

Extraction de connaissances dans les réseaux ad hoc inter-véhicules

Bruno Defude
TELECOM & Management
SudParis
9 rue Charles Fourier
91011 EVRY Cedex - France
Bruno.Defude@int-
edu.eu

Thierry Delot
LAMIH - Université de
Valenciennes
Le Mont Houy
59313 Valenciennes - France
tdelot@univ-
valenciennes.fr

Jose-Luis Zechinelli
Martini
Universidad de las Américas
Puebla - Mexique
joseluis.zechinelli@udlap.mx

Nicolas Cenerario
LAMIH - Université de
Valenciennes
Le Mont Houy
59313 Valenciennes - France
ncenerar@univ-
valenciennes.fr

Sergio Ilarri
IIS Dept, University of
Zaragoza
Maria de Luna 1
Zaragoza, 50018 - Espagne
silarri@unizar.es

ABSTRACT

Our work focus on data management in Vehicular Ad Hoc Networks (VANETs). Many pieces of information may be exchanged in such networks, for instance to warn drivers when a potentially dangerous event arises (accident, emergency braking, obstacle in the road, etc.) or to try to assist them (available parking spaces, traffic congestions, real-time traffic conditions, etc.). Existing systems only use the data exchanged to warn the driver. Then, the data is considered obsolete and is deleted.

In this paper, we rather propose to aggregate the data once it becomes obsolete. Our objective is to produce additional knowledge to be used by drivers when no relevant data has been communicated by neighbouring vehicles. For example, it becomes so possible to dynamically detect potentially dangerous road segments or to determine the areas where the probability to find an available parking space is high when none has been received.

Categories and Subject Descriptors

C.2.4 [Computer-Communication Networks]: Distributed Systems - distributed applications; H.3.4 [Information Storage and Retrieval]: Systems and Software - distributed systems

General Terms

VANETs, Stockage et agrégation des données

1. INTRODUCTION

Nos travaux se concentrent sur la gestion des données dans les VANETs. Ces réseaux sont constitués d'un ensemble d'objets mobiles qui communiquent entre eux à l'aide de réseaux sans fil de type IEEE 802.11, Bluetooth, ou Ultra Wide Band (UWB). Avec de tels mécanismes de communication, un véhicule peut recevoir des informations de ses voisins proches ou d'autres plus distants, grâce aux techniques de multi-sauts qui exploitent dans ce cas des objets intermédiaires comme relais.

De nombreuses informations peuvent être échangées dans le contexte des VANETs, notamment pour alerter les conducteurs lorsqu'un événement survient (accident, freinage d'urgence, véhicule quittant une place de stationnement et souhaitant en informer les autres, etc.). Au fur et à mesure de leurs déplacements, les véhicules sont ensuite "contaminés" par les informations transmises par d'autres. Dans ces environnements particulièrement dynamiques, l'accès aux données repose sur des techniques d'échanges d'informations entre les véhicules. Les données pertinentes sont ensuite exploitées localement pour alerter ou informer le conducteur ou bien stockées pour être interrogées ensuite. Ces dernières années, de nombreux travaux de recherche se sont intéressés à la transmission des informations entre les véhicules tout en contrôlant leur cohérence sur des critères spatiaux voire temporels.

Dans cet article, nous considérons l'exploitation des données de manière sensiblement différente par rapport aux travaux existants. Ces derniers visent en effet à utiliser les données échangées pour produire des alertes aux conducteurs. Une fois ces données utilisées, elles sont considérées définitivement obsolètes. Dans cet article, nous nous concentrons sur la génération dynamique de connaissances à partir des données collectées par les véhicules au cours de leur trajet. Notre objectif est de pouvoir fournir ainsi des informations aux conducteurs, y compris lorsqu'aucun véhicule communicant ne se trouve à proximité.

La suite de cet article est organisée de la manière suivante. La section 2 présente quelques travaux en lien avec les nôtres et expose notre problématique. La section 3 est consacrée à l'extraction de connaissances. Nous y décrivons notamment les structures et méthodes utilisées pour agréger l'information. Enfin, nous concluons dans la section 4.

2. CONTEXTE & PROBLÉMATIQUE

2.1 Etat de l'art

La communication inter-véhicules est un thème de recherche récent. Des contributions intéressantes ont toutefois déjà été proposées, en particulier en ce qui concerne les protocoles d'échange d'informations entre véhicules, pour palier aux limites des techniques dites d'inondation [7].

Les projets CartTalk [11] et FleetNet [5] visent à exploiter les communications inter-véhicules afin de sécuriser la conduite. L'objectif de ces projets est d'assister les conducteurs, via l'utilisation de communications sans fil, afin d'anticiper une situation accidentogène (freinage d'urgence, accident, obstacle sur la chaussée). Ces projets se concentrent sur les aspects communication entre véhicules. Ils utilisent des protocoles de communication Geocast [6, 10] dont le but est de spécialiser le routage des messages en fonction de critères de localisation géographique. Il devient ainsi possible de définir la zone géographique dans laquelle les messages doivent être acheminés, qu'il s'agisse de communications directes entre véhicules ou de communications multi-sauts.

La dissémination de données dans les réseaux inter véhiculaires a donc paru une bonne alternative et de nombreux protocoles spécifiques au V2V ont été proposés. Dans le projet *TrafficView* [12], différents protocoles de dissémination pour des environnements extra urbains sont comparés. Ces protocoles exploitent la direction relative de circulation des véhicules utilisés pour relayer les informations. Différentes métriques sont proposées pour évaluer ces protocoles comme la latence (temps nécessaire pour propager une information entre deux véhicules séparés par une certaine distance), la "performance" d'une diffusion (en terme de nouvelle zone géographique couverte par une diffusion), le pourcentage de véhicules ayant reçu l'information, la précision (en terme d'erreur moyenne dans l'estimation la localisation des autres véhicules sur la route) ou le taux d'utilisation (proportion d'information utile reçu par les véhicules).

Dans le projet *Mobi-Dik* [14], les auteurs proposent une solution de dissémination inspirée du domaine de l'épidémiologie et appliquée au partage d'informations sur des places de stationnement libres. Un véhicule détenteur d'une information se comporte comme le porteur d'une maladie et "contamine" des véhicules au cours de son déplacement. Des mécanismes d'évaluation de la pertinence des informations, sur des critères spatiaux et temporels, sont également proposés afin de déterminer si une information doit encore être diffusée.

Dans [3], les auteurs calculent une probabilité de rencontre afin de déterminer la pertinence des informations pour un véhicule. Cette probabilité de rencontre est exploitée à la fois pour déterminer si une information doit être communiquée aux conducteurs ou diffusée à d'autres véhicules. Dans [2], des mécanismes sont proposés pour contrôler les informations communiquées aux conducteurs.

Dans [8], les auteurs s'intéressent aux environnements urbains. Ils font notamment une distinction intéressante entre le transport des informations via locomotion (avec des stratégies dénommées store-and-forward [1] ou carry-and-forward [15] dans lesquelles les véhicules transportent une information dans une zone géographique où elle peut être diffusée) et via dissémination (diffusions successives). Cette distinction permet notamment de supporter des densités de véhicules différentes lors de la dissémination.

2.2 Problématique

Les différentes solutions présentées dans la section précédente se concentrent sur l'échange d'informations entre les véhicules. Ces travaux ne considèrent toutefois les données échangées entre les véhicules que comme un élément à transmettre. Une fois ces données reçues par un véhicule, leur pertinence est évaluée, généralement sur des critères spatiaux et temporels, et, le cas échéant, le conducteur est informé de la présence d'un événement. Aucune autre exploitation des données n'a été investiguée. Une fois les données utilisées, elles sont considérées obsolètes et détruites.

Dans cet article, nous montrons que les données reçues par un véhicule ne doivent pas servir uniquement pour produire un message d'alerte à destination du conducteur. Une fois stockées sur un véhicule, il est en effet possible d'utiliser a posteriori ces données collectées pour produire, au niveau du véhicule, des connaissances sur l'environnement exploitables par le conducteur. A titre d'exemple, les événements concernant les places de stationnement disponibles reçus par un véhicule peuvent être exploités, lorsqu'il n'y a aucune place disponible diffusée par les autres véhicules, pour déterminer l'endroit où la probabilité de trouver une place libre est la plus importante (en fonction du jour et de l'heure par exemple). Dans un autre contexte, grâce à la corrélation des différents messages reçus sur les accidents et les freinages d'urgence, les zones dangereuses peuvent être dynamiquement détectées et indiquées au conducteur, qu'elles soient d'ailleurs continuellement dangereuses ou seulement temporairement du fait des conditions climatiques par exemple.

3. AGRÉGATION DES ÉVÉNEMENTS

Au cours de son trajet, un véhicule produit et reçoit des événements. Dans la suite, nous considérons qu'un événement est défini comme un n-uplet (estampille, type d'événement, localisation). Lors de leur dissémination, ces événements peuvent avoir une description plus riche (valeurs associées au type d'événements, vecteur de direction, etc.) qui n'est prise en compte ici. Le processus d'agrégation doit avoir les propriétés suivantes :

- résumer les événements en favorisant les dimensions fondamentales que sont la localisation et le temps;
- le processus doit être incrémental et le volume de données stocké rester faible;
- chaque conducteur doit pouvoir choisir à quels types d'événements il/elle s'intéresse et avec quelle échelle spatiale et temporelle il/elle veut agréger;
- les données agrégées doivent pouvoir être échangées entre les véhicules pour enrichir leurs connaissances respectives;

- le processus d'agrégation doit être peu coûteux.

De nombreux travaux ont été faits sur l'extraction de connaissances ou les résumés de bases de données [13, 4, 9]. Dans la suite, nous proposons une approche adaptée au contexte des réseaux ad hoc inter-véhicules. Notre approche repose sur une agrégation simple du nombre d'événements selon les dimensions spatiale, temporelle et leur type.

3.1 Structure d'agrégation pour les événements

Chaque véhicule doit maintenant conserver, en plus des événements "actifs" reçus, un résumé des événements passés structuré de la manière suivante : à chaque type d'événement est associé une matrice à deux dimensions (spatiale et temporelle). Chaque cellule de cette matrice contient le nombre d'événements considérés et une valeur de confiance comprise entre 0 et 1 (voir figure 1).

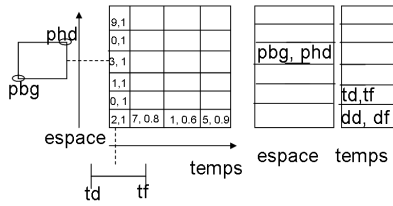


Figure 1: Structure de stockage des événements

Pour la dimension spatiale, nous considérons un découpage de l'espace en rectangles. Ce découpage n'est pas forcément complet (un conducteur ne s'intéresse pas à tout l'espace) et la taille des rectangles n'est pas forcément homogène (des zones requièrent une information plus précise que d'autres). Pour cela, à chaque matrice de type d'événements est associé un tableau définissant l'ensemble des rectangles. Chaque cellule du tableau donne le point bas gauche (pbg) et le point haut droit (phd) du rectangle. Pour la dimension temporelle, il faut choisir d'une part la fenêtre d'observation définie par la date et l'heure de début qui détermine les événements à agréger et d'autre part le découpage du temps qui peut aussi être non homogène. Un tableau associé au type d'événement définit également le découpage temporel choisi avec le temps de début (t_d) et le temps de fin (t_f) ainsi qu'en premier élément la date de début (d_d) et la date de fin (d_f) des agrégations. Par ailleurs, une valeur de confiance indique si l'information est sûre (les événements ont bien été observés dans cette cellule) ou approchée (les événements ont été observés dans une cellule plus grande selon l'une ou l'autre dimension). Pour être capable de gérer une fenêtre d'observation glissante des événements, nous proposons pour l'instant de stocker différentes versions de chaque matrice à des temps différents et de calculer par une interpolation simple les versions manquantes. Un exemple est présenté dans la figure 1. La première colonne de la matrice contient 6 agrégations distribuées dans l'espace au premier intervalle temporel défini. Toutes ces observations sont sûres (valeur de confiance à 1). La première ligne donne quant à elle l'agrégation des événements observés sur le premier rectangle spatial pour tous les intervalles de temps. Les valeurs de confiance peuvent cette fois y être inférieures à 1. Potentiellement, chaque matrice peut comprendre un nombre

important de 0. Des techniques de stockage de matrices creuses peuvent donc être utilisées pour les compacter.

3.2 Algorithmique de base

Les matrices d'agrégation sont utilisées pour stocker et manipuler des résumés d'un ensemble d'événements connus. L'opération de base sur ces matrices est la réduction d'une cellule relativement à une autre. Cette opération peut être utilisée pour l'échange de résumés entre véhicules, pour restructurer un résumé ou pour résoudre une requête sur un résumé. Elle consiste à comparer deux cellules qui n'ont pas forcément le même référentiel spatio-temporel.

fonction réduction ($c1, c2$) retourne sous-cellule

$c1$ et $c2$ sont des cellules de matrices concernant le même événement. $c1$ est la cellule cible et il faut donc le cas échéant éclater $c2$ en plusieurs sous-cellules dont une aura le même référentiel spatio-temporel que $c1$. Plus précisément les cas suivants sont à traiter :

- $c2$ est "include" dans $c1$: on retourne $c2$;
- $c2$ est disjointe de $c1$: on retourne la cellule vide;
- $c1$ est "include" dans $c2$: il faut découper $c2$ pour l'adapter à $c1$ et donc créer au plus cinq cellules pour la dimension spatiale (rectangle commun, rectangles à droite, à gauche, au dessus et en dessous) et trois pour la dimension temporelle (intervalle commun, intervalle précédent et intervalle suivant). La valeur agrégée de $c2$ est alors distribuée uniformément entre l'ensemble des cellules créées puisque l'agrégation nous a fait perdre l'information sur la localisation spatio-temporelle plus précise :

$$valeur(sous - c2) = \frac{valeur(c2)}{nb}$$

$$conf(sous - c2) = \frac{conf(c2)}{nb}$$

avec nb qui correspond au nombre de cellules créées. On retourne alors la sous-cellule *sous-c2*;

- $c1$ et $c2$ ont une intersection non vide : on se ramène au cas précédent en ne considérant que la zone intersection de $c1$ et $c2$.

3.3 Echange de résumés entre véhicules

Chaque véhicule peut décider de publier tout ou partie de ses résumés vers d'autres véhicules et réciproquement peut être intéressé à souscrire à tout ou partie des résumés des autres. Pour simplifier, nous ne considérons ici que des publications et souscriptions publiques (on publie/souscrit vers tous les véhicules). La publication consiste à définir quels sont les résumés publiés (éventuellement en les agrégeant). L'agrégation consiste ici à regrouper des cellules. La souscription consiste à définir des filtres indiquant les types d'événements qui nous intéressent en y ajoutant le cas échéant des critères spatiaux et temporels. Par exemple, je suis intéressé par le type "accidents" dans la zone "Paris" sur le mois dernier. L'échange d'informations entre véhicules peut ensuite se faire soit par le biais de relais (par exemple des serveurs localisés le long des routes), soit directement (en supposant que le

temps de rencontre soit suffisamment long pour que l'échange puisse avoir lieu). La première phase est la comparaison entre les publications d'un véhicule A et les souscriptions d'un véhicule B. Si une ou plusieurs matrices correspondent, le processus calcule la fusion des matrices correspondantes pour A et pour B.

procédure fusionmatrice(m1, m2)

m1 est la matrice cible du processus de fusion. On compare successivement toutes les cellules de m1 avec toutes celles de m2 en utilisant la fonction *réduction*. Si la valeur retournée est différente de vide on applique la fonction *fusioncellule*.

fonction fusioncellule(c1, c2) retourne c3

c1 et c2 ont le même référentiel spatio-temporel. Il suffit de fusionner les valeurs agrégées des deux cellules et de dériver le nouveau facteur de confiance. Les événements de c2 peuvent être intégrés dans c1, sans toutefois prendre en compte des doublons éventuels. Puisque qu'aucune information plus précise n'est disponible, nous approximations la valeur agrégée par réduction en prenant le max des deux valeurs¹ :

$$\text{valeur}(c3) = \max(\text{valeur}(c1), \text{valeur}(c2))$$

$$\text{conf}(c3) = \frac{\text{valeur}(c1) * \text{conf}(c1) + \text{valeur}(c2) * \text{conf}(c2)}{\text{valeur}(c1) + \text{valeur}(c2)}$$

3.4 Exploitation des agrégations

Chaque cellule d'une matrice peut être vue comme un événement abstrait en tout point semblable à un événement observé à la différence notable qu'il est incertain. Un événement observé a une probabilité de 1 alors qu'une cellule avec une valeur agrégée à 0 a une probabilité de 0. Pour les cellules non vides, la probabilité va être estimée à partir du nombre d'événements observés et la valeur de confiance. Les besoins en recherche d'événements peuvent être assez riches et vont de simples filtrages sur un seul type d'événements à des corrélations complexes sur plusieurs types d'événements. Dans l'état actuel de nos travaux, nous nous consacrons aux filtrages simples. Nous considérons que le filtrage peut être construit à partir d'une requête de base qui recherche le nombre d'événements observés dans une zone spatio-temporelle donnée. Nous utilisons pour cela les fonctions suivantes :

fonction requête(zone1, matrice1) retourne (valeur, conf)

Cette fonction applique la fonction *réduction* sur toutes les cellules de matrice1 afin d'obtenir toutes les sous-cellules de matrice1 ayant une intersection non vide avec zone1. La fonction *unioncellules* est alors utilisée pour faire l'union ces sous-cellules.

fonction unioncellules(ensemble de cellules) retourne cellule

La zone définissant la cellule résultat est un sous-ensemble de la zone définissant la requête. La valeur agrégée est la somme des valeurs des différentes cellules. Le facteur de confiance est la moyenne pondérée des facteurs de confiance.

4. CONCLUSION & PERSPECTIVES

Dans cet article, nous avons posé les bases d'une solution permettant d'extraire des connaissances en agrégeant des

¹Il est également possible de travailler de manière plus précise en définissant un intervalle $[\max(\text{valeur}(c1), \text{valeur}(c2)), \text{valeur}(c1)+\text{valeur}(c2)]$.

données échangées entre véhicules. De nombreux travaux restent nécessaires. Nous nous intéressons d'une part à l'étude de techniques de résumés plus fidèles et d'autre part à l'expression de requêtes portant sur des corrélations entre types d'événements.

5. REFERENCES

- [1] C. Adler. Information dissemination in vehicular ad hoc networks. Master's thesis, Univ. of Munich, 2006.
- [2] N. Cenerario and T. Delot. Evaluation continue de requêtes dans les réseaux de communication inter-véhicules. In *Atelier INFORSID sur la Gestion de Données dans les Systèmes d'Information Pervasifs (GEDSIP'07)*, 2007.
- [3] N. Cenerario, T. Delot, and S. Ilari. Dissemination of information in inter-vehicle ad hoc networks. In *Intelligent Vehicles Symposium (IV'08)*, 2008.
- [4] D. Chan and J. Roddick. Summarisation for mobile databases. *Journal of Research and Practice in Information Technology*, 37(3), 2005.
- [5] A. Festag, H. Füssler, H. Hartenstein, A. Sarma, and R. Schmitz. Fleetnet: Bringing car-to-car communication into the realworld. In *World Congress on Intelligent Transport Systems (ITS)*, 2002.
- [6] H. Fußler, M. Mauve, H. Hartenstein, M. Kasemann, and D. Vollmer. Location-based routing for vehicular ad-hoc networks. In *Int. Conf. on Mobile Computing and Networking (MobiCom'02)*, 2002.
- [7] W. R. Heinzelman, J. Kulik, and H. Balakrishnan. Adaptive protocols for information dissemination in wireless sensor networks. In *Int. Conf. on Mobile Computing and Networking (MobiCom'99)*, 1999.
- [8] C. Lochert, B. Scheuermann, M. Caliskan, and M. Mauve. The feasibility of information dissemination in vehicular ad-hoc networks. In *Conf. on Wireless On demand Network Systems and Services (WONS'07)*, 2007.
- [9] I. F. V. Lopez, R. Snodgrass, and B. Moon. Spatiotemporal aggregate computation: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 17(2), 2005.
- [10] C. Maihofer. A survey of geocast routing protocols. *IEEE Communications Surveys & Tutorials*, 6(2):32-42, 2004.
- [11] P. Morsink, R. Hallouzi, I. Dagli, C. Cseh, L. Schafers, M. Nelisse, and D. D. Bruin. Cartalk 2000: Development of a cooperative adas based on vehicle-to-vehicle communication. In *Intelligent Transport Systems and Services*, 2003.
- [12] T. Nadeem, P. Shankar, and L. Iftode. A comparative study of data dissemination models for VANETs. In *3rd Int. Conf. on Mobile and Ubiquitous Systems (MOBIQUITOUS'06) - Workshops*, 2006.
- [13] G. Raschia and N. Mouaddib. A fuzzy set-based approach to database summarization. *Int. Journal of Fuzzy Sets and Systems*, 129(2), 2002.
- [14] B. Xu, A. M. Ouksel, and O. Wolfson. Opportunistic resource exchange in inter-vehicle ad-hoc networks. In *5th Int. Conf. on Mobile Data Management*, 2004.
- [15] J. Zhao and G. Cao. VADD: Vehicle-assisted data delivery in vehicular ad hoc networks. In *Int. Conf. on Computer Communications (INFOCOM'06)*, 2006.