moreover most information entities are direction dependent. In this case, the shortest path for a vehicle turn can be used to approximate the changes.

As the transition matrix is information type dependent, it can be created using the specific distance function of an information type. The resulting transition matrix has similar requirements as the temporal relation matrix and is shown in equation 4.

$$T_d = \begin{pmatrix} d_{11} & \dots & d_{n1} \\ \vdots & \ddots & \vdots \\ d_{n1} & \dots & d_{nn} \end{pmatrix} \tag{4}$$

Using this matrix the current probability vector can be estimated using the information type dependent distance function $d(loc, loc_0)$. The result of the distance function is integral. This is achieved by defining the transition matrix for adequately small distances. Equation 5 shows the geographical adjustment performed on an information entity.

$$\vec{p}_t = T_d^{d(loc,loc_0)} * \vec{p}_{t_0} \tag{5}$$

D. Spatio-Temporal Dependency

The temporal and the geographical dependency are combined to a spatiotemporal dependency utilizing both transition matrices. This is possible as the temporal and the geographical dimension are independent of each other. The resulting probability vector calculation is shown in equation 6.

$$\vec{p}_t = T_t^{t-t_0} * T_g^{d(loc,loc_0)} * \vec{p}_{t_0} \tag{6}$$

IV. Accuracy

There is no uniform definition of accuracy in the literature. It is a term that is hard to describe. Wang et al. [12] define accuracy as the degree of which the data is error-free, accurate and reliable. Batini et al. [13] define accuracy as the "closeness between a value v and a value v'" and differentiate between the syntactic accuracy and the semantic accuracy. While syntactic accuracy only checks if a value is a valid entry in terms of possible values, semantic accuracy compares the measured value with the real value. Dependent on the used comparison function, different values might be accurate.

In this work, we want to introduce the concept of the application dependent accuracy. Some applications like navigation can handle imprecise information very well. Besides accurate information safety applications require the degree of accuracy of the provided information.

Assuming each application knows the degree of inaccuracy it can handle, information can be validated easily. In the case of insufficient accuracy, the application might request additional measurements to increase the accuracy.

Therefore the accuracy metric is defined as the probability that the state of the observed variable is within a certain range of states. This range might be either static or dynamic dependent on the current state. The accuracy is dependent on the expected state of the observed variable s_e. The expected state $s_e \in S$ maximizes the accuracy for the specific application.

Equation 7 displays the accuracy metric a for applications with static maximum error f. The function $\mathrm{index}(s \in S)$ returns the index of the state s in the probability vector $\vec{p}$.

$$a = \max_{s \in S} \left[\sum_{i=\mathrm{index}(s)-f}^{\mathrm{index}(s)+f} (p_i) \right] \tag{7}$$

For an application with state-dependent error, the static variable f is replaced by a state dependent error-function $f(s)$. Therefore equation 7 is adjusted to equation 8.

$$a = \max_{s \in S} \left[\sum_{i=\mathrm{index}(s)-f(s)}^{\mathrm{index}(s)+f(s)} (p_i) \right] \tag{8}$$

Equation 8 can be used to represent both static and state-dependent error-tolerance of applications. As described above, the expected state is the state $s_e \in S$ for which the accuracy is maximized.

V. Data Aggregation

The main advantage of the data representation presented in chapter III is the possibility to aggregate values in order to increase accuracy. Different to other approaches, this model is naturally able to include past values and values of different positions into the estimation.

Given a past probability vector $\vec{p}_{t_0}$ for the state of the observed variable at t_0 and a measured probability vector $\vec{q}_t$ at t while both being at the same location, the aggregation process works as follows:

First, a probability vector $\vec{p}_t$ for the period t is calculated by multiplying the transition matrix of the observed variable with the probability vector $\vec{p}_{t_0}$ as shown in equation 9.

$$\vec{p}_t = T_t^{t-t_0} * \vec{p}_{t_0} \tag{9}$$

As both probability vectors now describe the state of the variable in the same period, those can be aggregated. The aggregation process is based on the fact that both probability vectors describe the state of the same variable. The probability for each state in the aggregation is calculated using the conditional probability that both vehicles measure the state s given both measurements measure the same state. This calculation is shown in equation 10.

$$P\left(X = s\right) = P\left(X_p = s \cap X_q = s \,|X_p = X_q\right) \tag{10}$$

Therefore, the new probability vector $\vec{r}$ can be calculated with equation 11.

$$\vec{r} = \begin{pmatrix} r_0 \\ \vdots \\ r_n \end{pmatrix} = \frac{\begin{pmatrix} p_0 * q_0 \\ \vdots \\ p_n * q_n \end{pmatrix}}{\vec{p}_t \cdot \vec{q}_t} \tag{11}$$

To complete the aggregation process, the meta-data needs to be adjusted. For timestamp and location of the aggregated information, the latest timestamp and location is used.

The aggregated vector has a higher accuracy than the original one. Dependent on the dimension of the vector, it might be necessary to compress it for sharing with other vehicles. This can be done by compression algorithms, as most elements are expected to be 0. Moreover, it might be possible to represent the vector using well-known functions like the Gaussian function, as mentioned in [9]. In this case, standard deviation and average are calculated.

VI. Evaluation

In the following, we evaluate the performance of the proposed mechanism for different scenarios. Figure 3 shows the maximum achievable accuracy for different scenarios. A rate of change of 20% states that the probability of the measured variable to keep its state is 80%, while the probability of a state change is 20%. A rate of change of 80% corresponds to a system without the usage of past information. There are no outliers in that scenario.

It is depicted that maximum accuracy is dependent on the rate of change and the rate of incoming measurements. The higher the rate of change is, the more measurements per time interval (TI) are required to achieve a certain accuracy. For static information, it is possible to achieve an accuracy of roughly 100% in all cases. However, in the vehicular scenario, static information is not common. For non-static information, the maximum achievable accuracy is lower due to the mechanism described in chapter III. With increasing amount of measurements per TI, the accuracy increases for

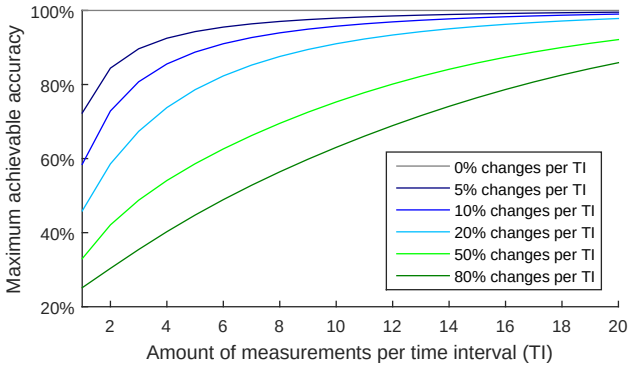

Figure 3: Maximum achievable accuracy dependent on the measurements per time interval (TI).

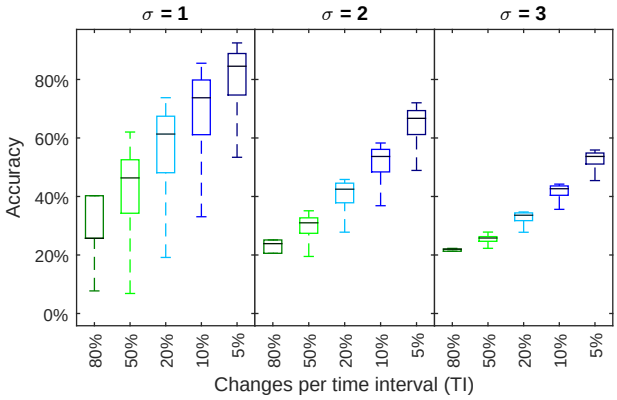

Figure 4: Behavior of the accuracy using a Guassian distributed variable.

all graphs. From this, it can be deduced that huge amounts of measurements per TI are required to achieve a high accuracy for a rapidly changing variable.

Figure 4 shows the accuracy distribution for a Gaussian distributed measurement. In the simulation, the average for each measurement has been chosen randomly based on the Gaussian distribution. Afterwards, the measurement vector itself is calculated based on the randomly chosen average and the standard deviation. The accuracy has been investigated for different rates of change and standard deviations. As in figure 3, a rate of change of 80% equals a system without usage of past information. The advantages of the proposed model can then be observed for the rates of change of less than 80%. With decreasing amount of changes per TI, the median of the accuracy increases. Moreover, the accuracy of a measurement with standard deviation of $\sigma = 1$ is higher than for $\sigma = 2$ and $\sigma = 3$, but simultaneously the confidence interval is bigger. This is caused by the influence that outliers have on the data. With increasing accuracy the influence increases.

VII. Conclusion

In this paper, we modeled the measurement process in the vehicular scenario. Instead of sharing only a single value, a probability vector is shared, which provides information about the sensor accuracy. Accuracy is important for various vehicular applications. Dependent on the application requirements, the amount of aggregated information can be adjusted. This leads to a smaller number of requested messages, as the information is only enhanced until it satisfies the accuracy constraints defined by the application.

The measured variables change over time and with increasing distance. To model this behavior, a Markov chain was used. By modeling the behavior of the measured variable, it is possible to enhance current and local measurements with past and distant data. The impact of temporal and geographical changes on the measured value is variable-dependent.

The evaluation shows that past information can be used to increase the accuracy of measurements by up to 200% dependent on the rate of change. In future work, we will enhance the proposed model with error handling to actively filter faulty information.

Acknowledgments

The work presented in this paper was partly funded by the LOEWE initiative (Hessen, Germany) within the NICER project.

References

[1] H. F. Durrant-Whyte, "Sensor models and multisensor integration," *The international journal of robotics research*, vol. 7, no. 6, pp. 97–113, 1988.

[2] E. F. Nakamura, A. A. Loureiro, and A. C. Frery, "Information fusion for wireless sensor networks: Methods, models, and classifications," *ACM Computing Surveys (CSUR)*, vol. 39, p. 9, 2007.

[3] S. Dietzel, J. Petit, F. Kargl, and B. Scheuermann, "In-network aggregation for vehicular ad hoc networks," *IEEE communications surveys & tutorials*, vol. 16, pp. 1909–1932, 2014.

[4] R. E. Kalman *et al.*, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, pp. 35–45, 1960.

[5] G. Welch and G. Bishop, "An introduction to the kalman filter," 1995.

[6] S. J. Julier and J. K. Uhlmann, "New extension of the kalman filter to nonlinear systems," in *AeroSense'97*. International Society for Optics and Photonics, 1997, pp. 182–193.

[7] R. Olfati-Saber, "Distributed kalman filtering for sensor networks," in *Decision and Control, 2007 46th IEEE Conference on*. IEEE, 2007, pp. 5492–5498.

[8] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *Proceedings of the 4th international symposium on Information processing in sensor networks*. IEEE Press, 2005, pp. 63–70.

[9] A. Boulis, S. Ganeriwal, and M. B. Srivastava, "Aggregation in sensor networks: an energy–accuracy trade-off," *Ad hoc networks*, vol. 1, pp. 317–331, 2003.

[10] G. V. Moustakides and V. S. Verykios, "Optimal Stopping," *Journal of Data and Information Quality*, vol. 1, pp. 1–34, 2009.

[11] R. Blake and P. Mangiameli, "The Effects and Interactions of Data Quality and Problem Complexity on Classification," *Journal of Data and Information Quality*, vol. 2, pp. 1–28, 2011.

[12] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, vol. 12, pp. 5–33, 1996.

[13] C. Batini and M. Scannapieco, "Data quality dimensions," in *Data and Information Quality*. Springer, 2016, pp. 21–51.