Learning Mobile App Design From User Review Analysis

doi:10.3991/ijim.v5i3.1673

E. Platzer[1] and O. Petrovic[2]
[1] evolaris next level, Graz, Austria
[2] Karl-Franzens-University, Graz, Austria

***Abstract*—This paper presents a new learning environment for developers of mobile apps that merges two quite different views of the same topic. Creative design and system engineering are core issues in the development process that are based on diverging principles. This new learning environment aims to address both points of view by not suppressing one of them but trying to benefit from both. User review content analysis is introduced as a tool to generate information that is useful for both aspects.**

***Index Terms*—application design; creativity tool; innovation support; user motive analysis.**

I. INTRODUCTION

Normally, Engineers are used to having clear specifications for developing software. Whole academic disciplines like system engineering follow a structured process with engineer-like thinking. If you look to mobile apps like that in Apple's AppStore, Google's Android Market or to the founder of Facebook, Mark Zuckerberg, you cannot find much of that engineering thinking and feeling. The challenge is: how can we build a learning environment for engineers to bring them to experiences in the field of real world, high emotional mobile apps which are loved by consumers?

The basic consideration of the learning environment is that due to app markets we have access to a broad range of apps in a very easy, fast and less expensive way. Thus, we are able to learn the development of mobile apps by browsing through and experiment with different apps. Additionally, we can also use the user-generated content in form of reviews and assessments together with download numbers to proof user acceptance and to deduce trends from that. The main aim of the learning environment currently under development is to enable engineers to explore existing mobile apps and related user-generated content in a semi-structured way and to experience critical success factors and current trends leading to high user acceptance. The goal is not to construct a specification robot but a learning environment for human beings.

In the first part of this paper a conceptual framework is presented that serves as a foundation for technical implementation of the system which is described subsequently. The system is then evaluated concerning its usability for developers of mobile applications. Results of this evaluation and a brief outlook on future research activities are provided at the end of the paper.

II. CONCEPTUAL FRAMEWORK

A. State of the Art

Three potential sources of information about user requirements and ideas for new mobile apps provide a basis for the conceptual framework for the suggested learning environment. Innovation support tools are often named synonymously with creativity techniques which provide more or less systematic instruments for idea generation. Basically there exist intuitive-creative methods like brainstorming, brainwriting or synectics and systematic-logic approaches like mind mapping or morphological analysis [10].

Technology acceptance research focuses on adoption and further usage of technology. Main concepts of acceptance research are Technology Acceptance Model [6] where "ease of use" and "usefulness" are key constructs and Task Technology Fit Model [7] that suggests strong influence of the fit between the challenging task and the technologies abilities to support the user with it on the behavioral intention to use a technology. The flow construct [4] is also a well-tested factor of technology acceptance. Next to that some compound models like the Unified Theory of Acceptance and Use of Technology [12] accumulate parts of existing models to form a new one.

User-generated content is a phenomenon that gained importance with the fast development of Web 2.0. Users publish their opinion concerning various aspects of life voluntarily in the internet. The form of publication ranges from product reviews to blogs. The rise of social networks like Facebook or mySpace added completely new possibilities of interaction between users that generate content on the web. The incredible amount of content that is available leads to initiatives like Folksonomies that aim to provide a user-generated taxonomy of previously unstructured information.

As shown in fig. 1 there are several approaches to combine innovation support tools, acceptance research and user-generated content but none of them addresses all three sources.

Dynamic models in technology acceptance research like Dynamic Approach for Re-evaluating Technologies and Compass-Model [1] include cyclic phases of technology design followed by acceptance research and redesign of technology. An approach to integrate end users in the innovation process are the Lead User concept [8] which is based on the assumption that certain people show pronounced needs that will be general phenomena in future.

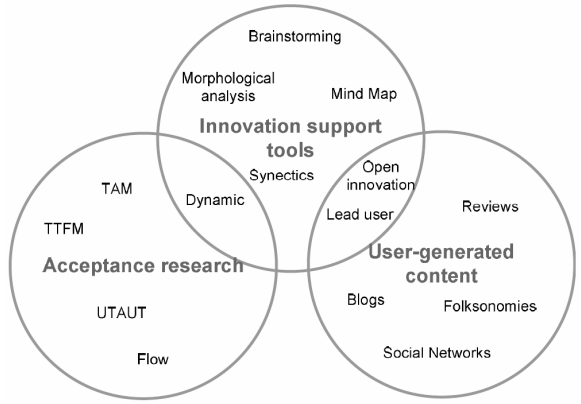

Figure 1. Interaction and integration of innovation support tools, acceptance research and user-generated content

A further development of this concept is the Customer as Innovator approach [11] that enables customers to create their own products by means of a toolkit based offer. Next to these rather market oriented approaches there also exist more cultural or society based concepts like Participatory Design or Design Anthropology [3].

B. Potentials for Improvement

The three sources of information and ideas are applied to different stages of the traditional design process. Innovation support tools assist the very first phase of idea generation whereas technology acceptance research takes place at the earliest when a prototype has been built which can be tested by users. Analysis of user-generated content is done at the end of the innovation process in order to find out what people think about the launched product or service and which changes they suggest. This procedure includes numerous potentials for further improvement.

Creativity techniques do not take into account acceptance factors but focus on the designers abilities to anticipate what users need. Acceptance research takes place when investment in infrastructure and product development has already been made. This fact often impedes fundamental changes of the product. Sometimes tested products or services are already in use and acceptance research is only made to fully understand why people use it or not without further consequences. In the opposite case when the product or service is not available for the respondents of the survey another problem will occur. The interviewees' answers are based on mere imagination instead of real experience.

The most commonly applied method of data gathering in technology acceptance research is survey with standardized questionnaires. Standardization of questionnaires limits the resulting reasons for acceptance to previously defined acceptance factors that must not cover or even include the real acceptance drivers. Moreover acceptance factors are commonly highly aggregated constructs in order to achieve a "good fit" of the tested model. This aggregation level causes fuzzy constructs that are not intersubjectively comprehensible as to say "ease of use" which is the most tested construct in technology acceptance research means different things to different people.

Moreover product life cycles in the mobile service market are quite short and surveys concerning technology acceptance take time when results should be at hand soon. Another potential lies in the analysis of user-generated content after market launch that will lead to incremental improvements of the existing product or service rather than to radical innovations.

To sum up the potentials for improvement found in the traditional process:

- There is a need to come up with the dynamics of development in the mobile service market.
- There is a need to enhance design relevance of provided information.
- There is a need to provide an environment that enables radical innovations.

C. Reshaped Process

The potentials presented above can be captured if acceptance research is done by means of user-generated content analysis and transferred into an innovation support tool that is integrated in the idea generation phase of the innovation process. This integration is possible if some preconditions are fulfilled. Firstly the analysis of user-generated content has to be done automatically in order to shorten the effort of time and money until results are at hand. The possibility to use information concerning acceptance factors immediately enables the designer to come up with the dynamics of the market. Moreover automation of the process allows continuous monitoring of acceptance factors and therefore avoids obsoleteness of information.

Design relevance is enhanced by providing information concerning basic motivations for usage of mobile applications and linking them to best practice examples. As in this framework acceptance research is done before the product or service is developed there is a need to redefine its goals. Traditionally acceptance research wants to find out why people adopt and use a certain service. In this case it should find out what makes people adopt and use successful mobile applications in general and then provide examples of mobile applications that where users emphasized these causes. It is very important to ensure that the user-generated content that is analyzed was produced by people who actually experienced the mobile applications that serve as best practice examples. This enables a shift from behavioral intention to actual behavior which makes results more valid concerning economic reality.

Radical innovations are possible as information is at hand before investment has been made which would prevent fundamental changes. The best practice examples can serve as a focused creativity tool. The system suggested in this paper is not a design tool that acts as a robot but a design support tool that acts as learning environment. Creative design is not replaced by automatically processed parameters of successful mobile applications but encouraged by providing some basic information concerning acceptance factors and examples of mobile applications that were successful in practice.

III. TECHNICAL IMPLEMENTATION

A. Data Source

Apples AppStore is used as the data source for prototypical implementation for several reasons. First of all it is the most used platform for distribution of mobile applications. In October 2010 more than 300.000 apps were available with more than 7 billion downloads performed. These usage numbers ensure reasonable amounts of avail-

able customer reviews for successful apps. Secondly AppStore allows reviews only if the reviewer has downloaded the app in question. Therefore it is ensured that each review is based on real experience of the app.

The US AppStore is chosen for analysis as it is huge and often used. Moreover most customer reviews are written in English which facilitates language processing. Analysis is limited to the 100 most downloaded apps of each category (free, paid and grossing). This is important because only successful services should be examined and the number of apps should be sufficiently high. The data that are used for analysis are app name, app ID in order to identify all other information, download rank in order to measure economic success of the app and all customer reviews related to the app in order to mine them for usage motives. The data are scraped automatically using a proxy and then filtering of relevant data and then saved in xml format to facilitate further processing.

B. Data Classification

There exist several options for classification of the customer reviews. Unsupervised clustering methods would provide a list of salient topics that are addressed in the reviews. This method is out of question as it would not lead to a learning system that automatically matches customer reviews with acceptance factors represented by usage motives and therefore acts as a forecasting tool of economic success and it would neither ensure design relevance of the information that is provided. Another way is supervised learning which could provide most accurate annotation of reviews with usage motives but is a too time consuming procedure in this case as the number of reviews is very high. Another approach is semi-supervised learning that enables automated annotation after training with manually labelled data. This method is most useful for the purpose of this research.

The first step is the manual annotation of a training data set of customer reviews with usage motives. Reiss model [9] is a very useful model of motivation that aims to cover all possible areas of motives for any human activity. The 16 basic desires listed in table 1 represent a canonical list that does not need adaptation or enlargement in case of technological development but remain validity.

The annotation of the training set is done by two independent annotators in order to ensure intersubjectivity of the data. For the machine learning process only data are used where manual annotation was the same for both annotators. These data are then annotated in GATE [5] and the precision of the machine based annotation is evaluated for the training set. This is done by splitting the training set and then comparing annotation results of the support vector machine [2] and the provided manually labelled data. Support vector machines learn a classification hyperplane in the feature space using the provided training data to find out maximal distance to all training examples. Generalization capabilities of support vector machines are usually good and outperform those of other distance- or similarity-based learning algorithms [2].

The machine learning model is applied to all data as soon as evaluation results like F-measures are satisfying. When all reviews are annotated the next step is to calculate frequencies of usage motives. These frequencies represent proportional importance of usage motives as addressed in the reviews.

TABLE I.
16 BASIC DESIRES OF REISS MODEL OF MOTIVATION

Motive name	Motive	Intrinsic feeling
Power	Desire to influence (including leadership; related to mastery)	Efficacy
Curiosity	Desire for knowledge	Wonder
Independence	Desire to be autonomous	Freedom
Status	Desire for social standing (including desire for attention)	Self-importance
Social Contact	Desire for peer companionship (desire to play)	Fun
Vengeance	Desire to get even (Including desire to compete, to win)	Vindication
Honor	Desire to obey a traditional moral code	Loyalty
Idealism	Desire to improve society (including altruism, justice)	Compassion
Physical exercise	Desire to exercise muscles	Vitality
Romance	Desire for sex (including courting)	Lust
Family	Desire to raise own children	Love
Order	Desire to organize (including desire for ritual)	Stability
Eating	Desire to eat	Satiation (avoidance of hunger)
Acceptance	Desire for approval	Self-confidence
Tranquility	Desire to avoid anxiety, fear	Safe, relaxed
Saving	Desire to collect, value of frugality	Ownership

C. Data Interpretation

Developers of mobile apps are provided with several forms of data interpretation. Firstly they get a ranking of usage motives that are currently important. The motives are arranged according to their frequency within the analysed reviews. Also their proportional importance regarding the other motives is displayed. As the system is planned to serve as a continuous learning environment it is also possible to compute changes within the motive structure over time.

Best practice apps are available for each motive. Best practice means that these apps address the motive best. This is indicated by disproportionately high frequency of the motive in question within the reviews related to the app.

Another functionality of the system is that certain apps can be monitored and analysed in comparison to the most successful apps. The motives addresses in reviews concerning the selected app and those in all the successful apps are juxtaposed and differences are calculated.

Next to annotation of usage motives the system will learn a machine learning model that matches customer reviews and download ranks that were provided in the xml files extracted from AppStore. This second learning model allows forecasting economic success of new apps by means of download rank prognoses. The download rank prognosis is computed by means of probabilistic heuristics.

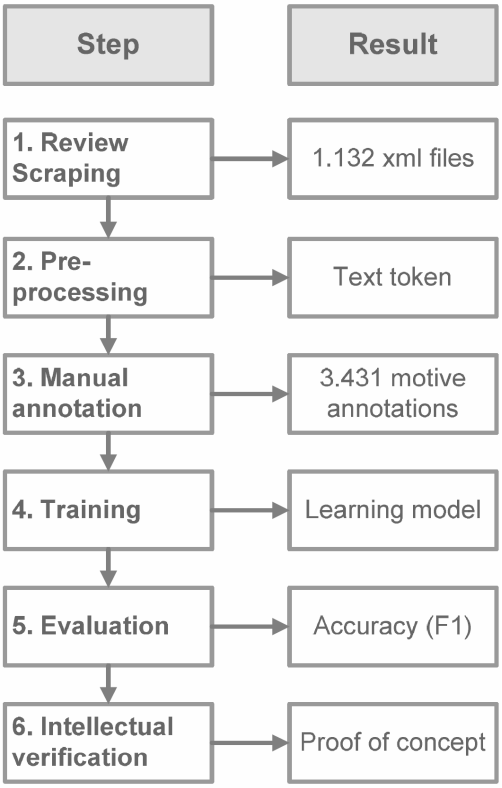

Figure 2. Procedure of the feasibility study including results of each step.

D. *Feasibility study*

A prototypical exemplary application of the system was developed in order to test the general feasibility of automated motive-based content analysis of user reviews. This was done in a six-step process that is depicted in fig. 2.

The first step was scraping the reviews concerning the top apps. This was done in form of a snap-shot at a given moment. The 277.345 reviews of the top 100 free apps, the top 100 paid apps and the top 100 grossing apps were then transformed into xml files including the needed meta data. In order to ensure balance for the further processing each file contained 200 reviews at most. This process resulted in 1.588 xml files that were then reduced by the doublets that occurred due to the fact that one app can be a top app in more than one category.

After that the remaining 1.132 xml files were preprocessed for the machine learning tasks. Finite state transducers were used for the tokenization of the text. 60 files containing 9.510 reviews were randomly chosen from the 1.132 files. These files served then as a training data set. Two annotators tried to manually annotate each of the 9.510 reviews with one motive that was salient in the text after a discussion concerning the meaning of the 16 motives in the context of mobile apps. There was also an option to annotate none of the motives because there was either no motive identifiable or several motives were mentioned and it was impossible to tell which one was dominant. The manual annotations were then compared and 3.431 corresponding annotations were found. This represents about one third of the total sample size. As the training data set was randomly chosen this leads to the assumption that it is possible to identify intersubjectively comparable motives in about one third of all reviews.

The manually annotated reviews were then used for the training of the learning model. Several engines were tested in order to find the most powerful one. Next to a support vector machine also a Naïve Bayes, C 4.5 decision tree, k-nearest neighbor were computed for reasons of comparison. As expected because of data base characteristics the support vector machine provided superior results to the other standard algorithms. Unigrams were used to obtain kernels for the machine learning. For the review classification task the multiclass problem of 15 motive classes ("idealism" was not present in the sample) was transferred into numerous binary problems that could be computed by the system. The threshold probability for classification was set 0.4. This level was supposed to be sufficiently high to keep classification results meaningful and also sufficiently low to obtain a satisfying number of classified instances. Motive kind was the classification target for each of the review instances.

A hold-out test where the training data set is split into two parts was carried out for evaluation of the machine learning model. A new model was learned from only two thirds of the training data and then applied to the remaining third. Then the results of the automated annotation were compared to those of the previous manual annotation. The overall accuracy level (F1 measure) of the learning model was 0,67. This is sufficient for the conclusion that it is possible to obtain meaningful classification results concerning motives when analyzing the content of customer reviews in AppStore.

To double-check the meaningfulness of the resulting annotations the leaning model was applied to some of the remaining xml files that were not annotated by hand. The annotations that were suggested by the system were then verified intellectually. It showed that in general the tested annotations were meaningful and comprehensible.

The concept of automated motive-based user review content analysis is therefore considered to be generally feasible.

IV. EVALUATION

A. *Methodology*

The evaluation of the presented system is executed in cycles. This first evaluation of design relevance shall provide information for further development of the system itself and also concerning its actual technical implementation. In a later evaluation cycle design relevance and usability of the system will be tested in a field study with more experts. An expert-based qualitative approach was chosen as it will lead to more in depth information. As the system is not fully implemented yet we used "scribbles" for the evaluation. These "scribbles" are draft-like virtual screens of the results the system will provide. The system was presented to three app developers from different areas of development – creative system design, technical implementation and user interface design - in form of the drafted screens which are depicted in fig. 3, 4, 5 and 6.

They were then interviewed separately concerning their perceptions of design relevance and usefulness respectively their suggestions for further improvement. Fig. 3 shows a fictitious pie chart of usage motives that were addressed in the customer reviews. In fig. 4 variation of these relative usage motives is depicted over time. Fig. 5presents the planned functionality of the system to com-

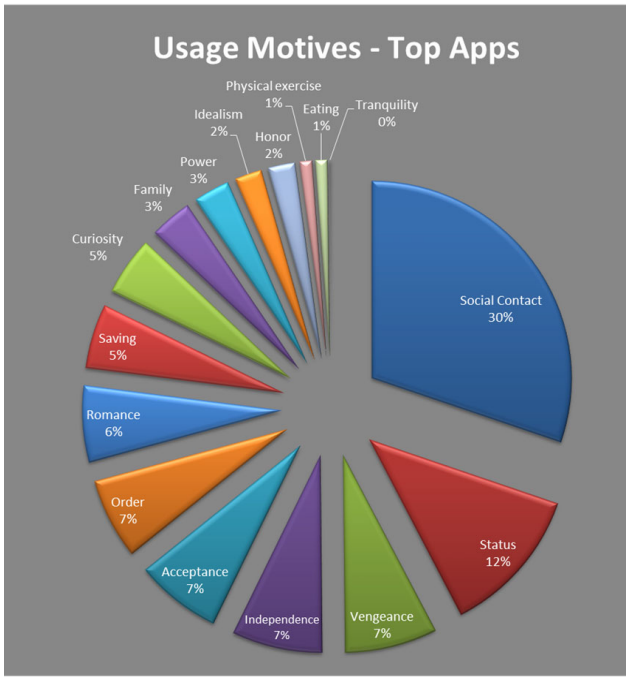

Figure 3. Screen 1: Relative usage motives in top apps (top 100 paid, top 100 free and top 100 grossing)

pare a certain app with the top apps. The app that was used as an example shows shortcomings concerning major motives whereas minor motives are over-represented. Fig. 6 finally gives an example of the "best practice"-section. Five apps are presented that addressed the motive in question best. A link to the AppStore enables immediate download of the app that will initiate a creative learning process.

B. *Results*

The review analysis approach is regarded as more useful and more design relevant than questioning as the reviews are "closer to reality" and reflect what "the user really experiences" and it ensures that the respondent is "really interested in the product". The results that acceptance research could provide – Technology Acceptance Model [6] and Task Technology Fit Model [7] were presented as the most often used models - would be also helpful if they were at hand when idea generation takes place. Moreover the information should be provided on a more detailed level (e.g. how to achieve "ease of use").

Screen 1 was regarded as useful for idea generation and optimization of existing apps. The offered information give "a direction for one's design objectives". It was not considered to be relevant for the design of user interfaces but for the development of own ideas. One expert emphasized the fact that "one can see at a glance what the world doesn't need".

The advantages of screen 2 are to be seen in its trend depiction as a designer could derive future importance of motives from their past development. It is expected to be very useful for idea generation where there "has always been a lack of such data".

The experts did not consider screen 3 to be as useful as the previous two screens. The comparison to all top apps is not design relevant if the own app aims to be a niche product. It would me more useful if successful apps with similar usage motive structure were provided. One expedient use case of the comparison is evaluation of target achievement regarding the motives that were intended to address and those that actually were addressed.

Screen 4 was regarded as most useful for graphic design and feature design. One expert named this screen as the most useful functionality of the presented system as it really allows learning from the best practice examples. The experts reported that they usually try to find apps similar to that they want to design and would be more than happy to get a thorough report on that without further research. Criticism that was passed on this feature was that all apps are more or less built the same way.

C. *Discussion*

As the results of this first evaluation cycle show each presented feature was regarded to be useful and design relevant for at least one aspect of mobile application development it is reasonable to adhere to the presented data interpretation and representation forms. All four functionalities will be implemented in the technical solution.

The editing will be very content-oriented according to the expert requirements. There is no need to focus on the graphical interface but instead provide the information in a purist design that does not influence the creative app design process too much.

In the course of the interviews two experts mentioned their strong need for a kind of "price finding support tool" that could possibly be implemented in the final system as an additional functionality. Such a tool could be "worth its weight in gold" as developers of apps often experience that a good app fails because of wrong prizing. The technical implementation could be computed as an additional machine learning model similar to the rank prognoses model where rank is forecasted based on reviews and realized ranks of the top apps. When the price of the top apps is added as additional information it would be possible to train a machine learning model that connects customer reviews and prices of top apps and then suggests a price for the new app based on its reviews. A difficulty in this plan is that it will be problematic to obtain customer reviews for the actual app before a price is set. It could harm the success of the app if the price is set to zero until there exist enough customer reviews to compute an optimized price and then raise the price without added value for users.

V. OUTLOOK

The next steps in the research process include technological implementation of the system on a ready to use level. As soon as this is done it will be possible to evaluate the usefulness of the system in practical use.

To further develop the system it will be necessary to evaluate its accuracy over time. The functional test of the system in the course of the feasibility study was executed in one run. In order to find out whether the system is able to keep up with the dynamic changes of the data base it will be useful to evaluate the system on the long-term. This is to say that the learning model is applied to updated data from AppStore at regular intervals and the results of automated annotation are compared to additional examples of manually annotated reviews. This comparison could uncover decline of accuracy over time. In this case it might be useful to implement active learning elements. Decreasing accuracy can occur when the text characteris-

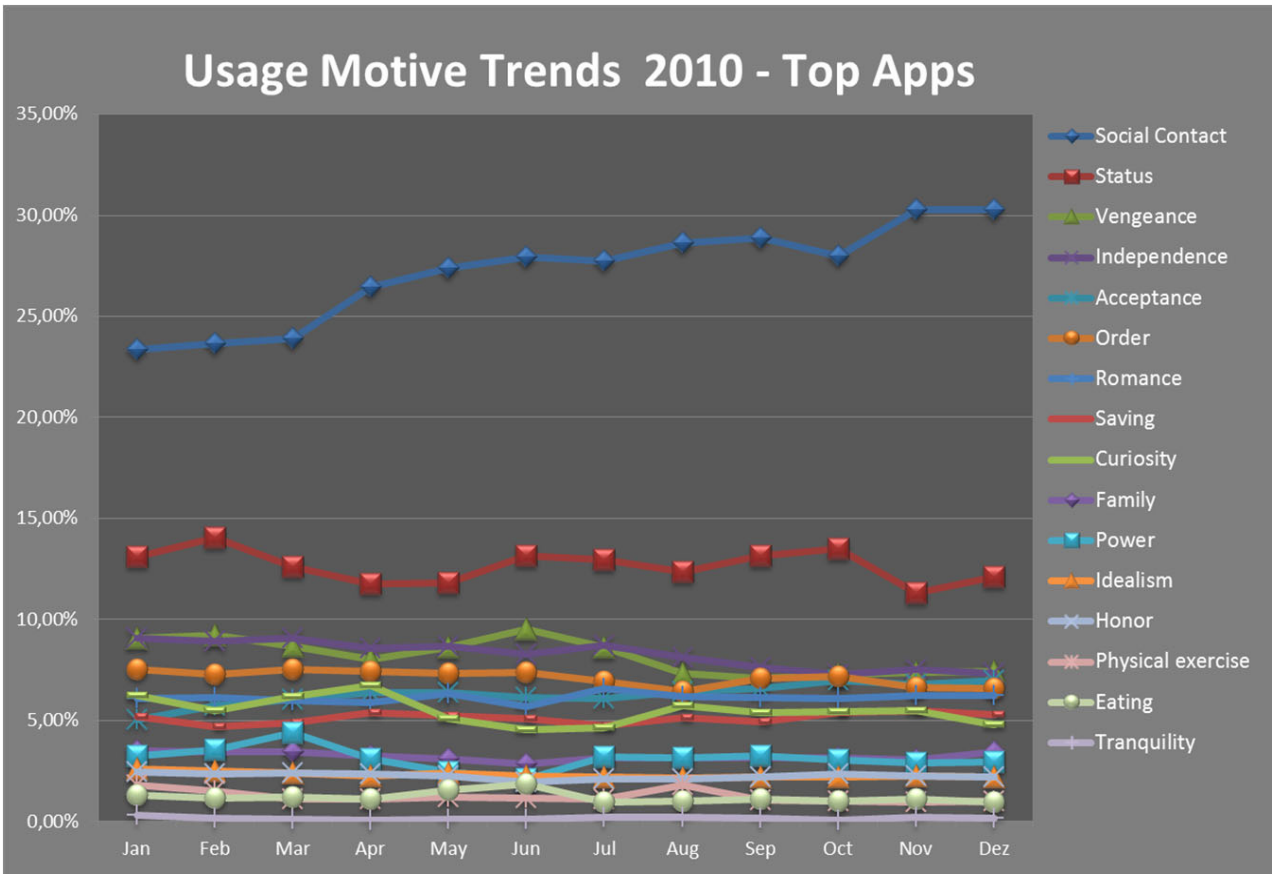

Figure 4. Screen2: Usage motive trends over a period of one year

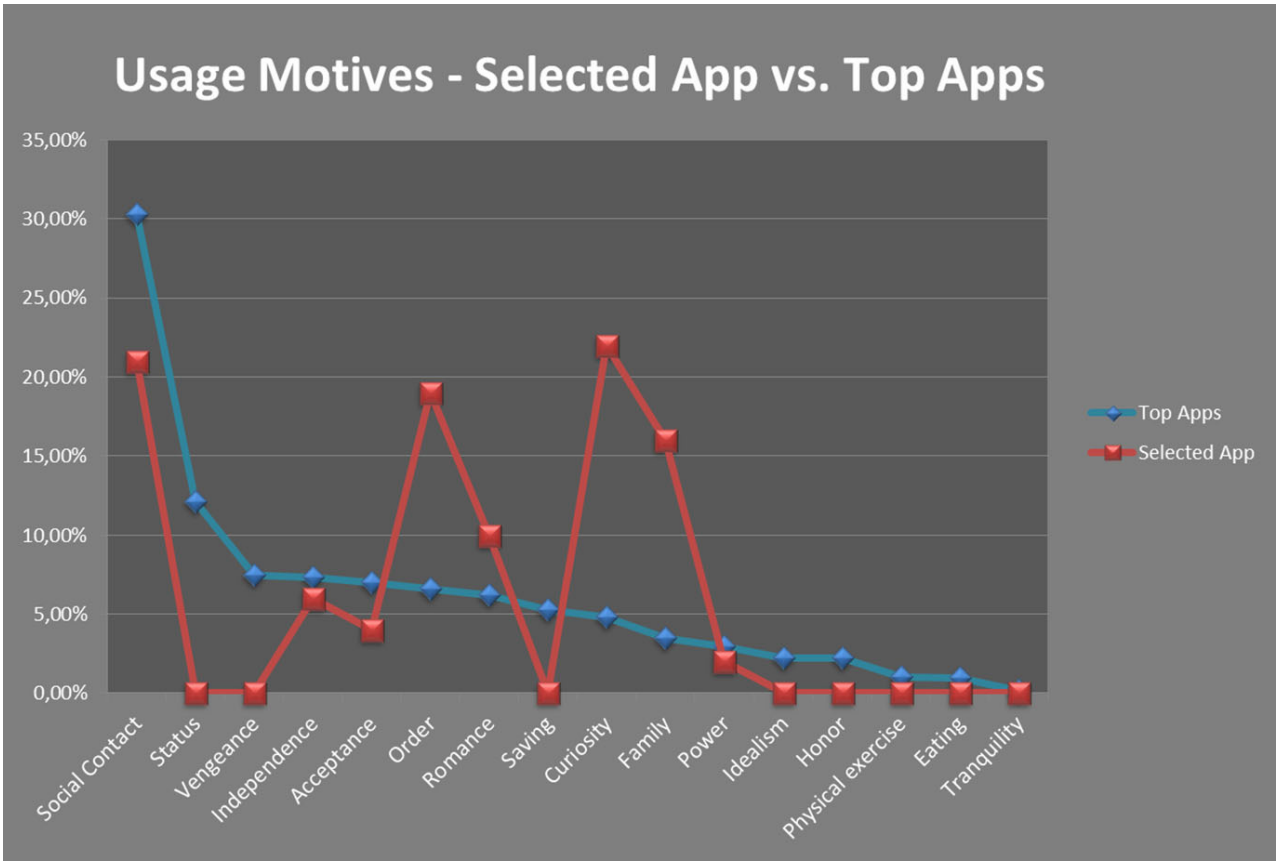

Figure 5. Screen3: Comparison of usage motives between top apps and a certain selected app

Figure 6. Screen 4: Example for presentation of best practice apps

tics that indicate classes do not change gradually but suddenly. Evaluation of the system's usefulness has been done by means of descriptive methods so far. It is necessary to continue the evaluation and include experimental and observational methods. This will be possible as soon as the system is ready for practical usage in app development processes. At the moment the system is implemented in form of a semi-automated prototype and trained for classification of reviews concerning the motives included in the motivational model by Reiss. Training of the system concerning other models is possible at any time by means of manual annotation.

Moreover it will be interesting to observe economic success of mobile applications that were developed supported by the learning environment presented in this paper in a long-term study. This further evaluation of the system can provide deeper insights concerning its usefulness in practice. An accompanying usability study with developers of mobile applications could support further development of the learning environment.

Another focus of future research will be applicability of the system to other data sources than AppStore or even other fields of products or services. The functionalities of the presented system are not bound to the mobile service market. Generalizability of the system will be tested in selected areas.

REFERENCES

[1] M. Amberg, M. Hirschmeier, J. Wehrmann, „The compass acceptance model for the analysis and evaluation of mobile services", International Journal of Mobile Communications 2(3), pp. 248-259, 2004.

[2] Y. Li, K. Bontcheva, H. Cunningham, "Adapting SVM for data sparseness and imbalance: a case study in information extraction," *Natural Language Engineering* 15(2), pp. 241-271, 2008. doi:10.1017/S1351324908004968

[3] J.Burr and B. Matthews, "Participatory innovation", *International Journal of Innovation Management* 12 (3), pp. 255-273, 2008. doi:10.1142/S1363919608001996

[4] M. Csikszentmihalyi, *Das Flow-Erlebnis: Jenseits von Angst und Langeweile im Tun aufgehen*. (The Flow-Experience: Being carried away with action beyond fear and boredom.), Stuttgart: Klett-Cotta, 1987.

[5] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, "GATE: A framework and graphical development environment for robust NLP tools and applications", in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL'02), Philadelphia, July 2002.

[6] F. Davis, *A Technology Acceptance Model for Empirically Testing New End-User Information Systems*, Massachusetts Institute of Technology, Sloan School of Management Thesis, 1985.

[7] D. L. Goodhue, R. L. Thompson, "Task-technology fit and individual performance", *MIS Quarterly* 19, pp. 213-236, 1995. doi:10.2307/249689

[8] E. v. Hippel, "Lead Users: A Source of Novel Product Concepts", *Management Science* 32 (7), pp. 791-805, 1986 doi:10.1287/mnsc.32.7.791

[9] S. Reiss, "Multifaceted Nature of Intrinsic Motivation: The Theory of 16 Basic Desires", *Review of General Psychology* 8(3), pp. 179-193, 2004. doi:10.1037/1089-2680.8.3.179

[10] G. Steiner, "Kreativitätsmanagement: Durch Kreativität zur Innovation (Creativity Management: To Innovation via Creativity)", in Strebel, H., Ed., *Innovations- und Technologiemanagement* (Innovation and Technology Management), Vienna: WUV Universitätsverlag, pp. 265-325, 2003.

[11] S. Thomke and E. v. Hippel, "Customers as innovators, a new way to create value", *Harvard Business Review* 80 (4), pp. 74-81, 2002.

[12] V. Venkatesh, "User acceptance of information technology: toward a unified view", *MIS Quarterly* 27, pp. 425-478, 2003

AUTHORS

E. Platzer was with the Department of Information Science and Information Systems at Karl-Franzens-University, Graz, Austria. She is now with evolaris next level (e-mail: elisabeth.platzer@evolaris.net).

O. Petrovic is with the Department of Information Science and Information Systems at Karl-Franzens-University, Graz, Austria (e-mail: otto.petrovic@uni-graz.at).

Manuscript received May 13th, 2011. Published as submitted by the authors June 9th, 2011.