Route Classification in Travel Time Estimation Based on Cellular Network Signaling

David Gundlegård and Johan M Karlsson

Linköping University Post Print

N.B.: When citing this work, cite the original article.

Original Publication:


©2006 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Postprint available at: Linköping University Electronic Press http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-50949
Route Classification in Travel Time Estimation Based on Cellular Network Signaling

David Gundlegård
Department of Science and Technology
Linköping University
Norrköping, Sweden
davgu@itn.liu.se

Johan M Karlsson
Department of Science and Technology
Linköping University
Norrköping, Sweden
johka@itn.liu.se

Abstract— Travel time estimation based on cellular network signaling is a promising technology for delivery of wide area travel times in real-time. The technology has received much attention recently, but few academic research reports has so far been published in the area, which together with uncertain location estimates and environmental dependent performance makes it difficult to assess the potential of the technology. This paper aims to investigate the route classification task in a cellular travel time estimation context in detail. In order to estimate the magnitude of the problem, two classification algorithms are developed, one based on nearest neighbor classification and one based on Bayesian classification. These are then evaluated using field measurements from the GSM network. A conclusion from the results is that the route classification problem is not trivial even in a highway environment, due to effects of multipath propagation and changing radio environment. In a highway environment the classification problem can be solved rather efficiently using e.g., one of the methods described in this paper, keeping the effect on travel time accuracy low. However, in order to solve the route classification task in urban environments more research is required.

Keywords- Travel time estimation, cell phone positioning, tracking, route classification

I. INTRODUCTION

The evolution of cellular based road traffic information estimation has so far mostly been driven by commercial companies. This means that few research publications are available describing the estimation approaches. This is, however, needed to give a thorough understanding of the methods and concepts, and hence raise the potential for greater improvements in the area. A few exceptions are the work of Zhi-Jun Qiu et al. [1-2], and Bruce Hellinga et al. [3-4]. In order to estimate road traffic information from cellular networks the following basic phases can be performed:

- Location data collection
- Terminal classification
- Map matching
- Route determination and
- Traffic state calculation

The location data collection phase involves how to gather the relevant data from the cellular network. The terminal classification focus on determining which of the mobile terminals that are located on the road network and in which type of transport mode they are. The map matching phase includes how to connect the location data to the road network, whereas the route determination process is used to determine the route of the vehicle, consisting of several road sections. The last phase performs the traffic state calculation. Depending on the traffic information of interest, e.g., long term traffic flow estimation, origin-destination (O/D) estimation, travel time estimation and incident detection, different approaches and location data will be used. In this paper we mainly assume travel time estimation using passive monitoring and handover data. The phases can be carried out in a different order, in combination with each other and there can also be iteration between the different phases.

Since the location data from the cellular network is relatively coarse, the map matching phase is quite important. Map matching and route determination are often combined. Depending on whether active or passive monitoring is used the map matching approach will differ. In active monitoring, due to the relatively slow sampling rate, an approach similar to map matching for regular GPS-based floating car data (FCD)

![Figure 1. The problem studied in this paper is to determine whether section S_A or S_B is traversed given the location events L_1 and L_2 together with the sequence of reports X_1-X_6.](image-url)
is used [5], i.e., variants of point-to-curve matching. Since the location accuracy is quite low, using previous locations and connections between different sections is more important compared to GPS-based systems. In passive monitoring, the sampling rate will typically be much higher, which means that some kind of curve-to-curve matching might be interesting. Another difference is also that the terminals are located spatially uniform, i.e., at the same place every time a travel time is calculated, which means that the map matching task is more predictable. If the task of map matching involves how to match one or more location samples to a road section, the route detection task is to determine the order of sections that has been traversed. This can be done using the terminal locations in combination with the network structure and road intersection decisions. Useful input to the route determination is also the measurement reports, since pattern matching techniques can be used to determine the most likely section or route. When determining the route and matching positioning measurements to points on the road, a radio map including handover locations and signal strengths is needed. This map can be estimated using propagation models, test drives or a combination.

Recently three extensive reports have been published summarizing the different projects and field trials that have been performed within the field of estimating road traffic information from cellular networks [6-8]. The conclusions from these reports are quite similar; the technology has large potential although more research is needed in order to use it as efficiently as possible. Promising results have been shown on highways, in arterials and even urban environments. As motivated in [9], it is reasonable to assume that the handover location accuracy in GSM and UMTS networks are accurate enough to make useful estimations of travel times, provided that the terminal classification and route determination tasks are performed well enough. In order to make a better prediction on the usefulness of travel time estimations based on GSM and UMTS signaling data, the aim of this paper is to examine the route determination task in more detail.

II. LOCATION DATA IN GSM AND UMTS

The type of location data available for a terminal in GSM and UMTS depends on the state of the terminal. The state, on the other hand, depends on how the terminal is used, e.g., used for surfing the web or making telephone calls. In both GSM and UMTS, terminals used for real-time services generate the most detailed location data and we will focus on this kind of data throughout the paper. Terminals used for real-time services continuously send measurement reports about the radio environment in order to assist the network in the handover decision. These measurement reports together with handover points can be used to track the route of the vehicle and calculate travel times. The measurement reports in GSM are sent every 480 ms and they contain the signal quality (RxQual) of the serving base station, signal strength (from now on denoted RxLev) of surrounding base stations and a timing advance (TA) value that gives a rough estimation of the distance to the serving base station. The measurement reports in UMTS are more flexible, the reports with constant interval are complemented with event-triggered reports. The reports will be sent more often in UMTS (implementation dependent, but typically 200 ms) and will also contain different information. The number of surrounding base stations that can be measured is increased from 6 in GSM to 32 in UMTS. A TA value is not calculated in UMTS (WCDMA) networks since it is not a TDMA based system, but other time alignment measurements are available, e.g., round trip time and time difference between base stations.

An important difference between GSM and UMTS is the possibility to use soft handover in UMTS. This means that a terminal can be connected to several base stations simultaneously in UMTS, whereas in GSM the terminal is only connected to one base station at the time. To track a vehicle, both measurement reports containing radio parameters and handover points can be used. When it comes to calculating travel times it is very important to have two accurate estimations of the vehicles position in order to make a good estimation of the travel time between those points. The crossings of cell boundaries are a good candidate to estimate those positions.

III. PROBLEM DEFINITION

The problem studied in this paper can be seen as a sub problem in estimating travel times based on cellular network signaling, but it can also be seen as a sub problem in the more general problem of tracking vehicle located cell phones in a road network, independent of the intended ITS application, e.g., travel time estimation, stolen vehicle recovery, fleet management and road user charging.

Consider a sequence of reports \((X_1, X_2, \ldots, X_n)\) generated by the cellular network. Each report is a vector containing measurements of \(m\) attributes, where \(m\) typically varies in consecutive reports and has an upper limit of \(M\). The general problem of tracking a vehicle located cell phone can be formulated as

\[
\text{Given a report sequence } (X_1, X_2, \ldots, X_n),
\]

\[
\text{determine the sequence of traversed road sections } (S_1, S_2, \ldots, S_k). \tag{A}
\]

A report \(X_i\) can include for example signal strength (RxLev) and timing events to serving and neighbouring base stations. \(M\) could be for example 9 (serving RxLev, RxQual, timing advance and 6 neighbour RxLev) in the GSM network. Considering the travel time estimation algorithm, we typically need two locations in order to calculate a travel time. An important advantage of a passive monitoring system is that these two locations can be chosen in positioning friendly locations, i.e. where the positioning accuracy is good. Suppose we have two location events \((l_1, l_2)\), which we use to estimate the travel time, the modified problem becomes
Given two location events \((L_1, L_2)\) determine the sequence of sections \((S_1, S_2, ..., S_k)\) that has been traversed between \(L_1\) and \(L_2\), based on the sequence of measurement reports \((X_1, X_2, ..., X_n)\)

Depending on the road and cellular network structure, a pair of location events can uniquely define the section or sections that have been traversed, or there can be ambiguity in transformations from location events to traversed sections.

Typically, there will be ambiguity in location events if we have close-by parallel roads or if we have a relatively dense road network compared to the cell sizes. The location events can be triggered by a sequence of reports or they can be explicit events in the cellular network, a commonly used location event in current systems is the handover event.

The problem studied in this paper includes the assumption that two location events uniquely defines a pair of parallel sections and the task is to determine which of the two possible routes that has been traversed, i.e.,

Given two location events \((L_1, L_2)\) determine which of the two parallel sections \((S_1, S_2)\) that has been traversed between \(L_1\) and \(L_2\), based on the sequence of measurement reports \((X_1, X_2, ..., X_n)\)

This problem will most likely occur in the cellular travel time estimation process when there are close-by parallel roads or in urban environments. The effects of the problem can be severe due to large differences in travel times on for example a highway and a small secondary road running in parallel. A high travel time on the highway can be misclassified to be on the secondary road and a normal travel time on the secondary road can be misclassified as a high travel time on the highway. The problem is illustrated in Figure 1.

IV. PREVIOUS WORK

The literature within general positioning of cell phones is extensive and much of the work has been performed the last five years. However, the area of tracking cell phones based on cellular network signalling has received much less attention. The previous literature within tracking cell phones based on cellular network signalling for the purpose of deriving road traffic information is, with a few exceptions, non-existent. This is mostly due to lack of academic research in the area.

Route determination error impact on travel time estimations is an almost unexplored, but important, area. In [11] hidden Markov models are used to estimate traffic information based on measurement reports. The performance is evaluated using a simulation and shows promising results in estimating the speed on a road section, which includes the route determination task. In [12] live measurement reports from a single base station in a city environment was used to determine the route of a vehicle. Hidden Markov models were used and 6 out of seventeen routes were classified correctly, six times the correct route was highly probable and three times the classification failed. Our research investigates the route determination problem further and explains it in terms of travel time estimation based on cellular network signalling. The difference in this paper compared to previous work is that neighbouring cells are included in the algorithm and that a different method for classification is used.

As described previously, we approach the problem as a classification problem based on a sequence of measurements. This means that we do not try to explicitly estimate the cell phone position in other points than the location events. However, the problem can also be solved by trying to estimate a position for every measurement report and then applying filtering and map matching methods in order to determine the traversed route. This approach is used in [13] based on a similar distance metric as in this paper in combination with a Kalman filter with promising results.

V. METHOD

The experiments will aid in determining the magnitude of the cellular travel time error caused by route misclassification. This is important input in order to determine the potential of travel time estimation based on cellular network signalling. The classification methods that are used are Bayesian classification method and a version of the weighted K Nearest Neighbour method. The classification methods will be assessed for the problem defined in (C), using data collected from GSM idle mode terminals in a highway environment with relatively close-by parallel road sections. The classification methods are evaluated using 208 test drives on six different sections where three pairs of sections are relatively close-by and parallel. 60 test drives are used for training the classification models. The six sections (S1-S6) are shown in Figure 2, where also the location events (L1-L6) are marked. The main part of the test drives are made with a standard GSM module that makes standard idle mode measurements and send the averaged measurements together with GPS coordinates to a server every two seconds.

![Figure 2](image-url) Location events with corresponding sections of the experiment area showing the length of one of the sections.
A couple of test drives are also made with the TEMS Investigation 9.0 tool provided by Ericsson AB in order to identify possible location events in the test area in terms of handovers in the GSM and UMTS networks.

In the problem definition (C), we assume that there will be location events, e.g., handovers, which can be used to identify the end points of the different sections. In order to justify this assumption for both the GSM and UMTS network, the test drives with the TEMS tool is used. Figure 3 shows the handovers (GSM) and radio link additions/removals (UMTS) that occurred during a test drive on two of the sections. It can be seen that several handovers are possible as location events and it can also be seen that in some cases the handover occurs at the same time for the parallel sections. The consistency and accuracy of GSM and UMTS handover are further discussed in [5, 11].

It is important to note that the data collection from the GSM modules are carried out in idle mode, which means that the GSM module makes less measurements of the radio environment compared to when the cell phone is engaged in a phone call. The sampling frequency is four times higher during a phone call compared to idle mode (2 Hz compared to 0.5 Hz). Current systems for travel time estimations typically use cell phones that are engaged in a call, which means that the results in this paper are based on a much lower sampling frequency and can hence be seen as a lower bound of the classification performance. Using idle mode modules also means that it was not possible to extract the Timing Advance value or the RxQual value from the module.

A. Nearest Neighbour Classification

The first method is based on nearest neighbour classification. The K Nearest Neighbour classification algorithm is a lazy learning algorithm, since it does not need a trained model in order to perform the classification task. Instead a distance metric is used to compare the distance between the evaluation sequence and the training sequences. If we use \( K = 1 \), the chosen class will be the class with the shortest distance between a training sequence and the evaluation sequence in terms of the specified distance metric.

Increasing \( K \) means that we pick out \( K \) closest neighbours and choose the class from these according to some rule, e.g. majority or the weighted sum of distances. We chose the following distance metric between an evaluation sequence \( p \) and a training sequence \( q \):

\[
D_{pq} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M_y} (X_{ij} - Y_{ij})^2, \quad \text{where} \quad (1)
\]

\( X_{ij} = \text{RxLev} \) for cell \( j \) in report \( i \) for training sequence \( q \)

\( Y_{ij} = \text{RxLev} \) for cell \( j \) in report \( i \) for evaluation sequence \( p \)

\( M_y = \text{largest number of cells in either } X_{ij} \text{ or } Y_{ij} \)

\( C_i = \text{number of measurements > 0 in both report } X_{ij} \text{ and } Y_{ij} \)

\( N = \text{number of reports that are compared} \)

For \( K > 1 \) we use two alternatives for classification within the \( K \) training sequences. The first one is based on majority voting and the second one is based on choosing the class with the largest sum of the weighted training sequences belonging to one class. The weight for training sequence \( q \) is defined as

\[
W_q = \frac{1}{D_{pq}}, \quad \text{where} \quad (2)
\]

\( D_{pq} \) is the distance between evaluation sequence \( p \) and training sequence \( q \).

B. Bayesian Classification

If the probability density functions (pdf) of the RxLev for the different cells and the probabilities of traversing the different road sections were known, it would be possible to choose the most probable road section using Baye’s formula.

Unfortunately neither the RxLev pdf nor the probabilities of traversing the different road sections are known. However, they can be estimated using the training data and can be used in a Bayesian classifier.

Figure 3. Possible location event candidates in terms of handover in UMTS (left) and GSM (right), respectively. The picture shows one pair of parallel sections (S3-S4) that were used in the evaluation.
For the Bayesian classification we define the a posteriori probability \( P(S_i | X) = \) probability that road section \( i \) has been traversed given the measurement vector \( X \).

For simplicity and in order to get more stable decision variables we start by using the measurement vector \( X = [X_1, \ldots, X_N] \), where
\[
X_j = \frac{1}{N} \sum_{n=1}^{N} X_{nj}
\]
(3)
is the average RxLev of cell \( j \) taken over all \( N \) measurements in a road section. Using Baye’s formula we get
\[
P(S_i | X) = \frac{P(X | S_i) \cdot P(S_i)}{P(X)}
\]
and the road section with the maximum a posteriori probability is chosen in the classification method. Since \( P(X) \) is constant for all sections, it can be excluded from the decision. \( P(X | S_i) \) and \( P(S_i) \) can be estimated from training data. Also \( P(S_i) \) can be excluded from the decision if we have an equal number of training measurements for all sections, or if we do not want to increase the probability of choosing a common (training) section. Assuming statistical independence between the RxLev of the different cells yields
\[
P(X | S_i) = \prod_{j=1}^{J} P(X_j | S_i)
\]
where \( J \) is the number of cells detected while passing road section \( i \). A second assumption is that the deviation of the average RxLev for one cell is normally distributed, i.e.,
\[
\bar{X}_j = X_j - \bar{X}_j \sim N(0, \sigma_j)
\]
(5)
\[
\bar{X}_j \approx \frac{1}{Q} \sum_{q=1}^{Q} X_{jq},
\]
(6)
\[
\sigma_j = \sqrt{\frac{1}{Q} \sum_{q=1}^{Q} (X_{jq} - \bar{X}_j)^2}
\]
(7)
\( X_{jq} \) is the average RxLev of cell \( j \) in training sequence \( q \).

The most probable road section is hence determined by
\[
\arg \max_i P(S_i | X) = \prod_{j=1}^{J} P(X_j | S_i) \cdot P(S_i)
\]

However, using the measurement vector \( X = [X_1, \ldots, X_N] \), where \( X_j \) is the average RxLev over all measurements in one road section (3), results in poor classification performance. This is due to the fact that a large part of the information identifying a specific route is contained in the RxLev variation for one cell during the road section. To improve the performance of the Bayesian classifier we need to incorporate the RxLev variation over the section in the model. This can be done for example using the measurement matrix
\[
X = \begin{bmatrix}
X_{11}, X_{12}, \ldots, X_{1M} \\
\vdots \\
X_{M1}, X_{M2}, \ldots, X_{MN}
\end{bmatrix},
\]
where \( M \) is the number of measurement reports that are reported while passing the road section, and hence \( X_{ij} \) is the average RxLev for cell \( j \) in measurement report \( i \). Using the previous assumptions, the updated Bayesian classification model chooses the most probable section by determining
\[
\arg \max_i P(S_i | X) = \prod_{j=1}^{N} \prod_{i=1}^{J} P(X_j | S_i) \cdot P(S_i)
\]

In order to adjust the Bayesian classification to the current problem, a couple of modifications are made when calculating the a posteriori probabilities:

- Measurements available in the training set, but not in the evaluation set are not included in the probability calculation
- Measurements available in the evaluation set, but not in the training set generate a penalty probability
  \[
P(X_j | S_i) = p_{\text{pen}}
\]
  where the punishment term is varied between 0.0001 and 0.01 in the experiments
- In order to avoid extremely small probabilities, the a priori probability for cell \( j \) in measurement report \( i \) is changed to
  \[
P(X_j | S_i) = \alpha (P(X_{ij} | S_i) + p_{\text{pen}})
\]
  where \( \alpha \) is set to 10 in the experiments
- When calculating the a posteriori probability \( P(S_i | X) \), only the \( b \) most probable cells are used, i.e.,
  \[
P(S_i | X) = \prod_{j=1}^{M} P(X_j | S_i) \cdot P(S_i)
\]
  where \( B \) is the set of \( b \) most probable cells.

VI. RESULTS

The results of both classification algorithms are rather optimistic taking into account that a relatively low sampling rate is used compared to cell phones engaged in a call. However, the field measurements and the results also tell us that the variations in path loss due to dynamic factors, such as multipath propagation and moving objects, makes the problem non-trivial to solve.

Figure 4 shows the percentage of classifications that are correct in a slightly more difficult problem, i.e., given a report sequence \( (X_1, X_2, \ldots, X_N) \), determine which of the six possible sections \( (S_1, S_2, S_3, S_4, S_5, S_6) \) that has been traversed. Section S1-S6 are shown in Figure 2, where it can be seen that section S1-S2 are slightly more separated compared to S3-S4 and S5-S6. The result is illustrated as a function of nearest neighbours \( K \) and the section average RxLev of a cell that is required in order to be included in the distance calculation. From the figure we can see that in general K-values below 10 give good results on this data set and we can choose \( K=1 \) for simplicity. For \( K>1 \), the difference in performance between the majority classification and the sum weighted with the distance, the distance weighted sum performs slightly better. We can also see that removing weak cells from the calculations can improve the performance unless we choose a too high threshold. In this case the performance drops at approximately average RxLev = 20 and is good in the range of 5-10. The best result is 92.3% correct classifications using half of the 268 files for training and the rest for evaluation. Using 5 training
files for every section reduces the best result to 90 %, i.e., the classification method is not very sensitive to reductions of the training data set.

Figure 4. The classification performance as a function of the classification parameters number of nearest neighbours (K) and minimum average RxLev of a cell in order to be included in the distance calculation.

Table 1 summarizes the results of the classification problem defined in (III.C) for the three pairs of sections (shown in Figure 2). The most separated sections, S1-S2, were possible to classify 100% correct if the parameters were assigned in a good way.

In the second and third case, where the sections are separated by the shortest distance (~250 m open terrain) a few sections were misclassified independent of parameter settings. It should be noted that although the classification methods seems to give good results already on fairly small training sets, increasing the training set will also improve the classification performance.

The performance of the two classifiers is very similar for the analysed data set. However, an advantage of the Bayesian classifier is that the classification task can be implemented more efficiently by utilising a predetermined statistical model for classification, instead of calculating the distance to all training sets. The assumption of independent signal strengths made in the Bayesian classifier is most likely not very close to reality, which means that it could be possible to improve the classifier by implementing a Bayesian belief network, instead of the naïve Bayes classifier used in this paper.

Table 1. Classification results (in % correct classifications) using the two different classification methods for the three pairs of sections evaluated in the field experiments. The result is the best one obtained with the parameter settings described previously.

<table>
<thead>
<tr>
<th>Sections</th>
<th>Nearest Neighbour</th>
<th>Bayesian</th>
<th>Evaluation test runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1-S2</td>
<td>100 %</td>
<td>100 %</td>
<td>50</td>
</tr>
<tr>
<td>S3-S4</td>
<td>99 %</td>
<td>94 %</td>
<td>79</td>
</tr>
<tr>
<td>S5-S6</td>
<td>97 %</td>
<td>99 %</td>
<td>79</td>
</tr>
</tbody>
</table>

VII. CONCLUSIONS

A conclusion of the result is that the problem of route classification based on cellular network signalling is not trivial even in a highway environment due to variations in measurements caused by multipath propagation and environment changes. However, if adequate methods are used, the route classification in a highway environment should not affect the performance of cellular based travel time system to any larger extent. Nearest neighbour classification and Bayesian classification seem to perform well enough to avoid major contribution to the error in travel time estimation based on cellular network signaling.

In order to manage the cellular based route classification in urban environments more research is required. Most likely the UMTS network will perform better compared to GSM when tracking the route of cell phones and this can used in order to develop cellular based road traffic information also in urban areas.

Further work includes testing the classifiers in urban environments using also including measurements from the UMTS network.

REFERENCES