Information discovery and combination from divergent data sources for Traveler Information Systems

Frank P. Terpstra[1], Geleyn R. Meijer[2], Arnoud Visser[1]

[1] University of Amsterdam, Department of Computer Science, Kruislaan 403 1098 SJ Amsterdam, The Netherlands
ftrpstra, arnoud@science.uva.nl
[2] LogicaCMG Unwired Concepts, Laan van Kronenburg 14, P.O. Box 133 1180 AC Amstelveen, The Netherlands
Geleyn.Meijer@logicacmg.com

Abstract. In this paper we describe the development and use of knowledge discovery processes in the area Traveler Information Systems. Our view is that the next generation of these systems will provide travel advise that employs not only the current traffic situation but also on the future state of traffic. We show with actual traffic data that the impact of external influences (such as weather and large public events) are key to achieve accurate predictions. We propose an approach to model these influences in detail and describe an information discovery and combination process supporting this approach. It will combine real-time and historical information on traffic and its influencing factors by means of correlation of data clustered in both space and time. We show our development and experimentation environment which offers solutions to the problems encountered in the development of the information discovery and combination process. Finally we discuss the progress and the challenges ahead.
keywords: Traveler Information System, application, temporal and spatial clustering, information discovery and combination.

1 Introduction to Pervasive Computing for Traveler Information Systems

The project described in the paper deals with development of pervasive computing systems for use in the Intelligent Transport Systems domain. Our view is that in particular the further development of Traveler Information Systems is heavily depended on context aware capabilities and effective combination of historical and realtime data. This latter issue being the subject of this paper.

The employment of Traveler Information Systems has received a great deal of attention in the recent past and it is believed by both the research and commercial communities that traveler information is key to dealing with transport challenges in our congested societies. Current applications for instance are dynamic road displays which keep the traveler up to date with the current traffic

situation. These displays give information about the length of traffic jams, about capacity reduction due to road works or lane closures or provide actual travel times over a given stretch of road. The research in this area is now focusing more on providing alternative route choices based on multi modal traveling. Furthermore, the traveler will not just receive information from dynamic roadside displays but also in the car itself. In fact, personalised information services require some form of machine intelligence to deal with the specific context of the user. Is the user traveling by car or public transport? What are the time constraints on a trip? What are the user interface capabilities and preferences? Research in this field is for instance carried out by the project PALS Anywhere[12]. Here, the development is studied of a personal assistant capable of supporting a user in different situational contexts by adapting the user interface and offering on-line help. For mobile accessible services, the momentary users' needs and usage contexts change over time. A fundamental question is how to realise an adequate individual interaction process. PALS (Personal Assistant for onLine Services) focuses on a generic solution: a personal assistant, which tunes the interaction to these needs and context (e.g. adjusting the information presentation and navigation support to the current device and interests of the user). The PALS project carries out experimental usability research with the PALS functionality on mobile interfaces.

An other relevant aspect for context awareness research is the underlying motivation for people to adopt and use new technologies. Individual characteristics are very different and determine for a large extend when innovations (like a context aware service running on a mobile device) are being appreciated. To understand for instance the relative importance of usability aspects versus the value of the information delivered by a service, insight is needed in the individual preferences. A relevant line of research in this respect is the well-established theory of the diffusion of innovations. It can be used to determine the potential success of mobile services and the adoption rate one might expect for a given service. Everett Rogers[2] has over the course of many years studied the process of diffusion of innovations and the social behaviour of individuals in adopting innovations. Among others, Parker[3] has applied the original diffusion theory in the telecommunications domain and gained experience with forecasting models. His ACCORD analysis is a useful methodology. In 2002, LogicaCMG and UvA started to apply a a quantified model based on the ACCORD methodology to the field of value added services for transport and traffic[4]. The work showed that quantification can be efficient by choosing a well-defined domain and a "one-step-at-the-time" focus. In particular the definition of user oriented quality factors related to the application, is crucial. Further research in this field is ongoing and will be reported elsewhere.

Apart from the context awareness research, there is also a great demand for a better understanding of combining historical and real-time data to make predictions of future travel conditions. These predictions are necessary when providing alternative routes to drivers in order to avoid congestion. The traffic situation on the roads is a constantly changing process, thus a piece of road which is free

from congestion at the time the route advise is given can be congested by the time a driver arrives there. This problem among others is dealt with by research commissioned by the FHWA[3]. In particular within DynaMIT[1], which is part of this research, the problem of prediction is addressed. It focusses on the problem created when a lot of travelers start taking alternative routes based on a predicted future state of traffic. The drivers reaction to the advise he is given will have to be taken into account when making predictions. Informed drivers will have to be modeled in the prediction process. Another area that is recognised within DynaMIT as being important for the prediction of future traffic are external influences such as weather, accidents and large public events (sports events, pop festivals, parades etc..). Within its model there is one factor representing all of these influences however this factor is not the result of explicit modeling. It does not use direct observations of these influences instead it uses regression analysis on historical traffic data to identify distinct patterns of delays caused by these influences. We propose that a detailed model of external influences can give a significant improvement to predictions of the future state of traffic. In the rest of this paper we will explain why we think this is the case and how we plan to implement such a detailed model. First though we will briefly explain where we see the development of this model within the field of Knowledge Discovery. Figure 1 shows a segmentation of the Knowledge discovery field. It makes a distinction between data, information and knowledge, and between discovery and comparison where the first three are objects and the latter two actions which can be performed on them. The research in this paper falls into the information grid category, which we understand as dealing with the combination of structured and abstracted information. We however do not intend to use ontologies and other language based semantics as used in the semantic web. The rest of the paper will be structured as follows In section 2 we will explain what influences are important, what the properties of different influences are and for some of them illustrate this with real world data on delays in traffic. After that, in section 3, we will put forward an information discovery and combination process which uses external influences for generating predictive models. This is followed in section 4 by a description of the development environment. Finally we will briefly discuss some of the issues to be faced in this research and draw some conclusions.

2 Importance of external influences

In this section we will look at the correlation between external influences (weather, special events and incidents) and the pattern of traffic. Furthermore we will analyse the implications of this correlation for the accurate (longterm) prediction of traffic and congestion. Over time traffic patterns have a large amount of variability. There are many causes for this variability, in order to explore these further we divide these causes into three distinct categories.

[3] U.S. Department of Transportation Federal Highway Administration http://www.fhwa.dot.gov

	Data	Information	Knowledge
Discover	Data Mining	Information Mining	Knowledge Mining
Compare	Data Grid	Information Grid	Semantic Web

Fig. 1. the place of the information grid within the field of knowledge discovery

First there is nature, which is almost entirely accounted for by the weather. Different weather conditions such as rain, fog and snow cause different driver behaviour and therefor different traffic patterns.

Second there are the human causes, these consist of two sub-categories of global and local influence. The global events are calendar related phenomenon such as the difference between weekends and work-days but also less regular events such as holidays. These events have effect on the society or country in which this calendar is used unlike the second category which is bound to one geographical location. Large public events such as pop festivals parades or demonstrations can cause an unusually high number of people to come to one place. Furthermore in case of demonstrations or parades streets can be closed causing a reduction in capacity.

The third category are incidents and accidents, these are a separate category because they have a bidirectional relationship with traffic patterns. They cause extra traffic to happen, but the amount of traffic influences the chance of accidents occurring.

For all of these factors at least some degree of prediction is possible. The likelihood of accidents is the most difficult to predict as it has many dependencies upon other uncertain information. The weather however can be accurately predicted up to 5-7 days ahead. Both the global calendar related events and the local events can be known for a time period reaching much further into the future. To illustrate the importance of external influences we present real world data on several calender based events both with a global influence and with a very local influence. The data was obtained from our partner info.nl who are currently building a detailed database of all traffic-jams occurring in the Netherlands. Figure 2 shows the total length of traffic-jams in the Netherlands during the Good Friday holiday with two typical Fridays included for comparison. The graphs show that traffic on good Friday is very different from a typical Friday, the morning rush hour is missing instead there is a prolonged peak later in the morning. In Figure 3 a similar situation is shown this time for Easter Monday. During daytime there are virtually no traffic-jams, however there is a prolonged peak in the evening. This in stark contrast to typical Mondays which show very distinct morning and evening rush hours. These two examples showed a very

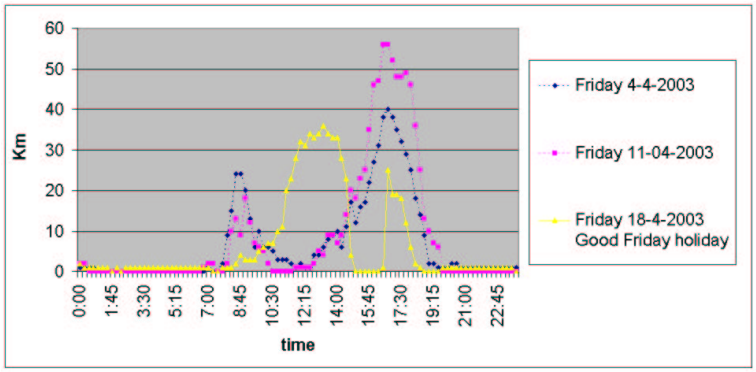

Fig. 2. Total length of traffic-jams in the Netherlands on the good Friday holiday and two typical Fridays

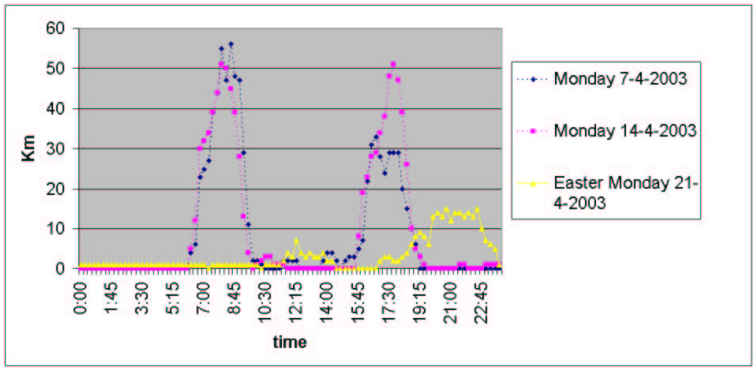

Fig. 3. Total length of traffic-jams in the Netherlands on the Easter Monday holiday and two typical mondays

strong global influence on traffic. It should be noted that these figures present an abstraction of many local events the majority of which is an effect of the global influences. Local events can also occur without the existence of a global influence. Figure 4 shows the length of traffic jams at the Amsterdam ring-road near the RAI exhibition centre. At Saturday 5-4-2003 a exhibition was being held which attracted 240.000 people over a nine day period. The graph shows a traffic-jam starting just before 11:00 the opening time of the exhibition. It is very difficult to predict the traffic-jams shown in these graphs without knowledge about their causes. Both Easter and Good Friday have different dates each year, although there is an exact pattern in their occurrence. Well attended exhibitions do not follow any exact pattern in time at all. Predicting these events only on the basis of historical traffic data is unnecessarily difficult and will give less certain and accurate results than correlating traffic data with its influences.

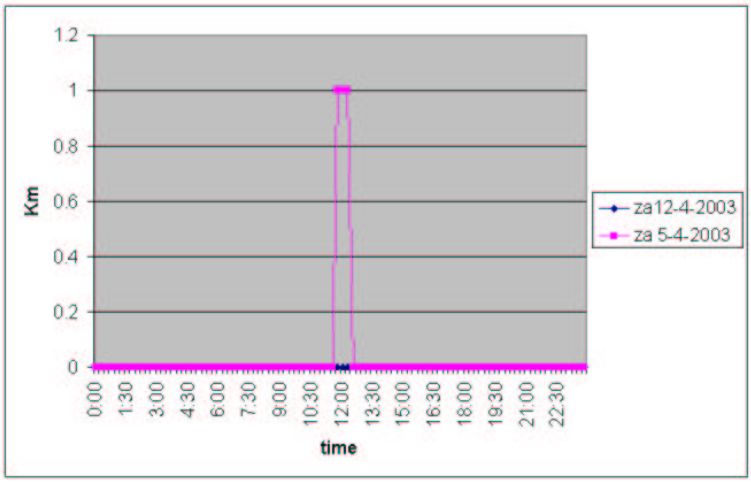

Fig. 4. Length of traffic-jams near the ring-road exit for the RAI exhibition centre in Amsterdam during a well attended exhibition and on a typical saturday

3 Information discovery and combination

The information discovery and combination process described in this section takes a central role in our research. We focus on this subject not only because we believe it plays a central role in improving Traveler Information Systems but also because we believe that it has applications outside of this domain. This section describes how such a system will work and touches on some of the issues that will need to be resolved. In general what happens in this process is the creation of a model predicting the behaviour of autonomous entities, car-drivers in the application described in this paper. The prediction made by this model is based on several data sources.

- Historical data on behaviour of the autonomous entities
- Historical data on factors influencing the behaviour
- current data on behaviour
- current data on factors of influence
- predicted future data on factors of influence

To ensure the reusability of this information discovery and combination process we will use a standard such as PMML[4][9] to define our models. This allows for the use of our models in other domains where they adhere to such a standard. Furthermore to keep our solution general it has to be independent of the type of database used. The VLAM-G virtual laboratory [5] wich is being developed at the University of Amsterdam offers a common database interface which can solve this problem.

[4] Predictive Model Markup Language a standard defined by the Data Mining Group http://www.dmg.org

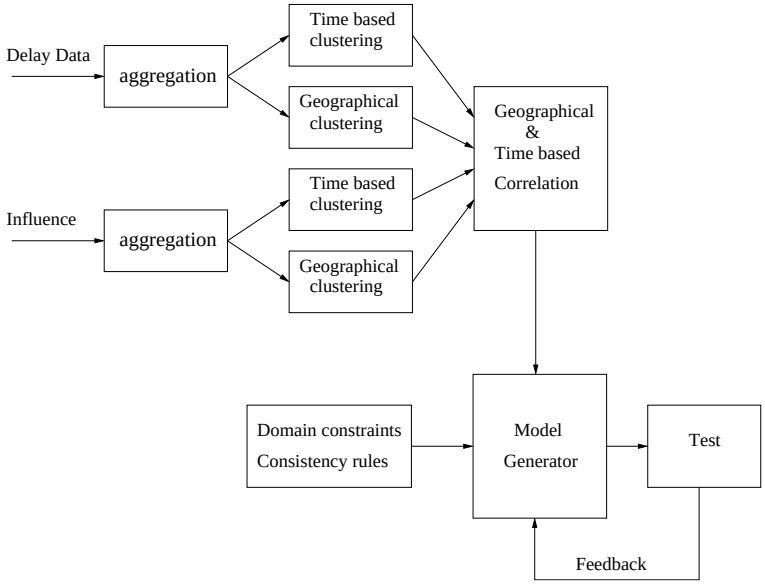

Fig. 5. Proposed model for data combination

Our approach to data combination is outlined in Figure 5. The raw data used as input is usually a series of observations from different locations at different times. By clustering this data one tries to identify distinct patterns, for instance within traffic and weather. Other data such as calendar data and information about events is discreet by itself so it does not need this step performed on it. When performing the clustering the next step in the process, correlation, has to be taken into account. As the examples of section 2 already showed, different influences act at different granularity both in space and time. Therefor correlation has to take place with data that is aggregated over the appropriate temporal an spatial intervals. This requires many aggregations to be performed on the incoming data, research by Johannes Gehrke et. al. [7] offers solutions in this area.

The model generator is the centre of the knowledge discovery process, it has to determine which pieces of knowledge gained from correlation should be used for prediction given the relevant current observations. Furthermore the information used has to be weighted, for instance when predicting the traffic in the vicinity of a pop festival, data about the last time it was held is far more important than data form last week, while for an ordinary day last weeks data might very well be the most relevant. To achieve this the learning process described in Figure 6 is used. This process can use the historical database as a training-set, a very similar process can be employed once the system is in use. Training on the historical database starts with mining the first half of the database, until timestamp T, then using the knowledge gathered to make a prediction about T+X where X equals the amount of time ahead the prediction has to be. The prediction is

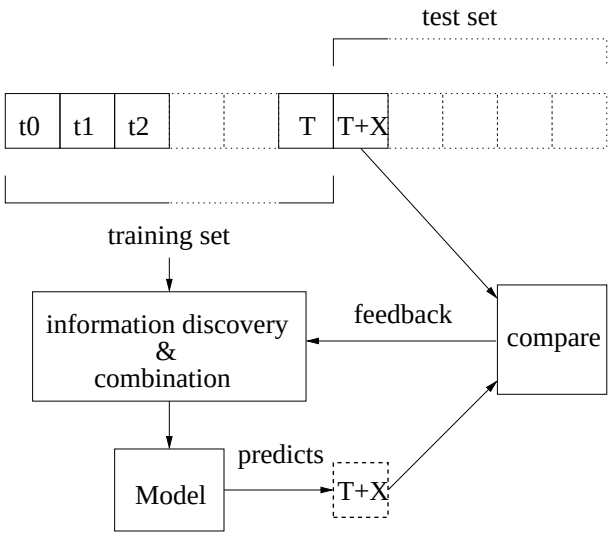

Fig. 6. Learning process using historical data.

checked with the actual data at timestamp T+X, the model generator tries to adjust to the feedback received, after which the process is repeated with T equaling T+X. To ensure that the models which are generated are consistent with reality, domain constraints and consistency rules have to be adhered to. In the traffic domain these can consist of information about the shape of the roadnetwork, and rules such as there can be only one trafficjam at a specific time and place.

When employed in a real time system the process will work similarly to the learning process described above. The model which is the result of training is used, it will use current data on traffic and influences as well as available predictions for influences as input and produce a prediction of the future state of traffic. This will then be used to advise the user about the optimal route to take. When the time of the prediction is reached the data on the actual situation is used as feedback resulting in a constantly adapting system. To realise this Information discovery and combination process within our Traveler Information System research we will take an experimental approach which we will outline in the next section.

4 RAIDO environment

A common complaint in research dealing with the traffic information systems is the lack of data. The problem is both in quantity and quality. The first problem arises from the fact that data collection usually starts at beginning of a project, which will certainly give a shortage of good data at the start of development and even at the end there is not as much as researchers would like. If large quantities

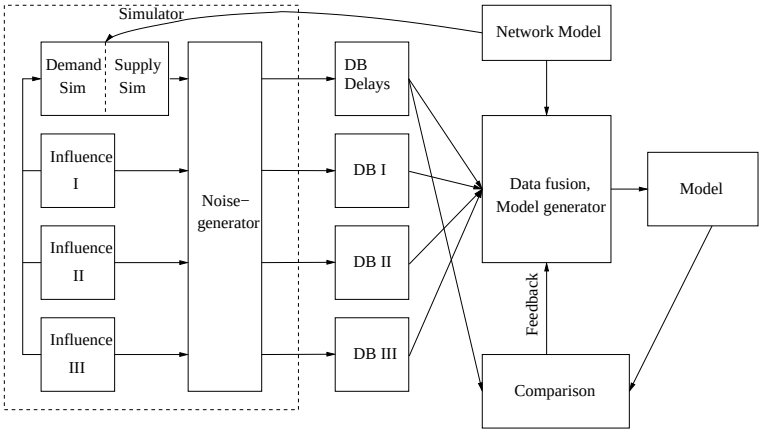

Fig. 7. Raido simulator environment

of data are available it is often in an unstructured form or in a form totally unsuitable to the purposes of the project. Structuring this data requires time and therefor it will not be available for experiments at the beginning of a project. Yet even at the start of development we want to work on the knowledge discovery process, so a way was devised to have data to work with even in the early stages of the project. We have chosen for an experimental approach and are focussing initially on establishing an effective experimental environment employing simulation technology to help gathering a relevant and substantial source of data. The use of simulation in traffic applications is something we have experience with. In a previous project described in [10, 11] we succesfully employed traffic simulation to design and evaluate an automatic debitting system.

For the research described in this paper we have designed an environment in which a simulator generates the data needed for developing the data discovery and combination process. The simulator consists of a traffic simulator, several fabricated data sources for external influences and a noise generator. The traffic simulator is an existing program consisting of two interacting parts:

- supply contains information about the topology of the area being simulated in combination with information on lane closures due to for instance road works or accidents, which is used to determine the capacity of roads.
- demand describes the number of planned trips their origin and destination in so called Origin Destination Matrices

The interaction between supply and demand results in a simulation of traffic patterns. Statistics resulting from the simulation, in our case information about delays, can be stored in a database. The noise generator can be used to pollute the data-set by mutating data or leaving out some data altogether. This allows for experiments determining the minimum quality of data needed both in terms of accuracy and availability. Meanwhile we are also working to establish a source

for and a database of historical data on traffic-jams. Furthermore we are currently investigating the possibility of obtaining data on external influences from the web. There are many different sources on the web which provide information on public events. There are also many web-services which offer weather forecasts. A web-crawler is in development which will obtain this information with a high level of autonomy. We will report our findings on this at a later date. When real data on either delays or influences becomes available it can replace the corresponding simulated data sources.

5 Discussion & Conclusion

In this paper we described our research into the next generation of Traveler Information Systems. In particular we have looked at how we can enable accurate travel advise which is based on prediction of the future state of traffic. It should be clear that we are at the start of our research in this particular area and that much work needs to be done before this is a working system. For instance, problems can arrive in managing computational complexity and exploiting the computational resources needed to run our proposed sytem in real time. We also believe that developing the information discovery and combination process as a component that can be used outside of this domain will add much to the value of our research. In this paper we describe the effect of external influences on traffic and the need to model them using the data available from different sources describing these influences. We presented an information discovery and combination process which can use data on these external influences together with current and historical data on traffic to produce a model which predicts the future state of traffic. Furthermore we have shown the experimental environment we will use for development of this information discovery and combination process. It is our strong belief that the approach described in this paper is the way to reach the accuracy of prediction necessary to make the next generation of Traveler Information System work.

Acknowledgements: The authors would like to thank Peter Schouten and info.nl for providing the traffic data used in this paper.

References

1. Moshe Ben-Akiva, Michel Bierlaire. Didier Burton, Haris N. Koutsopoulos, Rabi Mishalani, *Network State Estimation and Prediction for Real-Time Traffic Management*, in Networks and Spatial Economics 1: 2001 293-318, Kluwer Academic Publishers
2. Everett M. Rogers, *Diffusion of Innovations, fourth edition*, New York:Free Press,1995.
3. Philip Parker, *Aggregate Diffusion Forecasting Models in Marketing: A Critical Review*, International journal of forecasting, 10, 1994, p. 353-380.
4. Geleyn R. Meijer, Jacco Samuels (CMG Unwired Concepts) and Frank Terpstra (University of Amsterdam), *Modeling user acceptance and technology adoption: is*

there a case for value added services?, ITS World Conference on Intelligent Transport Systems, Chigaco october 2002.
5. H. Afsarmanesh, R.G. Belleman, A.S.Z. Belloum, A. Benabdelkader, J.F.J. van den Brand, G.B. Eijkel, A. Frenkel, C. Garita, D.L. Groep, R.M.A. Heeren, Z.W. Hendrikse, L.O. Hertzberger, J.A. Kaandorp, E.C. Kaletas, V. Korkhov, C.T.A.M. de Laat, P.M.A. Sloot, D.Vasunin, A. Visser and H.H. Yakali. *VLAM-G: A Grid-based virtual laboratory*, in Scientific Programming (Special issue on Grid Computing), 10(2), pp. 173-181, Ronald H. Perrott and Boleslaw K. Szymanski editors. IOS Press, 2002, ISSN 1058-9244.
6. Vladimir Estivill-Castro, Ickjai Lee. *Fast spatial clustering with different metrics and in the presence of obstacles*, in Proceedings of the ninth ACM international symposium on Advances in geographic information systems, pp. 142-147, 2001 ACM press, ISBN:1-58113-443-6
7. J. Gehrke, F. Korn, D. Srivastava. *On Computing Correlated Aggregates Over Continual Data Streams*, SIGMOD 2001, Santa Barbara, CA, 13-24.
8. Howard J. Hamilton, Dee Jay Randall. *Data Mining with Calendar Attributes*, In Proc. International Workshop on Temporal, Spatial and Spatio-Temporal. Data Mining, TSDM2000, Lyon, France, pp. 117-132, J. F. Roddick and K. Hornsby, Eds., Springer. Lecture Notes in Artificial Intelligence. 2007.
9. Robert L. Grossman and Mark F. Hornick and Gregor Meyer, *Data mining standards initiatives*, Communications of the ACM, vol 45, nr 8, 2002, pp. 59–61, ACM Press ISSN 0001-0782.
10. Hoekstra, A.G., Meijer, G.R., Hertzberger, L.O., *ADS-SIM: A Generic Simulation Environment to Evaluate and Design Automatic Debitting Systems*. Proceedings 4th World Congress on Intelligent Transport Systems, ICC Berlin, Germany, 21-24 October, 1997.
11. Meijer, G.R., Hertzberger, L.O., Wijk, D.P.van. *Design and Operation of ERP by Simulation*. In Proceedings of the International Conference on Transportation in the Next Millenium , pp 197-222, Singapore , September 9-11, 1998.
12. Eelco Herder and Betsy van Dijk. *Personalized adaptation to device characteristics*, Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, May 2002. Springer LNCS 2347, pp. 598-602