Empirical Analysis of Passenger Trajectories within an Urban Transport Hub

Benjamin Heuer
Hochschule der Medien (HdM)
benjamin.t.heuer@googlemail.com

Heiko Roßnagel
Fraunhofer IAO
heiko.rossnagel@iao.fraunhofer.de

Jan Zibuschka
Fraunhofer IAO
jan.zibuschka@iao.fraunhofer.de

Johannes Maucher
Hochschule der Medien (HdM)
maucher@hdm-stuttgart.de

ABSTRACT

In this contribution we present an analysis of passenger trajectories in an urban transportation hub. We collected an extensive amount of empirical data consisting of both gate and individual stalking observation in the central station of Cologne. Three different data mining algorithms are used to analyze this data, producing both data that may be used as input for simulation frameworks, and, as an aside, visualizations of passenger movements that could be of high interest to transport and emergency managers.

Keywords

Passenger Trajectories, Urban Transportation, Data Mining, Empirical Data

INTRODUCTION

Large transport hubs, like central urban railway stations, are especially vulnerable to accidents or attacks due to the large number of people that can be targeted there, and their accessibility to the public. In situations where traffic peaks, for example before and after large gatherings such as music festivals or popular sports events, the crowds within the station are both especially hard to control, and especially attractive targets (Roßnagel et al. 2011). From an emergency management perspective, this is a very challenging setting, because of the high number of persons that are concentrated within rather small areas during the event (Roßnagel et al. 2010)

In such situations, disaster management systems, especially those supporting warning systems and training activities, have been identified as critical infrastructural investments to mitigate disaster effects (Johnston et al. 2007). Especially during large public events, crowd dynamics play a crucial role (Chertkoff and Kushigian 1999). Agent based simulation can help to identify extreme crowd densities and congestions of areas beyond obvious choke points, which rate among the most important factors that determine the severity of disasters (Chertkoff and Kushigian 1999). These capabilities can be useful for egress routing during an incident, for the design of egress routes, or for post-event analysis (Roßnagel et al. 2010).

However, a simulation is only helpful if its results (largely) align with reality. For such a simulation, it is necessary to have a significant amount of input data gathered under realistic conditions (Hand et al. 2008). However, as human behaviour during critical situations can deviate widely from behaviour that would normally be observed, and at the same time is determined by the nature of the event, the environment, as well as individual factors (Fritz and Marks 1954), the only way to gather truly realistic data sets for our scenario would be to observe a critical incident during a large public event at a transport hub, and even then it would be hard to generalize the results. This approach seems questionable both from an ethics perspective, as we could certainly not cause such an event nor observe it without getting involved, and from a feasibility perspective.

In this contribution, we present an approach for integrating information gathered in a non-critical situation at a specific public event into a system for agent-based simulation of crowds. Our aim is to separate the behavioural aspects of disaster situations, which the system already covered (Roßnagel et al. 2010), from the specifics of the transport hub and concrete public event. We present a set of empirical information consisting of both gate and individual stalking observation, an approach also taken by e.g. (Chang 2002), analyze it using data mining approaches and visualize the results, and finally indicate how the integration with a larger simulation framework is realized. Using the data mining approach probability models are constructed that can serve as input parameters for the agent-based simulation.

The remainder of the paper is structured as follows. We first provide a literature survey of related work and outline the setup of the study describing the algorithms used for analysis and the methods of data collection. Then we present and discuss the results of our analysis and give an outlook on further research, before we conclude our findings.

RELATED WORK

Simulation has long been recognized as an essential tool for emergency management (Pidd et al. 1996). With regard to what kind of simulation is appropriate to use in this field, there is a broad consensus that agent-based simulation offers many advantages over the discrete event-based simulations (Siebers et al. 2010). Consequentially, during the last few years, several agent-based systems have been reported on by researchers and have been made available in the marketplace (Railsback et al. 2006). Examples include (Pan et al. 2007), (Berrou et al. 2007) and (Roßnagel et al. 2010).

One of the main advantages of agent-based simulation models is the decentralization of control (Siebers et al. 2010). Each agent can decide on its next action individually, allowing for emergent behavior (Siebers et al. 2010) and agents with individual capabilities, such as the capacity to receive notifications via specific communication channels (Roßnagel et al. 2010). This requires modeling the behavior of agents with different characteristics under different environmental conditions (Pelechano et al. 2007). This is recognized by researchers and practitioners, who have built generic simulation frameworks that can accommodate for the integration of a wide set of factors, especially considering the trajectories of agents and their speed along those trajectories (Klüpfel et al. 2005). Examples for such frameworks that were recently reported include Legion (Berrou et al. 2007), MAGS (Moulin et al. 2003) and CAST (Roßnagel et al. 2010).

However, there is still a lack of a broad basis of data collected from a variety of environments that are of high relevance to disaster management. There have been some first steps made into that direction by recent publications such as (Klüpfel 2007), (Berrou et al. 2007), who model the behavior of crowds at large public events. This is very close to what we are presenting in this contribution; however, we are concerned with the behavior of crowds at public transport hubs before, during and after large public events. As we could not identify one single algorithm in this area that is considered to be the state of the art, results obtained with several different approaches are provided, two of them based on Markov chains as (Andrade et al. 2006) suggests, albeit for automatic analysis of video material.

STUDY SETUP

In this section we present the setup of our study. We will first present the data mining algorithms that we use for our analysis of passenger trajectories and then describe our method of gathering the empirical data.

Data Mining Algorithms

We use three different data mining algorithms to analyze the empirical data. Each of the algorithms pursues a different objective and provides different results that can be used as input for an agent based simulation or provide relevant insights to predict the behavior of passengers.

Generalized Sequential Patterns (GSP)

The GSP-Algorithm (Srikant and Agrawal 1996) is an algorithm used for sequence mining. It discovers all sequential patterns even those who are not directly consecutive. So if a passenger moves from A to B to C and finally to D the algorithm will be able to identify A to D as a sequential pattern. The GSP Algorithm makes multiple database passes. In the first pass, all single items (1-sequences) are counted. From the frequent items, a set of candidate 2-sequences are formed, and another pass is made to identify their frequency. The frequent 2-sequences are used to generate the candidate 3-sequences, and this process is repeated until no more frequent sequences are found. Due to its ability to identify indirect sequential patterns the algorithm is especially suited to identify the probabilities of which exits the passengers will eventually use based on their position within the train station. While the GSP can be used to predict correlations between locations and therefore is suitable to predict where a passenger will leave the simulation space, it is not capable to predict the actual next step based on the sequence of places already visited by the passenger.

Consecutive Pattern Extraction (CPE)

To address this issue we developed our own algorithm that we call Consecutive Pattern Extraction (CPE) for the analysis of sequential patterns. This algorithm only extracts patterns which are strictly consecutive without gaps. This is essential for creating a conditional probability distribution depending on the places already visited. The algorithm transfers all sequences to a tree. Every node of this tree holds a specific pattern, the frequency of its occurrence and the probability distribution of the places which are able to reach in the next step. The algorithm creates two trees; a prefix-tree and a suffix-tree. In the prefix-tree every child node adds a suffix to the pattern-sequence of the father node. In the suffix-tree every child node adds a prefix to the pattern-sequence of the father node. The trees are created by an incrementally increasing sliding window which slides over each sequence and counts the specific sequence. The probability distribution of a node can easily be computed by considering all child nodes of the node in the prefix-tree. The suffix-tree provides the agent in the simulation with the necessary information for making its choices. As a result the CPE provides the probability distribution for a chosen length of the Markov chain. This information can directly be used as an input for the simulation. A limitation of this algorithm is its high demand for memory, which limits the potential depths of the Markov chains. While this algorithm enables researchers to make statistical projections of passenger behavior in general, it does not allow the clustering of individual sequences into homogenous clusters, because the individual subsequences are aggregated into one single tree.

Group Movement Pattern Mining (GMPMine)

To be able to group the sequences into clusters we used the GMPMine algorithm (Tsai et al. 2011), which has been developed to cluster animal movements tracked via sensor networks. The strength of this algorithm is that it provides a similarity measure to cluster sequences. The algorithm creates a probabilistic suffix tree (PST) for every sequence. This PST represents a Markov model of variable length. Every node in that PST is represented by a string, which is a significant pattern. Additionally every node in the PST holds the conditional probability distribution of all places. Every child node adds exactly one place before the rest of the string but only if the probability distribution differs significantly. These PSTs make it possible to compare a pair of sequences. For a detailed description of this specific algorithm please refer to (Tsai et al. 2011). The next step is to transfer the cluster problem to a graph. In the graph a sequence corresponds to a node and an edge between two nodes means that the similarity of these sequences exceeds a certain threshold. After that the graph is split up with help of the FastCut-algorithm (Har-Peled 2002) into strong connected sub graphs. Each of these sub graphs represents a cluster.

Data Collection

As a basis for our data analysis we collected an extensive amount of empirical data consisting of both gate and individual stalking observation in the central station of Cologne. We performed the data acquisition on two separate occasions: during a particular large event ("Kölner Lichter 2010") and on a normal working day. This allows us to compare the results and to analyze differences in the behavior and the amount of passengers. Figure 1 shows the setup for our measurements including our partitioning of the station into several segments that we labeled based on their location within the station (A1-A4, B1, D1-4, U1). All exits (E1-E14) and service points (S1-S48) were also labeled.

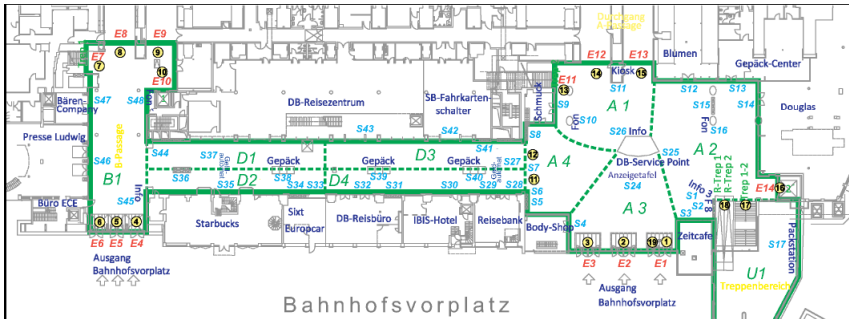

Figure 1: Map of Cologne train station including the segments used for data analysis

During "Kölner Lichter" we counted the number of passengers entering and leaving the station at all exits during a defined time frame of two and a half hours using 10 minute intervals of measurement. Overall we counted 47,196 incoming and 47,328 outgoing passengers during this time frame. On the normal working day

the time frame was extended to 3 hours and we counted about 39,400 passengers entering and leaving the train station during this time frame. Figure 2 shows the distribution of incoming and outgoing passengers during Kölner Lichter. The breadth of the arrows indicates the amount of passengers using the exit.

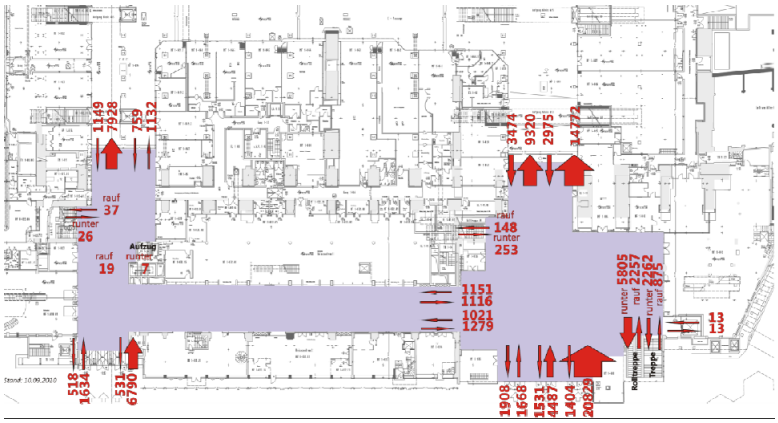

Figure 2: Incoming and outgoing passengers during Kölner Lichter 2010

In order to acquire information about the particular plans of passengers, we used the method of single person pursuit. The researchers picked random passengers to follow and to record their activities during their stay in the central train station. Using this method we recorded the activities and movements of 1042 passengers during Kölner Lichter and of 1251 during the normal working day.

EMPIRICAL RESULTS

Results of the GSP-algorithm

The focus on this algorithm was to generate the distribution of used exits in dependence of a specific current location. Therefore, all possible 2-tupel combinations were generated and we selected those combinations where the second element was an exit. So for every area a chart could be generated for both days.

A-Passage (Main hall of the station)

The A- Passage (see Figure 1) shows quite similar results for both days. Passengers coming from A4 strongly tend to leave the area through exit E12 and the ones coming from A2 are likely to exit through E13. Both exits are equally frequented while exit E11 was remarkably little frequented. Looking at the exits E1, E2, E3 the results are quite similar. Passengers coming from A4 leave the hall through the left exit E3 und coming from A2 through the right exit E1. All three exits are equally frequented. Exits E18 and E19 also show a slightly higher frequency. The other exits are frequented negligible. Based on these results we are very confident to state, that persons in the A-passage mostly use the exits to the railways or to the "Kölner Domplatte".

B- Passage (Side hall of the station)

The B- Passage also shows similar results for both days. Of the three exits to the railways (E7, E8, E9), E8 is disproportionately frequented. Of the three exits to the "Domplatte", E4 is frequented about twice as often as E6.

D- Passage (Connection between the halls)

The results show that the D- Passage is very quiet on both days and passengers tend to leave the hall through the same area they entered.

U- Passage (Connection to the subway)

As in the other areas there was no significant difference in the results of both days. Most of the persons coming from the subway-station leave the railway station through exit E13. Passengers going to the subway tend to take exit E18. At the workday there is also a higher frequentation of exit E20, which leads to the "Domplatte". Figure 3 shows the distribution of passengers to exits for the area U1.

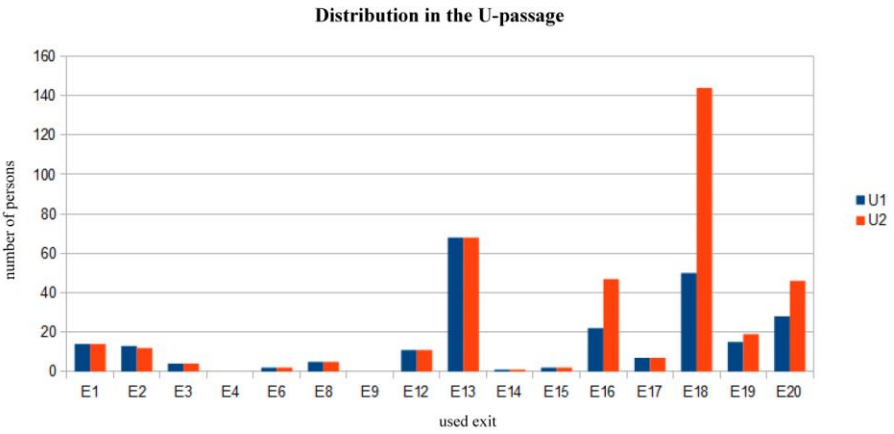

Figure 3: Distribution of passengers to exits in the area U1

Result of the CPE-algorithm

The CPE-Algorithm enables to generate a probability distribution dependent on how many previous visited states are considered. This data can find direct access into the agent simulation to calibrate the behaviour of the single agents corresponding to the empirical data. To visualize the results a modell was constructed which represents a probabilistic automat. The states of the model are represented by the places A, B, C, D, S1-S46, E1-E20 (Figure 1) and the transition between two places is the likelihood for going from one place to another. Using the visualization we can conclude:

- Both days show a quite similar likelihoods
- Remarkably difference in both days for exit E4 – workday(5%) & Kölner Lichter (20%)
- Service Points except of the infopoint S24 are rarely frequented
- Likelihood to go from B-passage or A-passage to the D-passage is quite low
- People tend to use the escalator instead of stairs going to/coming from the U-passage
- Observations of used exits correspond to those from GSP-algorithm
- Many exits are extremely low frequented (E5,E7, E9-E11, E15, E17, E19)

Figure 4 shows an example of a visualized modell of the B-Passage. For reasons of clarity the main areas were colored red, the service-points were colored green and the exits/entrances were colored yellow. To further enhance the clarity trivial transitions were removed. Trivial transisions are fixed transitions where no other transition is possible. For example the transition from E1 to A3 (see Figure 1) is such a trivial transition. Figure 4 clearly shows the behaviour of persons in the B-Passage. As already mentioned service points are negligible visited. Also few persons tend to go into the D-Passage. Most people leave the B-Passage through exit E8.

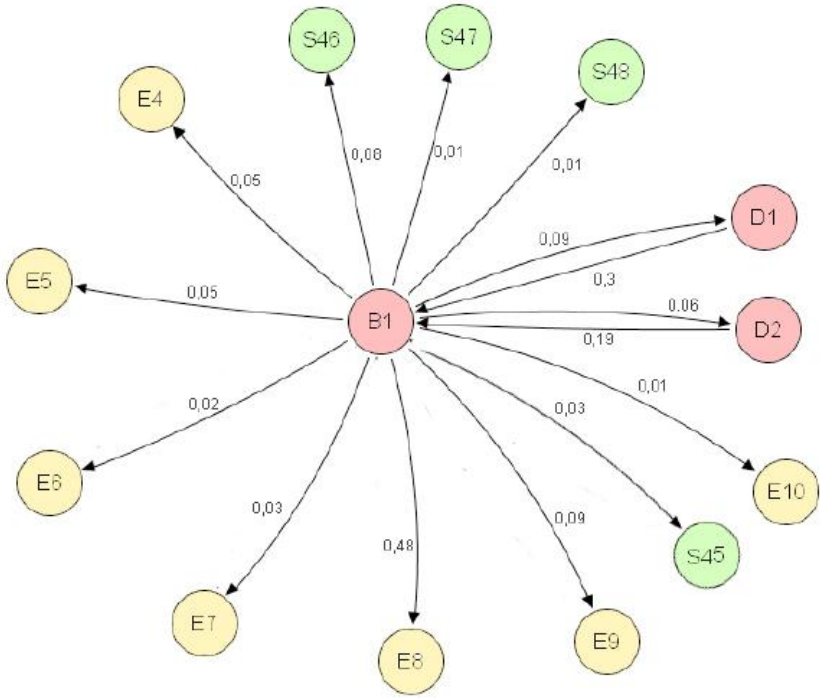

Figure 4: Example of a markov modell of the B-Passage

Result of the GMPMine-algorithm

The results from the clustering depend on the choice of the given parameters of the GMPMine-algorithm. The parameters influence the number and size of resulting clusters. To test the quality of resulting clusters a sensivity analysis with varying parameter settings was done. The results of this sensitivity analysis show that dependent on the parameter settings, the number and size of the clusters can vary quite a bit but the resulting clusters remained very stable over the different runs. Figure 5 illustrates the resulting clusters.

Cluster 1 shows a quite stable cluster in the B-Passage. This cluster primary consists of patterns from E8 to E4-E6 and vice versa. Another frequent pattern is the one from E7 to E8. This cluster clearly represents a tendency from railways to exits.

Cluster 2 contains patterns from E1 over A2 to the railways via E13. Passengers in this cluster often made a stop at service point S24, which represents an information point about train schedules. This pattern is completely reasonable as it shows a way on which passengers enter the railway station, inform themselves about time and platform of departing trains and head towards the railways.

Cluster 3 consists of patterns showing ways from the A-Passage to the U-Passage. It can be seen, that persons coming from the railways via E13 leave the railway station immediately through E1 or they take the longer way into the U-Passage to leave the railway station through E20.

Cluster 4 appeared in all evaluations and represents movement in the subway station. There are many tracks from E20 to the railways over E16 and E18. These are passengers which used the lower entrance to directly access the subway station. The cluster also includes movement from E16 to E18.

Cluster 5 shows a strong tendency from the entrances E1, E2 or E3 via A4 to the railways via exit E12 or E13.

The cluster study of the workday shows quite similar results in dependence of the chosen parameters. This algorithm also confirms the assumption, that the behavior of the passengers at the cologne railway station is independent of the external conditions tested in our study.

Originally the GMPMine-Algorithm has been developed to track animal movement. The animal movement was tracked in a certain terrain via senor networks which leads to a certain amount of movement repetitions. The concerns using this algorithm were, that passengers moving through the railway station usally do not move in repetitive patterns so the resulting PSTs might not been suitable for a comparison between sequences. The sensivity analysis with different parameters however resulted in relatively constant clusters. Dependent on the chosen parameters (especially similarity threshold and Markov-Order) the count of resulting clusters varied from one to five. The quality of the resulting clusters however remained relatively undependent of the chosen

parameters. A great disadvantage of the algorithm is that dependent on the chosen parameters the runtime increased dramatically which is a problem of extracting the minimum cuts out of the graph to extract the string subgraphs. However there is potential for decreasing runtime by parallizing this part of GMPMine.

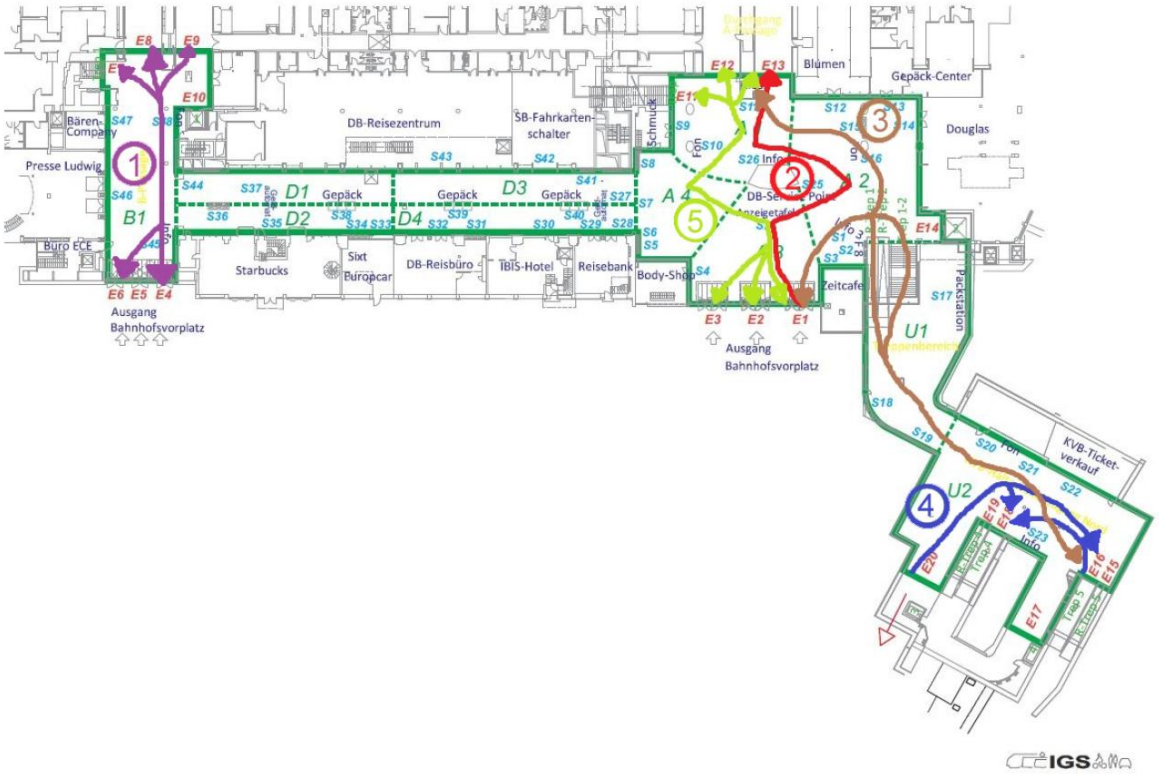

Figure 5: Resulting clusters from theGMPMine algorithm

DISCUSSION AND OUTLOOK

Simulation and analysis of quantitative data can provide an important contribution to the management of emergencies, as it could help emergency managers to make more informed decisions and therefore improve operational effectiveness and response times. It could aid emergency planners in their development of contingency plans and thus help to save human lives and protect critical resources.

In this contribution we presented the analysis of passenger trajectories within one specific public transport hub during large public events. The results of our clustering may be applicable to transport hubs with very similar layouts and services, however, (Gelhausen et al., 2008) found significant differences in passenger preferences for passengers travelling to the airport and passengers at the airport in the Cologne region, demonstrating a significant complexity of the underlying behavioral drivers. We illustrated several approaches to generalize from the available data, producing both data that may be used as input for simulation frameworks, and, as an aside, visualizations of passenger movements that proved to be of high interest to transport and emergency managers. The results of the GSP-algorithm provide a distribution of the use of exits based on the position of the passengers. This could help to identify exits that are less frequent. This information could be helpful for dedicated cell broadcast messages during passenger egress (Roßnagel et al. 2011). The CPE results in a conditional probability distribution for the next step depending on the sectors already visited by the passenger. This information can directly be used as an input for the simulation. The cluster analysis of the GMPMine-algorithm provides clusters of passenger movements, which might also form a basis for later transfer to other scenarios. As there is no systematic approach for selection of the appropriate data mining algorithms for all situations, and data mining tools are constantly evolving (Kantardzic, 2011), the application of the presented modern algorithms should be considered explorative, and their applicability is also a result of this contribution. This result should have a wider applicability than the clustering results for our specific setting.

However, an evaluation of the value of this information for those stakeholders has not been done as of yet. The visualizations have also not yet been optimized for that purpose, e.g. strong links are not emphasized yet. Also, due to the sheer volume of results that were produced, we could only give the coarsest overview of passenger

movements within the station in this contribution. Another open issue for future research is to find out how the achieved results for the cologne train station generalize to other public hotspots.

CONCLUSION

In this contribution we analyzed empirical data on passenger trajectories gathered during a large event and a normal working day in an urban transportation hub. Three different data mining algorithms were used for this analysis resulting in probability models that can serve as input parameters for agent-based simulation. Furthermore, the analysis provided visualizations of which exits are used by passengers based on their cluster position, probability distributions of likely next steps based on previous actions and clusters of passenger movements. These results could be of interest to transport providers and emergency managers for the preparation of egress plans or emergencies in general. In addition, this information is of high value for simulation experts for the calibration of their models and may contribute in the long run to a more general understanding of passenger movements, contributing to the scientific knowledge base.

ACKNOWLEDGMENTS

Research was performed in the program “Research and civil protection” of the German Authority for Research and Education (BMBF), as part of a high-tech strategy of the German government, for the funded research project VeRSiert. We gladly acknowledge the contribution of our colleagues who participated in the measurements at Cologne Station.

REFERENCES

1. Andrade, E., Blunsden, S. and Fisher, R. (2006) Hidden Markov Models for Optical Flow Analysis in Crowds, *ICPR 2006*, Hong Kong, IEEE, 460-463.
2. Berrou, J., Beecham, J., Quaglia, P., Kagarlis, M. and Gerodimos, A. (2007) Calibration and validation of the Legion simulation model using empirical data, Pedestrian and Evacuation Dynamics 2005, Part 3, Springer, Berlin Heidelberg, 167-181.
3. Chang, D. (2002) Spatial Choice and Preference in Multilevel Movement Networks, Environment and Behavior, 34, 5, 582 -615.
4. Chertkoff, J. M. and Kushigian, R. H. (1999) Don’t Panic: the Psychology of Emergency Egress and Ingress, Praeger Frederick, Westport.
5. Fritz, C. E. and Marks, E. S. (1954) The NORC Studies of Human Behavior in Disaster, Journal of Social Issues, 10, 3, 26-41.
6. Gelhausen, M. C., Berster, P., Wilken, D. (2008) Airport choice in Germany and the impact of high-speed intercity train access: The case of the Cologne region, Journal of Airport Management, 2, 4, 355-370.
7. Hand, J. W., Crawley, D. B., Donn, M. and Lawrie, L. K. (2008) Improving non-geometric data available to simulation programs, Building and Environment, 43, 4, 674-685.
8. Har-Peled, S. (2002) Minimum Cut in a Graph: 497 - Randomized Algorithms, http://valis.cs.uiuc.edu/~sariel/teach/2002/a/notes/min_cut.pdf, 2002-09-03.
9. Johnston, D., Becker, J., Gregg, C., Houghton, B., Paton, D., Leonard, G. and Garside, R. (2007) Developing warning and disaster response capacity in the tourism sector in coastal Washington, USA, Disaster Prevention and Management, 16, 2, 210-216.
10. Kantardzic, M. (2011) Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, 2011.
11. Klüpfel, H. (2007) The simulation of crowd dynamics at very large events: Calibration, empirical data, and validation, Pedestrian and Evacuation Dynamics 2005, Part 3, Springer, Berlin Heidelberg, 285-296.
12. Klüpfel, H., Schreckenberg, M. and Meyer-könig, T. (2005) Models for Crowd Movement and Egress Simulation, Traffic and Granular Flow ’03 Part 4, Springer, Berlin Heidelberg, 357-372.
13. Moulin, B., Chaker, W., Perron, J., Pelletier, P., Hogan, J. and Gbei, E. (2003) MAGS Project: Multi-agent GeoSimulation and Crowd Simulation, COSIT 2003, Springer, Berlin Heidelberg, 151-168.
14. Pan, X., Han, C. S., Dauber, K. and Law, K. H. (2007) A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations, AI & Society, 22, 2, 113-132.
15. Pelechano, N., Allbeck, J. M. and Badler, N. I. (2007) Controlling individual agents in high-density crowd simulation, Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation, San Diego, CA, USA, 99-108.

16. Pidd, M., de Silva, F. N. and Eglese, R. W. (1996) A simulation model for emergency evacuation, European Journal of Operational Research, 90, 3, 413-419.
17. Railsback, S. F., Lytinen, S. L. and Jackson, S. K. (2006) Agent-based Simulation Platforms: Review and Development Recommendations, SIMULATION, 82, 9, 609-623.
18. Roßnagel, H., Zibuschka, J. and Junker, O. (2010) Agent-Based Simulation for Evaluation of an Mobile Emergency Management System, Proceedings of the 16th Americas Conference on Information Systems, August 12-15, Lima, Peru.
19. Roßnagel, H., Zibuschka, J. and Junker, O. (2011) On the effectiveness of mobile service notifications for passenger egress during large public events, Proceedings of the 8th International ISCRAM Conference, Lisbon, Portugal.
20. Siebers, P. O., Macal, C. M., Garnett, J., Buxton, D. and Pidd, M. (2010) Discrete-Event Simulation is Dead, Long Live Agent-Based Simulation!, Journal of Simulation, 4, 3, 204-210.
21. Srikant, R. and Agrawal, R. (1996) Mining Sequential Patterns: Generalizations and Performance Improvements, Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology.
22. Tsai, H., Yang, D. and Chen, M. (2011) Mining Group Movement Patterns for tracking Moving Objects Efficiently, IEEE Transactions in Knowledge and Data En*gineering*, 23, 2, 266-281.