The Behaviour Assessment Model for the Analysis and Evaluation of Pervasive Services

Bernhard Klein, Ivan Pretel, Ulf-Dietrich Reips,
Ana B. Lago, and Diego Lopez-de-Ipiña

DeustoTech –Deusto Institute of Technology, Universidad de Deusto
Avenida de las Universidades 24, 48007, Bilbao, Spain
{bernhard.klein,ivan.pretel,reips,
anabelen.lago,dipina}@deusto.es

Abstract. Several mobile acceptance models exist today that focus on user interface handling and usage frequency evaluation. Since mobile applications reach much deeper into everyday life, it is however important to better consider user behaviour for the service evaluation. In this paper we introduce the Behaviour Assessment Model (BAM), which is designed to gaining insights about how well services enable, enhance and replace human activities. More specifically, the basic columns of the evaluation framework concentrate on (1) service actuation in relation to the current user context, (2) the balance between service usage effort and benefit, and (3) the degree to which community knowledge can be exploited. The evaluation is guided by a process model that specifies individual steps of data capturing, aggregation, and final assessment. The BAM helps to gain stronger insights regarding characteristic usage hotspots, frequent usage patterns, and leveraging of networking effects showing more realistically the strengths and weaknesses of mobile services.

Keywords: Mobile services, user acceptance, human-computer interaction, technological acceptance model, log data analysis, service design, living lab.

1 Introduction

User acceptance in field trials is still mostly evaluated through questionnaires and focus interviews. Mobile applications are, however, much stronger related to real mobile behaviour as people carry their devices with them. Because of the dependency of mobile applications' functionalities on the user situation answers to general questions about the application can often not easily be given.

A relative new approach for mobile services is the observation of application usage through data loggers. A data logger records application events or errors jointly with other usage or system related metadata. To support daily activities successfully, mobile applications should not interrupt the activities, provide a reasonable interaction/benefit ratio to the user, and provide community leveraging beyond exploitation of personal experience. Analysing usage hotspots, usage frequency and usage type allow researchers to speculate about potential strengths, weaknesses or even problems for the surveyed service.

Á. Rocha et al. (Eds.): *Advances in Information Systems and Technologies*, AISC 206, pp. 1129–1140.
DOI: 10.1007/978-3-642-36981-0_107

In this work, we present a behaviour assessment framework that describes the systematic collection of behaviour data and guides researchers in their log data analysis. With such an analysis approach researchers can gain more insights about first and long term service impressions, acceptances issues correlated with the user experience and the success of subsequent product evolvement steps.

The article is structured as follows. The next section discusses related works. Section 3 describes the method how to capture, aggregate, and represent data. In Section 4 the behaviour assessment model is defined. A preliminary case study is given in Section 5. Section 6 concludes the article.

2 Related Work

In order to perform a study focused on mobile services the first step is to compare, analyse and find the differences and connections between existing data loggers, concepts and conclusions related to the mobile services evaluation field.

Lab-based evaluation frameworks log information in a controlled environment using specific devices and specific users. The main advantages of the lab-based frameworks are the highly controllable environment and the collection of data, which is cheap and easy. However, the context, which is the most influential factor in the mobile services field, is not considered and it can hardly be simulated. Many simulation tools produce highly inaccurate results because of the context. Furthermore, several agents also alter the results of user experiments. The experts who lead the experiment and the tasks performed by the users can not only alter the execution of experiments but also evoke situations that would never happen in real environments. The users may also add biased results during the execution of the experiments [1] because they suffer several problems such as test-anxiety [1]: during the task performance the highly test-anxious person divides his attention between self-relevant and task-relevant variables; due to the self-focussed attention the user of the mobile service may not show real behaviour. Further, in many tasks such as phone calls, it would be subjectively annoying for many users to be in a room with observing researchers.

On the other hand the field-based evaluation frameworks (see Table 1) capture information in real environments. They commonly use added cameras and human observers to capture information from the interactions. Furthermore, this kind of framework tries to bring the lab to the field. For example, the *Usertesting* platform [2] not only brings methods like the think-aloud verbal protocol but also records the user's feedback with a webcam; finally it reproduces the interaction again enabling the annotations during it. Using this kind of techniques means that although the task is performed in real environment, it is changed and consequently, the interaction altered. Another tool related to *Usertesting* is the *Morae Observer* [3] tool. It captures all the interaction data and indexes it to one master timeline for instant retrieval and analysis; it generates graphs of usability metrics. Both tools are focused on the interaction because they are centred on capture of screen interaction and the user's feedback through filming the face or recording comments. Another group of tools such as *ContextPhone* [4] and *RECON* [5] are focused on the context capture. They capture

the surrounding environment through mobile sensors. This capturing technique retrieves a lot of real data without influencing the interaction but the user's feedback is lost. In order to fill the lack of the user's feedback other tools like *MyExperience* [6] and *SocioXensor* [7] use techniques like self-reports, surveys and interviews mixed with the context capture. These tools are quite powerful and flexible because the user has at any time the complete control about when participate in an application acceptance survey. In case, he has been interrupted in the survey he can resume it to a later point of time.

Table 1. Properties of the logging tools

Tool	Capture techniques	Data	Report
Usertesting	Screen, webcam and microphone	Interaction, user information and user's feedback	Reproduce the screen interaction
Morae Observer	Screen, webcam and microphone, observer	Interaction, user information and user's feedback	Reproduce the interactions and calculate graphs
ContextPhone	Mobile sensing and interaction event logging	Interaction, device status and environment	Mobility patterns detection
RECON	Interaction event logging and mobile sensing	Interaction, device status, user information, user's feedback, and environment	Trace Data analysis Engine
MyExperience	Wearable hardware sensing, mobile sensing, audio recording and user surveys	Interaction, device status and environment	Performance analysis, SMS usage and mobility analysis
SocioXensor	Interaction event logging, survey, interview	Interaction, user, device status and environment	SQL database

To sum up, to acquire valid interaction data about mobile services, it is essential to capture objective information to solve questions like when, where, how long, etc. users are really interacting with a service. These questions can hardly be determined with a lab-based framework. The field-based evaluation frameworks can provide deeper and more objective information, but the added agents such as cameras and invasive evaluation methods (e.g. think-aloud verbal protocols) have to be removed. In order to do so, the best way to capture interaction data is by registering information through a mobile device using a tiny capture tool. This tool should log the context via the built-in mobile sensors and logging the key interaction events.

3 Mobile Service Assessment through Behaviour Analysis

A framework for automatically logging and processing data for evaluation has been developed. In the following we briefly explain the different behaviour capturing and aggregation phases and the architectural requirements.

3.1 Data Logging and Aggregation Overview

As can be seen in Fig. 1 the framework distinguishes four main phases:

1. *Data Capture*: A data logger component installed separately on the mobile device records event and error data triggered by the mobile service. Examples for logging data are: service start and stop times, UI events e.g. buttons pressed, screen transitions, any changes in settings and erroneous data entries, exceptions and any unexpected system behaviour. These data are complemented with additional user contexts (e.g. provider and subscriber data), service information (e.g. queries/results, content data, screen stay duration) and device contexts (e.g. location data) for further evaluation.
2. *Transfer Protocol*: Logging data is periodically (e.g. daily) transferred to an analysis component hosted on the Internet. To minimize the influence on mobile service performance the transfer process is only started if the mobile device remains in an idle execution state.
3. *Data Aggregation*: The analysis component parses the incoming logging data and interprets the raw data log format with a parser. A filter process removes out-of-bound values, spatio-temporal inconsistencies, and entries that do not conform to preset criteria. Following this filtering step the log data are aggregated through clustering analysis.
4. *Data Visualization*: From the results tables, graphs and diagrams are generated for the researcher. Furthermore, the entire log is automatically annotated so that each entry is written out for human readability and annotated to get basic derived information such as duration and transitions.

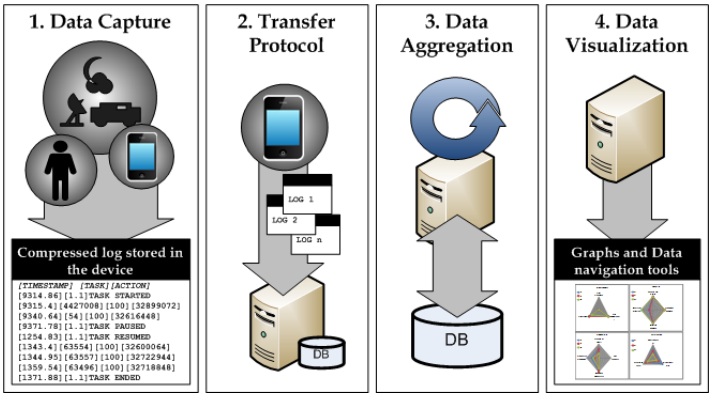

Fig. 1. Graphical description of the process

3.2 System Architecture

The Neurona evaluation framework [8] was extended to meet the BAM requirements. This platform shown in Fig. 2 is based on three main components: the mobile device component, the connector component, and the analysis server component.

The Mobile Device component is software installed in the user's mobile and logs user interactions; it is formed by the Logger/App interface, Logger Module and Context Information Module. The Logger/App interface is a tiny software library used to send interaction events to the logger module. The logger module stores the interaction data and shows brief questionnaires about the interaction experience to capture the user's feedback; these questionnaires are shown at the end of the interaction to not disturb the experience. Another element is the context information module, which provides context information acquired from the built-in mobile sensors and the mobile Operative System.

The Analysis Server component is hosted in a web server; this component is formed by the Data Aggregation Module, the Visualization Module, the Applications Manager and the Usergroup Administration. The Data Aggregation Module receives logged data and calculates normalized information to store it in the system database. The expert who wants to check the normalized information can do it using the Visualization Module; which shows advanced graphs. The Applications Manager enables the expert to register into the system, update and remotely configure prototype applications. The Usergroup Administration module registers users and devices, assigns applications and exposes several administration options related to the relations between users, applications and experts.

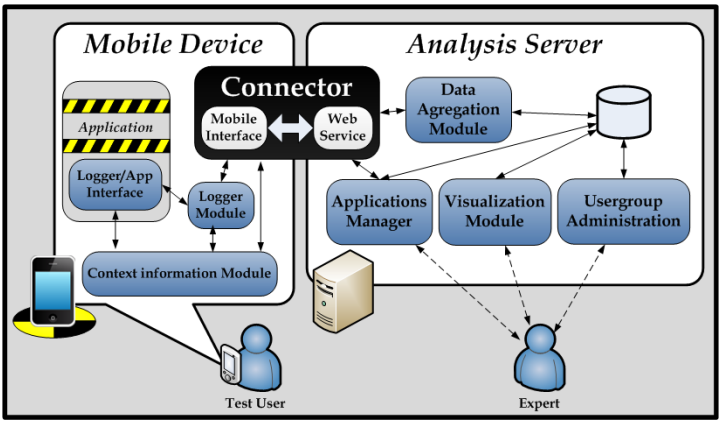

Fig. 2. System architecture

Finally, the connector between the explained elements transfers the logged information generated by the Mobile Device component to the Analysis Server component. It is divided in two main elements: the mobile interface and the server web service. Basically the mobile interface checks the state of the device and if the user is not interacting with the device it sends logged data to the web service hosted in the Analysis server. To minimize the required transfer bandwidth logging data is encoded in memory saving format and decoded later to a human readable format when the logging data has been received by the Analysis Server.

4 The Behaviour Assessment Model

A proven performance assessment method considering concurrent aspects has been the Balanced Scoreboard (BSC) approach. Aligning each of the dimensions systematically helps get a better impression about different influencing factors.

4.1 Dimensions of the Behaviour Assessment Model

The BAM is based on balanced set of behaviour categories which are orthogonal:

- *Planned and spontaneous execution scenarios*: According to Ajzen [9] people in unfamiliar situations often prefer to plan their activities, whereas people in familiar environments rely on their capabilities and thus react more spontaneously. Adequate mobile services have to support both scenarios; offering a remote and location based execution of their services (see Colbert [10]).
- *Service actuation and service interaction*: It is generally accepted that a seamless integration of mobile services in daily life is only given if mobile services raise users' attention in periods where the user is not interrupted, if the interaction efforts keeps a reasonable balance with it intended service benefit.
- *Central provision and community networking*: Mobile services targeting a broad proportion of the audience are better provided through a central provider. However, with peer-to-peer infrastructures people can also provide services to smaller user groups e.g. friend group or only provide them for a short time frame.

This leads to following six different dimensions illustrated in Fig. 3:

1. *Remote activity discovery*: This dimension is based on the categories Planned Execution Scenario and Service Actuation. In order to fulfil end-users need to plan activities ahead of a trip; users require the capability to explore the service offer according to given properties. The retrieval quality depends on the query power e.g. different search concepts and the query success rate. An example is a map based discovery tool, which retrieves services according to locations selected on a map.
2. *Situation-aware activity recommendations*: The dimension founds on the categories Spontaneous Execution and Service Actuation. As mobile services are much stronger correlated with the daily life of end-users an important requirement is to raise their attention to an adequate service offer in a seamingless manner. A successful implementation depends on the reasoning power (that compare the current users' context and the intended service context) and the number of directly consumed services (reasoning success).
3. *Mobile activity creation*: The categories Planned Execution and Service Interaction define this dimension. Complex mobile services require often too much knowledge from the user to execute them easily on the spot. Therefore, services should offer any type of service creation, personalization or reservation functionality so that they can be consumed better in time constrained situations. The editing complexity and the service content quality are important indicator examples to determine this dimension.

4. *On-the-spot activity support*: The dimension is constructed through the categories Spontaneous Execution and Service Interaction. Since users on the move often follow other real-world activities it is important that the attention needed to execute the service is kept to an absolute minimum. The navigation complexity (effort) and the quality of the content provided by the service are important indicator examples.
5. *General platform activity services*: This dimension stems from the categories Planned Execution and Central Provision. All general service aspects influencing the provision quality e.g. power consumption and error handling account for this dimension.
6. *User-created activity services*: This dimension is founded on the category Spontaneous Execution and Community Networking. Tools that consider community behaviour can help in structuring the knowledge space further and lead to more transparency in the community. Examples are best-of ranking lists, member reputation lists and content recommender systems. For instance car sharing opportunities can be more easily evaluated by users and improve their selection. Suitable example indicators are the lurker ratio (active community participation) and the degree of community transparency achieved with previously mentioned community services.

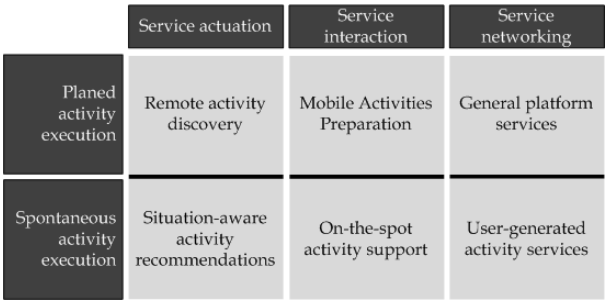

Fig. 3. Dimensions of the behaviour model

4.2 The Balanced Scoreboard Assessment Approach

These six dimensions focus on realistic service usage. This emphasises the valuation of a service by the way how end-users apply services to solve given problems. Such behaviour patterns have the potential to tell us about underlying reasons why specific service fail or become well accepted. Recording such behaviourally relevant data also allow the emulation of service usage in respect to given user´s context. Both aspects are important for developers to continuously improve the service. According to the BSC approach, the intention is to find a few aggregated indicators that quantify a given dimension. The indicator must meet the requirements of reasonability and measurability. A general problem of social surveys is to translate the indicators into precise measures. The abstract classes of measurement types, correspond hereby with different event and error logging data types. To achieve comparability between different numerical scales of measurements e.g. an event/error frequency scale, a

function has to be defined which maps selected scale areas on specific quality rating values. Since humans perceive the influence of various indicators for a given dimension differently, weight coefficients are used to balance the influence of individual indicators. Both mapping function properties and weight coefficients can be obtained through a profiling questionnaire prior to the field trials.

Finally, the results of an analysis and evaluation are typically held in a spreadsheet for detailed analysis and visualised by a radar chart for a summarised representation (see Fig. 4). For visualisation by a radar chart, the six dimensions are equally arranged. The scaling is adapted appropriately according to the distribution of the measurement results with its positive orientation towards the origin. For a better visualisation of the consequences of the results, each scale can be subdivided in fulfilled (positive centre areas), and not fulfilled (negative edge areas).

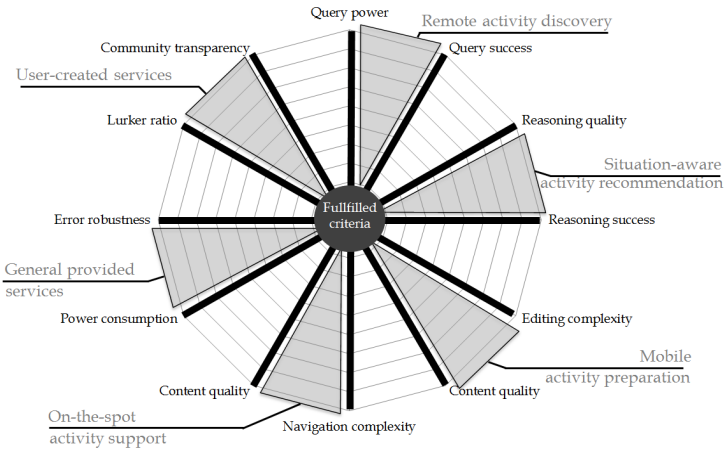

Fig. 4. Visualization of the behaviour model with six dimensions (grey colour) and example indicators (black colour)

5 Case Study of the MUGGES System

Mobile User Generated Geo Services (MUGGES) is a European research project with the goal of evaluating peer-to-peer service concepts based on Global Positioning Systems for mobile phones. MUGGES provides an infrastructure to create, publish, provide and consume mobile micro-services directly from mobile devices (see Klein et al. [11]). As part of the project four application prototypes were developed, which allow the description and sharing of places and routes between users. A field trial was conducted with early adopters in real environments. In the following we will demonstrate how the BAM framework can help to identify benefits and best practices for MUGGES type of applications.

5.1 The Assessment Process

Applying the BAM analysis technique requires specific preparation steps. These include the definition of indicators for each dimension, correlating them with available logging data, appropriate balancing of these measurements with weight factors, the execution of field trials and representing the results. In the following the assessment process is explained in more detail for each phase (see Fig. 5):

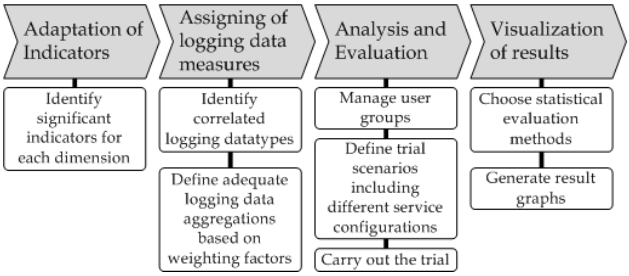

Fig. 5. Different phases of the assessment process and correlated activities

- *Adaptation of indicator structure*. First, adequate indicators have to be identified which represent a specific dimension of the BAM model. Since this is an intuitive process it is recommended to develop this in collaborative process among developers and potential end-users. The design is completed with a proof of plausibility of the defined criteria.
- *Assignment of logging data types*. In this phase, logging data types are correlated with the indicators. In order to achieve a balanced view of all indicators and logging data types weighting factors are applied. Questionnaires can be used to determine the individual importance of each dimension and corresponding logging data types to modify the weight factors accordingly.
- *Analysis and evaluation*. In order to learn more about the characteristic service behaviour it is recommended to conduct a series of experiments. Goals for every single indicator should be determined before the actual start of the field trials, in order to compare these to the empirical results. Then, initial and long-term service usage should be compared in order to identify entry barriers. In a subsequent step, the logging data of trial users shall be clustered according to the technical diffusion model from Rogers [12] to see how technical experience can influence mobile service usage. By comparing innovators, early adopters, early majority, late majority and laggards a reasonable priority list of future service modifications can be determined. Finally it is also important to analyse logging data from different trials in order to see to what extend applied service modifications have lead to an acceptance improvement.
- *Visualisation of evaluation results*. The results of the test group are analysed and evaluated with statistical methods and visualised according to the radar graph approach.

Generally an iterative evaluation approach is recommended starting from early prototypes up to the final mobile service. In order to compare the development

progress it is important not to vary the measurement criteria. It is assumed that the explanatory power of the BAM model increases with a stronger concretization of the mobile service during the development cycle.

5.2 Visual Evaluation of the MUGGES System

We demonstrate the advantages of the BAM with a small example based on the field trial executed for the MUGGES project. In this field trial logging data from 30 potential end-users have been collected during a 2 week period. Each study participant was given a mobile phone with the MUGGES software installed. The users were given specific tasks e.g. describing their favourite shopping route or leisure places. The connector component transfers periodically event data to the analysis server for further evaluation.

For the dimension remote activity discovery and situation-aware activity recommendation query power (measured as average number of applied search keywords or average distance of user/service location at the query time) and query success (measured as average search-consumer ratio) is relevant indicators. The dimension mobile activity preparation, on-the-spot activity support is determined by the indicators editing/navigation complexity (measured through the average screen stay duration per service) and service content quality (measured by the average content length, average number of comments and average update duration per service). The dimension General Platform Services is defined by the indicators power consumption (measured as consumed energy units per day) and error recovery quality (reciprocal number of occurred errors per day). And finally, the dimension user-created activity services are defined by the lurker ratio (measured as provider-consumer ratio of a consumed service) and the consumed service quality (measured through average rating of consumed services).

The MUGGES system has been generally be well accepted as the average mugglet creation rate dropped only insignificantly after the trial kick-off and stayed roughly at about 60 mugglets created per day. Still the MUGGES infrastructure revealed some weaknesses. Applying the BAM approach (with a rating range from 0 – very good till 3 very bad) a service provider can come, for example, to the following simplified conclusions concerning the following dimensions illustrated by Fig. 6:

- *Remote activity discovery*: With increasingly more created mugglets users applied more sophisticated search approaches (from simple template, keyword-based and map-based search) to compensate the small screen size. The discovery function seem to work well for the majority of the trial users (rating 1.5).
- *Situation-aware activity recommendation*: The overall distance between the mugglet location and the trial user has been quite far (up to 1 km). Besides the sparse distribution of the mugglets another reason has been the bad performance of the location technology. Provider could conclude that the recommendation service is not sufficient (rating 2.5) for the current spontaneous usage scenario.
- *Mobile activity preparation*: The mugglet creation process took a lot of time, not short enough to create mugglets on-the-go. People compensated this by

distributing the creation process in several phases. The mobile activity preparation is not sufficient (rating 3.0) in the current development stage.

- *On-the-spot activity support*: The mugglets in general have high information intensity for the user, as they come with a environment map, text descriptions, comments and photos. Above that, the real-time notification feature helped people to stay up-to-date. Mugglets thus have been very useful (rating 1).
- *General provider services*: The peer-to-peer service sharing approach has lead to an high power consumption and the error rate has been quite high. Service provider may conclude that device-to-device sharing is a bad option (rating 2.5) and moving mugglet sharing into the Internet cloud may be a better option.
- *User-created services*: MUGGES usage has been high since users could create their own personal service based on the offered service templates. Especially in later stages during the project service ratings have been found very useful (rating 1) to identify popular services.

The radar graph shows some important weaknesses. Recommender systems, the mugglet creation process and the provider infrastructure still make an everyday usage difficult (see Fig. 6). Comparing these logging data results with the questionnaires conducted after the trial backs these findings. But more importantly, user perceptions were not always clear enough to pinpoint the exact problems with the MUGGES infrastructure. The evaluation with the BAM is more differentiated and considers some critical aspects that influence the acceptance of this mobile service significantly.

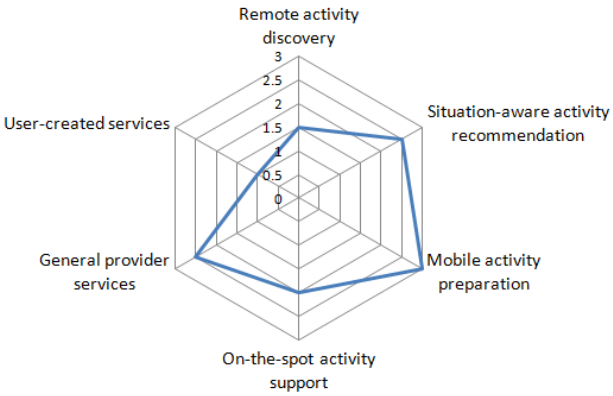

Fig. 6. Evaluation of the MUGGES System

6 Conclusions and Future Work

We introduced the BAM as an instrument for the analysis and evaluation of the user acceptance for mobile services. The BAM is characterised by a structure that helps to identify systematically a balanced set of important, individually measurable and

independent acceptance criteria. The application of the BAM is guided by a process model that supports all phases from the development of acceptance criteria over the measurement of relevant indicators to the evaluation and visualisation of the derived results. Using the BAM reveals several insights:

- *First and permanent usage patterns*. Analyzing the radar graph at the beginning of the trial and later phases of the trial shows can show entry barriers of the mobile service. Results obtained in later stages show how people exploit mobile service strengths but also compensate potential weaknesses of the service.
- *Usage patterns for different technical adoption groups*. According to Rogers technical diffusion model user groups are divided in innovators, early adopters, early majority, late majority and laggards. Clustering logging data according to these groups may reveal interesting insights how the technical experience influences service usage. These observations are especially valuable to define a priority of feature improvements for the mobile service.
- *Behaviour changes in different product development stages*. As the development of the mobile service evolves comparing results with earlier trials can help to confirm if the applied feature modifications fulfil the intended improvements.

References

1. Reips, U.-D., Stieger, S.: Scientific LogAnalyzer: A Web-based tool for analyses of server log files in psychological research. Behavior Research Methods, Instruments, & Computers 36, 304–311 (2004)
2. Cassady, J.C., Johnson, R.E.: Cognitive Test Anxiety and Academic Performance. Contemporary Educational Psychology 27, 270–295 (2002)
3. UserTesting.com - Low Cost Usability Testing, http://www.usertesting.com
4. Morae usability testing tools from TechSmith, http://www.techsmith.com
5. Raento, M., Oulasvirta, A., Petit, R., Toivonen, H.: ContextPhone: A Prototyping Platform for Context-Aware Mobile Applications. IEEE Pervasive Computing 4(2), 51–59 (2005)
6. Jensen, K.L.: RECON: Capturing Mobile and Ubiquitous Interaction in Real Contexts. In: Proceedings of MobileHCI 2009, Bonn, Germany (2009)
7. Froehlich, J., Chen, M.Y., Consolvo, S., Harrison, B., Landay, J.A.: MyExperience: A System for In situ Tracing and Capturing of User Feedback on Mobile Phones. In: MobiSys 2007, June 11-14, pp. 57–70. ACM, San Juan (2007)
8. ter Hofte, H., Otte, R., Peddemors, A., Mulder, I.: What's Your Lab Doing in My Pocket? Supporting Mobile Field Studies with SocioXensor. In: CSCW 2006, Conference Supplement, Banff, Alberta, Canada, November 4-8 (2006)
9. Pretel, I., Lago, A.B.: Capturing Mobile Devices Interactions Minimizing the External Influence. In: Proc. of UBICOMM, pp. 200–205 (2011)
10. Ajzen, I.: The theory of planned behavior. Organizational Behavior and Human Decision Processes 50(2), 179–211 (1991)
11. Colbert, M.: A diary study on rendezvousing: Implications for position-aware computing and communications for the general public. In: Proc. of Group Conference (2001)
12. Klein, B., Perez, J., Guggenmos, C., Pihlajamaa, O., Heino, I., Ser, J.: Social acceptance and usage experiences from a mobile location-aware service environment. In: Ser, J., et al. (eds.) Mobile Lightweight Wireless Systems, vol. 81, pp. 186–197. Springer, Berlin (2012)