UMAP 2013 Late-Breaking Results and Project Papers

Eelco Herder[1] and Olga C. Santos[2]

[1] L3S Research Center, Leibniz University Hannover, Germany
herder@L3S.de
[2] aDeNu Research Group, Artificial Intelligence Department, Computer Science School, UNED, Spain
ocsantos@dia.uned.es

Abstract. Late-breaking results contain unpublished accounts of innovative research ideas, preliminary results, system prototypes or industry showcases. Project papers introduce newly started research projects. These proceedings contain 9 late-breaking results and 4 project descriptions that were accepted for publication at UMAP 2013. The papers span a wide range of topics related to user modeling and personalization, and have been peer-reviewed by experts in the field.

Preface

UMAP is the premier international conference for researchers and practitioners working on systems that adapt to their individual users, or to groups of users, and collect and represent information about users for this purpose. The conference spans a wide scope of topics related to user modeling and personalization, including construction of user models, adaptive responses, tailoring of search results, recommending products, Web usage mining, collaborative and content-based filtering.

In addition to the regular research paper track, in which substantive new research is presented, the UMAP conference provides the opportunity to present innovative research ideas, preliminary results, system prototypes or industry showcases in the form of a poster or a demonstration. The first call for posters and demos received 27 submissions, of which 10 were accepted for publication in the main proceedings of UMAP 2013, published by Springer.

This year, UMAP 2013 offered an additional opportunity to submit late-breaking results, which were submitted only shortly before the conference. Similar to posters and demos, late-breaking results present innovative research ideas and preliminary results. Further, contributions were requested in which newly started research projects were introduced. In total, we received 18 submissions, of which 9 late-breaking results and 4 project descriptions have been accepted. These accepted papers are included in these online extended proceedings of UMAP 2013.

We thank all authors who submitted contributions to the Poster, Demo and Late-Breaking Result tracks of UMAP 2013. Accepted submissions are presented

at the UMAP 2013 Poster and Demo Reception, during which the conference attendees can vote for the best poster, the best demo and the most inspiring contribution.

We are also grateful to the following members of the Program Committee for their support and reviews of the submitted posters, demos, late-breaking results and project papers.

Program Committee

- Kenro Aihara, National Institute of Informatics, Japan
- David Albrecht, Monash University, Australia
- Liliana Ardissono, University of Torino, Italy
- Mathias Bauer, mineway GmbH, Germany
- Shlomo Berkovsky, NICTA, Australia
- Richard Burns, West Chester University, U.S.A.
- Federica Cena, University of Torino, Italy
- David Chin, University of Hawaii, U.S.A.
- Mihaela Cocea, University of Portsmouth, U.K.
- Alexandra Cristea, University of Warwick, U.K.
- Paul De Bra, Eindhoven University of Technology, Netherlands
- Marco De Gemmis, University of Bari, Italy
- Vania Dimitrova, University of Leeds, U.K.
- Peter Dolog, Aalborg University, Denmark
- Benedict Du Boulay, University of Sussex, U.K.
- Stephanie Elzer-Schwartz, Millersville University, U.S.A.
- Cristina Gena, University of Torino, Italy
- Bradley Goodman, The MITRE Corporation, U.S.A.
- Eduardo Guzmn, Universidad de Mlaga, Spain
- Neil Heffernan, Worcester Polytechnic Institute, U.S.A.
- Geert-Jan Houben, TU Delft, Netherlands
- Dietmar Jannach, TU Dortmund, Germany
- W. Lewis Johnson, Alelo Inc., U.S.A.
- Judy Kay, University of Sydney, Australia
- Tsvi Kuflik, The University of Haifa, Israel
- Mark Maybury, MITRE, U.S.A.
- Gordon McCalla, University of Saskatchewan, Canada
- Carlo Tasso, University of Udine, Italy
- Nava Tintarev, University of Aberdeen, U.K.
- Michael Yudelson, Carnegie Mellon University, U.S.A.

Encountering the Unexpected: Influencing User Experience through Surprise

Alice Gross[1], Manfred Thüring[2]

[1]Technische Universität Berlin, Graduiertenkolleg prometei, Berlin, Germany
agross@zmms.tu-berlin.de
[2]Technische Universität Berlin, Institut für Psychologie und Arbeitswissenschaft, Fachgebiet Kognitionspsychologie und Kognitive Ergonomie, Berlin, Germany
manfred.thuering@tu-berlin.de

Abstract. When purchasing an interactive product, users nowadays seek more than a flawless functionality and a comfortable ease of use. Products need to be enjoyable and exciting to have a unique selling point. User Experience (UX) is constituted by the instrumental qualities as well as the hedonic qualities of a product and impacts on the user's overall appraisal. One way to improve product appraisal is the use of surprise as a design element. Surprising product design has been shown to be beneficial for the user and the rating of a product. By using the classical computer game Tetris, the impact of surprise on UX ratings of a digital, interactive computer game was investigated. The results of our study stress two points. First, unexpected events with *undesirable* consequences lead to negative surprises which in turn impede users' information processing and have a bad impact on user experience. Second, whether unexpected events with *desirable* consequences lead to positive surprises, mainly depends on the interaction context and on the kind of system under consideration.

Keywords: User Experience, Surprise, Usability, User Centred Design.

1 Introduction

For many years, usability issues, such as effectiveness and efficiency [1], have dominated research and development in the domain of interactive systems. But due to technical advancements and the growing importance of user centered design, good usability is no longer something to be excited about. Instead, it has turned into a quality feature that is almost taken for granted. Today's customers are on the lookout for products that are not only easy to use, but that are exciting and pleasurable [2]. As Norman states "...the emotional side of design may be more critical to a product's success than its practical elements."[3, p.5]

Exciting products motivate customers to prefer one product over another [4]. Classical product design strives to create excitement and interest by adding surprise features to a product [5]. Because such products do not match the expectations of their users, they are more interesting, easier to remember, and elicited increased word-of-mouth than similar, conventional products [5]. These insights raise the question

whether similar effects can be attained by furnishing interactive products with surprising aspects because surprise may arouse interest and intensify user experience (UX).

2 Expectation, Surprise and User Experience

UX can be defined as "a person's perceptions and responses that result from the use and/or anticipated use of a product, system or service" [6]. Expanding this definition, the CUE model (Components of User Experience) by Mahlke and Thüring [7] describes the emergence of UX in more detail: When users interact with a product, they perceive its various instrumental and hedonic features, get impressions of its strengths and weaknesses and gradually form an opinion about it. These cognitive activities are accompanied by emotions which may be positive or negative depending on the quality of the interaction. Together, cognition and emotion constitute the users' overall experience that evolves from their actions and the responses of the system.

Some authors highlight the relevance of expectations which arise in the course of interaction. For instance, Pohlmeyer, Hecht and Blessing [8] emphasize the importance of anticipated experience for UX, and Karapanos states that even a person who has never interacted with a particular product may have expectations about its behavior when in use [9].

According to Reisenzein, there is a direct connection between expectations and emotions. In his belief-desire theory of emotion (BDTE), he claims that "emotions are the product of cognitions (beliefs) and motives (desires)" [10]. The result of an unfulfilled belief (or expectation) is surprise. If expectations are disconfirmed and this disconfirmation co-occurs with desire fulfillment, the result is a pleasant surprise. An unpleasant surprise results from a disconfirmation of expectations which co-occurs with desire frustration. In both cases, a prolongation in reaction times (RTs) can be observed, which may be used in an experiment to check whether an attempted surprise manipulation was successful or not[10].

Product designers have made use of the benefits of pleasant surprise for instance when designing tangible products [11]. They were able to demonstrate the beneficial effect of surprise by creating products that had similar visual appearances but differed in their tactual characteristics [5]. By creating these visual-tactual incongruities, they were able to provoke surprise reactions.

While pleasant surprise has been studied extensively in classical product design, not many researchers have actively explored it as a design factor for digital, interactive products. Although some studies refer to surprise related concepts, like WOW, delight or appraisal [4, 12, 13], most research was constrained to non-interactive products. In contrast, we investigate surprise in the context of interactive products. To clarify how surprising behavior of digital products might influence UX, we address two issues: 1) Does UX differ between two products which are basically identical but elicit either pleasant or unpleasant surprises? 2) Is a surprise event still surprising when it occurs more than once? To answer these questions, we carried out an experiment in which three groups of participants played three differently surprising Tetris games.

3 Method

3.1 Participants and Experimental Design

A total of 60 persons took part in the study, (14 female and 46 male). Their average age was 24.6 years (SD = 4.2). All of them were familiar with the game.

Two independent variables were manipulated in the experiment. The first one was a between-subjects variable called *'group'*. It had three levels. The *'bonus group'* unexpectedly received 50 additional points during the game, while the *'minus group'* suffered an unexpected loss of 50 points. The third group served as *'control'* and played the game without any surprising incidence. The participants were randomly distributed over the groups with 20 persons per group ensuring a similar male/female ratio per group. The within-subjects factor *'event'* served as second independent variable. It consisted of three treatments, i.e., the first, second and third time an unexpected event occurred (e1 to e3). Four different measures were employed as dependent variables. Reaction times (RTs) were measured for processing a Tetris stone that was accompanied by a surprising event. UX was assessed using three questionnaires: (a) the self-assessment manikin (SAM), a 2-item 9-point non-verbal instrument for the evaluation of emotions measuring the dimensions *valence* and *arousal* (SAM) [14], (b) the AttrakDiff questionnaire, a 28-item semantic differential with the subscales *pragmatic quality*, *hedonic quality identification*, *hedonic quality stimulation*, and *attractiveness* [15], and (c) a self-developed single-item questionnaire for judging the overall UX on a 6-point non-verbal scale showing a thumb down at one end and a thumb up on the other (see [16]).

3.2 Hypotheses

Three effects of the independent variables *group* and *event* are expected:
H1: For the factor *group*, a main effect on reaction times is predicted. Mean RTs for the *bonus group* and for the *minus group* are longer than for the *control group* because surprises increase processing time.
H2: For the *bonus group* and the *minus group*, mean RTs will decrease from event 1 over event 2 to event 3 because the extent of surprise diminishes when an unexpected event is encountered more than once. Therefore, an interaction effect of *group* and *event* is predicted for the reaction times.
H3: Since a positive surprise will lead to an improvement of UX, ratings of the *bonus group* will be better than those of the *control group*. Also, ratings for the *minus group* will be worse than for the *control group* because negative surprises impair UX.

3.3 Procedure

Participants played a game of Tetris and were instructed to reach a certain amount of points within 5 minutes, gaining 10 points for every stone they placed on the square board. All participants played the same sequence of 66 stones. They were not informed about the possible occurrence of any surprises beforehand. To motivate them

to play as ambitiously as possible, they were rewarded 7 Euros for participating in the experiment and received an additional 3 Euros for reaching the required goal.

To induce surprises, a message flashed on the computer screen at three different times during the game (i.e., simultaneously with the appearance of stone 38, 47 and 51). While the *minus group* saw "!!!Abzug: -50 !!!" (Abzug=Reduction), the *bonus group* saw "!!!Bonus: +50 !!!", see figure 1. The c*ontrol group* played the game without encountering any surprising message. RTs were measured via key log from the first simultaneous appearance of a stone and a message until first key stroke.

Fig. 1. Surprise events for *minus* group (left) and b*onus group* (right)

4 Results

Of all 60 participants, six were not able to finish the game, resulting in a game over. To avoid this negative experience having any impact on UX ratings, these participants were excluded from further analysis. A 3x3 analysis of variance (ANOVA) of the RTs was carried out with event as a within-subject factor and group as a between-subject factor. There was a main effect for the factor group ($F(2,47)=6.46$, $p=.003$, partial $\eta 2 = .216$). Figure 2 illustrates that the control group was faster than the bonus group which in turn required less time than the minus group to react under the surprising conditions. Contrasts revealed that participants in the minus group were significantly slower than participants in the bonus group ($p=.047$) as well as in the control group ($p=.001$). The difference between the bonus group and the control group, however, was not significant.

There was also a main effect for the factor event ($F(2,94)=3,338$, $p=.040$ partial $\eta 2 =.066$), indicating that RTs decreased from e1 over e2 to e3. Contrasts showed that e1 differed significantly from e3 ($p=.016$). There was no significant interaction between group and event ($F(4,94)=1,413$, $p=.236$, partial $\eta 2=0.057$).

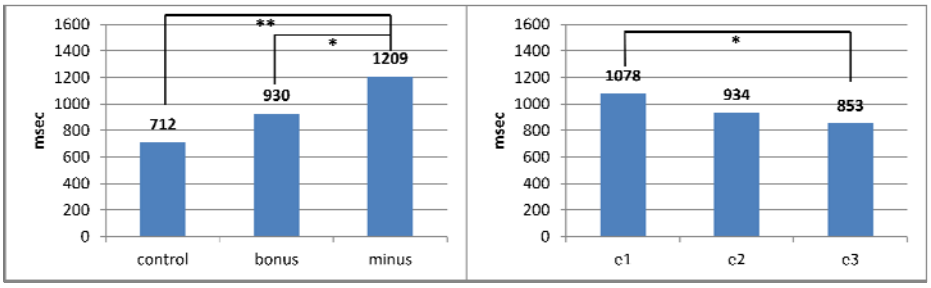


Fig. 2. Mean RTs per group (left), and mean RTs per event in milliseconds

Mean ratings for the dimensions of the UX-questionnaires are shown in table 1. To investigate surprise effects on UX ratings, a one-factorial MANOVA was carried out with '*group*' as between-subjects factor (all values z-transformed). The MANOVA revealed a significant main effect ($F(30, 66) = 1{,}851$, $p=.019$; Wilk's $\Lambda = 0.295$, partial $\eta 2 = .46$). Significant effects were found for the SAM subscale Valence ($F(2,47)=4{,}662$, $p=.014$, partial $\eta 2 = .166$) and the AttrakDiff subscale Hedonic Quality Identification (HQI), ($F(2,47)=4{,}647$, $p=.014$, partial $\eta 2 = .0165$). Contrasts showed that participants in the *minus group* gave significantly worse ratings than participants in the *bonus* and c*ontrol group* on both of these scales. Furthermore, contrasts revealed that participants in the *minus group* rated the game significantly worse than participants in the *bonus group*.

Table 1. Untransformed questionnaire ratings (Overall: 1=thumbs down, 7= thumbs up; SAM Valence: 1=happy, 9=sad SAM Arousal: 1=aroused, 9=calm) per group (AD: AttrakDiff).

Group	Overall	SAM Valence	SAM Arousal	AD-PQ	AD-HQS	AD-HQI	AD-Attraction
Control	5,86	2,31	4,75	5,14	4,09	4,24	5,29
Bonus	5,75	3,47	4,65	5,19	4,21	4,48	5,47
Minus	6,24	2,06	4,76	5,03	3,51	3,71	4,96

5 Discussion

Research on surprise as a design strategy has shown beneficial effects on the appraisal of a variety of non-digital artifacts [5]. The goal of our study was to test the influence of surprise on UX with digital, interactive products.

To induce different surprises in the course of a Tetris game, a *bonus group* received unexpected additional points, whereas a *minus group* suffered an unexpected loss of points. We predicted an increase of RTs in these two groups for trials, in which a surprise occurred, compared to a control group (H1). In support of this hypothesis, a significant effect of the factor '*group*' was found. However, single comparisons revealed that there was no significant difference between the *bonus group* and the *control group*. Only the differences between the *minus group* and the other two groups proved to be statistically relevant. Since the prolongation of reaction times

is a good indicator for surprise, we cannot be sure that the unexpected bonus worked as intended. An explanation for this result might be that a bonus in a game is not that unusual and hence not very surprising. On the other hand, a sudden and arbitrary reduction of points is rather uncommon and might therefore come as a real surprise.

Our second prediction concerned the change of reaction times over time (H2). It was assumed that an unexpected event loses its surprising character when it is encountered for a second or even a third time. In accordance with this hypothesis, reaction times decreased from the first to the third occurrence of the unexpected event (see right side of figure 2).

To measure the impact of surprise on UX, a number of rating scales was used. Our results do not fully support H3. However, it revealed that emotional valence as well as HQI were affected by the factor *group*. This effect resulted from the impact of negative surprises in the *minus group*. Mean ratings differ between this group and the other two groups in the expected direction. But similar to the results of the reaction times, no difference between the *bonus group* and the *control could* be substantiated.

In summary, it seems that our manipulation of surprise was only partially successful. Apparently the unexpected bonus was not as surprising as we had intended. This interpretation is supported by both, RTs as well as UX ratings. The unexpected loss of points though had the predicted effect. Trials with unpleasant surprises took longer to process and the ratings of the respective group indicate a less positive UX.

With respect to UX, our results stress two points. First, unexpected events in the course of human computer interaction which entail *undesirable* consequences should be prevented under all circumstances. They lead to negative surprises which in turn impede users' information processing and have a bad impact on UX. Second, whether unexpected events with *desirable* consequences lead to positive surprises, mainly depends on the interaction context and on the kind of system under consideration. As our experiment shows, an unexpected bonus in a game may not be as surprising as one might suppose. For other systems and in different contexts, such as software in a working environment, an unexpected and beneficial system response may prove as more surprising. Therefore, more research is required to investigate the causes and effects of positive surprise.

From a marketing perspective, our study raises the question whether all positive features of a system should be immediately apparent or whether some of them should be covered. Is it more beneficial to tell customers all positive aspects to prompt them to purchase the system? Or is it better to let them discover some surprising extras later which might pay off in the long run by increasing brand loyalty? Obviously, our results are not far-ranging enough to provide a sound answer, but future investigations may shed more light on this issue.

References

1. ISO. ISO 9241-11: Ergonomic requirements for office work with visual display terminals, The international organization for standardization (1996)
2. Jordan, P.W.: Putting the pleasure into products. IEE Review 249-252 (1997)

3. Norman, D.: Emotional Design: Why we love (or hate) everyday things. Basic Books, New York (2004)
4. Desmet P., Porcelijn R., van Dijk M.: Emotional design; Application of a Research-Based Design Approach. Know Technol Pol 20, 141–155 (2007)
5. Ludden, G. D. S., Schifferstein, H. N. J., Hekkert, P.: Visual-tactual incongruities in products as sources of surprise. Empirical Studies of the Arts, 27(1), 61-87 (2009)
6. ISO 9241-210: Ergonomics of human-system interaction -- Part 210: Human-centered design for interactive systems. The international organization for standardization (2010)
7. Mahlke, S., Thüring, M.: Studying antecedents of emotional experiences in interactive contexts. In: Proc. Conference on Human-Computer Interaction, pp. 915-918. ACM Press, New York (2007)
8. Pohlmeyer, A.E., Hecht, M., Blessing, L.: User Experience Lifecycle Model ContinUE [Continuous User Experience]. In: Proc. of BWMMS 2009, Berlin, Germany, pp. 314-317. (2008)
9. Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.-B.: User Experience Over Time. An Initial Framework. In: Proc. CHI 2009. pp. 729-738. ACM Press, New York (2009)
10. Reisenzein, R.: Emotions as metarepresentational states of mind: Naturalizing the belief-desire theory of emotion. Journal of Cognitive Systems Research. 10, 6-20 (2008)
11. Ludden, G. D. S., Schifferstein, H. N. J., & Hekkert, P. (2008). Surprise as a design strategy. *Design Issues*, *24*(2), 28-38.
12. Väänänen-Vainio-Mattila, K., Palvianen, J., Pakarinen, S., Lagerstam, E., Kangas, E.: User Perception of Wow Experiences and Design Implications for Cloud Services. In: DPPI ’11 Proceedings of the 2011 Conference on Designing Pleasurable Products, ACM Press, New York (2011)
13. Mori, H., Inoue, J.: Jigsaw Panel: A Tangible Approach for Delightful Human-Computer Interaction. Proc. of SICE Annual Conference, Sapporo, Japan. pp. 1579-1582 (2004)
14. Bradley, M. M. and Lang, P. J.: Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry, 25(1), pp. 49-59 (1994)
15. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Ziegler, J., Szwillus, G. (eds.) Mensch & Computer 2003. Interaktion in Bewegung. B.G. Teubner, Stuttgart, Leipzig, pp. 187–196 (2003)
16. Gross, A. & Bongartz, S.: Why do I like it? Investigating the product specificity of User Experience. In: Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design (NordiCHI '12). ACM, New York, 322-330 (2012)

Board Recommendation in Pinterest

Krishna Y. Kamath[1], Ana-Maria Popescu[2] and James Caverlee[1]

[1] Texas A&M University, College Station TX 77840, USA,
krishna.kamath@gmail.com, caverlee@cse.tamu.edu
[2] Research Consulting
anamariapopescug@gmail.com

Abstract. In this paper we describe preliminary approaches for content-based recommendation of Pinterest boards to users. We describe our representation and features for Pinterest boards and users, together with a supervised recommendation model. We observe that features based on latent topics lead to better performance than features based on user-assigned Pinterest categories. We also find that using social signals (repins, likes, etc.) can improve recommendation quality.

Keywords: recommendation, social network, interest network

1 Introduction

This paper focuses on the task of *recommending relevant boards to Pinterest users*. Pinterest is a fast-growing interest network with significant user engagement and monetization potential. One of the important aspects of Pinterest is encouraging pinning activity by recommending relevant, high-quality information to the site's users. We use a content-based filtering approach and report encouraging initial results. More specifically, we focus on three aspects of content recommendation for Pinterest boards. First, we describe our representation and features for Pinterest boards and users. Second, we describe our computation of potential board *relevance* to a user based on given features. Finally, we describe a supervised recommendation model which incorporates various relevance scores for good overall performance.

2 Related Work

Item recommendation is a well-studied problem [1]; general recommendation approaches include collaborative filtering [2], content-based filtering [11] or hybrid approaches [9]. Recently, recommender systems for users and content (tweets, topics, tags, etc.) in social networks have become an active area of interest [7, 8, 17, 14, 5, 15, 12]. Our work focuses on a particular recommendation task specific to Pinterest, a newer interest network, and leverages insights from both content-based filtering and from user modeling for social content recommendation. Pinterest is receiving additional attention from the research community, with recent work investigating other aspects such as global site analysis [3], gender roles and behaviors [13] and initial content quality measures [6].

3 Boards, Users and Board Relevance

In the following we describe our representation for Pinterest boards and users as well as our approach for assessing user-specific board relevance.
Let U be the set of Pinterest users, B the set of boards and $B_u \subseteq B$ be the set of boards created by user u. Each board is represented by means of a vector $\mathbf{b}$:

$$\mathbf{b} = \langle f_1, f_2, \ldots, f_n \rangle \tag{1}$$

where, $f_1, f_2, \ldots, f_n$ are features of $\mathbf{b}$ extracted from Pinterest data. Each user is represented as the mean of the board vectors for his set of boards B_u:

$$\mathbf{u} = \frac{1}{|B_u|} \langle \sum_{B_u} f_1, \sum_{B_u} f_2, \ldots, \sum_{B_u} f_n \rangle \tag{2}$$

To make sure this method of representing a user accurately captures his interests we exclude *community* boards from B_u. Pinterest users can turn a board into a *community* board by allowing others to pin to it. In previous experiments related to our recent work [6], we found that community boards have very high topical diversity and do not necessarily reflect the user's category-specific interest.
Given the above representation for a user u and board b, we can compute a measure of b's relevance to u by computing the cosine similarity between their corresponding vectors.

4 Feature Space

This section gives an overview of the features used to represent boards and users. We employ both *local* features (derived from a single board) and *global* features derived by leveraging a large set of boards.

4.1 Local Feature Extraction Methods

We use two methods to extract features directly from a given board by employing the user-supplied category label and, respectively, the board's pins.
Features From Board Category: When creating a board, users can assign to it one of 32 fixed categories (e.g., Art, Technology). Each board can be represented as a vector in a 32-dimensional space -e.g., a board in the Art category can be represented by a vector $\langle 1, 0, 0, \ldots, 0 \rangle$, where the first dimension corresponds to Art. A user vector is derived by combining board vectors as in (2).
Features From Pin Descriptions: Pins usually have free-text descriptions. A board can be represented as a vector using the bag-of-words model based on the content of the descriptions for all the board pins. Board vectors are again used to derived a final user vector as in (2).

4.2 Global Feature Extraction Methods

We next describe the use of information outside of a given board's content for feature extraction: (i) we account for Pinterest users interacting with a board and its owner; and (ii) we annotate a board with latent topics from a set learned from a collection of Pinterest boards.

Features From Social Interactions: We are interested in the social impact of a candidate board which may indicate the board is useful and recommendation worthy. We define the board *social score* as a linear function of its social impact (S_b) and the board user's social impact (S_u): $SocialScore(b) = w_b \cdot S_b + w_u \cdot S_u$. In later experiments we use $w_b = 0.9$ and $w_u = 0.1$. S_b is determined using social annotations from other users, in the form of repins, likes and follower count [3].

$$\begin{aligned} S_b = {} & w_{\text{re-pins}} \cdot \mathcal{F}(\text{mean re-repins for b}) \cdot \mathcal{F}(\text{std. re-repins for b}) + \\ & w_{\text{likes}} \cdot \mathcal{F}(\text{mean likes for b}) \cdot \mathcal{F}(\text{std. likes for b}) + \\ & w_{\text{followers}} \cdot \mathcal{F}(\text{\# of board followers}) \cdot \frac{\text{\# of board followers}}{\text{\# of user's followers}} + \\ & w_{\text{pins}} \cdot \mathcal{F}(\text{\# of pins on board}) \end{aligned}$$

where, $\mathcal{F}$ is a function maps which maps a real number to a value in $[0, 1]$ and the weights sum to 1. We experimented with logistic and double logistic functions for $\mathcal{F}$. Using this definition for S_b, we determine user's impact as:

$$\begin{aligned} S_u = {} & w_{\text{board scores}} \cdot [\text{Mean of social impact } (S_b) \text{ for all boards of } u\ ] + \\ & w_{\text{followers}} \cdot \mathcal{F}(\text{\# of user's followers}) + w_{\text{followers}} \cdot \mathcal{F}(\text{\# of boards}) \end{aligned}$$

Features From Latent Dirichlet Allocation (LDA): We previously described using Pinterest's board categories. However, users frequently skip the labeling step[4]. Additionally, generic categories (Outdoors, DIY & Crafts) lead to only a surface understanding of the board's content. These two reasons motivate us to also use features based on latent or hidden topics present in a board. Inspired by past work [18], we experiment with a *LDA-based topic discovery* method [19]. We generate one document per board by concatenating the board description, title, and pin descriptions. Topics are learned from a training set of 25,000 boards ($>$ 9 pins each), and the learned model is used to label test boards. We compared LDA methods with two different values for number of topics - 100 and 200 and found that LDA with 200 (LDA-200) topics discovered latent topics on Pinterest better [6]. Hence, we used it to extract features from Pinterest to represent board vectors. Given a board, we first find board topics using LDA-200. We then represent the board as a vector in 200 dimensions, each for one topics in the LDA model. The user vector is then determined using (2).

5 Supervised Board Recommendation

We now describe our initial results for the task of recommending boards to Pinterest users. We describe our dataset, the *supervised board recommendation* framework and two sets of experiments.

[3] We include information about board size to penalize very sparse boards

[4] In our experience, with $>$ 290,000 crawled boards, 47% lacked a user assigned category.

5.1 Data

For our analysis, we started with a sample of 4032 users and sampled 18,998 of their boards. We then extracted features from these boards, using the four methods we described in Section 4, and built the corresponding board vectors. While using LDA-200 to discover topical features, we found that we could determine vectors for only 14,543 (or 72%) of the boards. We analyzed the remaining boards and found that they were either very sparse (61% of the rest had at most 5 pins) or too incoherent; in some cases, topics outside of the learned set were required (e.g., a WWE board). Note that given the output of the LDA inference step for a test board, we only retain *core topics*, i.e. topics whose probability is greater than a threshold (0.05). Hence, for our experiments we used a dataset consisting of 4032 users and 14,543 boards.

5.2 Supervised Board Recommendation

We now describe our board recommendation approach. Initially, we directly used the cosine similarity score to determine board-user similarity and recommend boards to users. However, this approach was not very effective, especially when combining different types of information (e.g., pin descriptions and LDA topics). Hence, we experimented with a supervised approach to board recommendation.

Generating Labeled Data: For scalability purposes, we automatically generated labeled data. We used a balanced data set with 50% positive and 50% negative recommendation examples. A second evaluation of a model trained on such data and used to produce recommendations judged *manually* will confirm the quality of the automatically derived labeled set. To obtain the labeled data, we first generate a set of similarity scores for each available (board, user) pair. Each corresponds to a class of basic features (e.g., LDA topics,etc.). We then select top$-k$ and bottom-k board-user pairs for each type of similarity score as *positive* and respectively *negative* examples. For each example in the final set, the attributes are represented by the similarity types (and their values by the similarity scores). Given a specific k, we generate a labeled dataset with $2k \times 4 = 8k$ labeled instances. For the experiments below, we set $k = 1000$ to generate a balanced set of 8000 recommendation examples.

Learning Recommendation Models: We employ the labeled data for learning recommendation models. We experimented with an SVM-based regression model; given a test example, the model will assign a score indicating a potentially good recommendation (if close to 1) or a bad recommendation (if close to 0). Potential board suggestions can be ranked according to the predicted score. For one of our evaluations we also used SVM-based classification to make a binary decision about a board being a good or bad suggestion for a user.

5.3 Experiments

We evaluate the value of the various feature classes (and their combinations) for board recommendation. In addition to methods testing the 4 feature classes in Section 4, we evaluated 2 other methods combining feature classes. The first, (*non-soc*), combines features based on board categories, pin descriptions and LDA topics, while the second (*all*) assesses the added impact of social features.

Table 1. Results: Feature classes' contributions to board recommendation quality. Combining feature classes and including social signals improves performance.

Method	F1	AUC	UIM	Compare
Board category (cat)	0.60 (0%, 1.00)	0.69 (0%, 1.00)	0.33 (0%, 1.00)	cat
Pin description (pin)	0.70 (17%, 0.00)	0.70 (1%, 0.05)	0.13 (-61%, 0.00)	cat
Social metrics (soc)	0.73 (22%, 0.00)	0.66 (-4%, 0.00)	0.12 (-64%, 0.00)	cat
LDA-200 topics (lda)	0.76 (27%, 0.00)	0.78 (13%, 0.00)	0.16 (-52%, 0.00)	cat
Non-social features				
non-soc: pin+cat+lda	0.83 (9%, 0.00)	0.84 (8%, 0.00)	0.22 (38%, 0.00)	lda
All features				
all: non-soc+ soc	**0.87 (5%, 0.00)**	**0.88 (5%, 0.00)**	**0.21 (-5%, 0.09)**	non-soc

We perform two types of experiments: (i) an evaluation using the automatically constructed 8000-example dataset and (ii) a second evaluation in which learned recommendation models are used to recommend boards for a small set of test users. The suggested boards are manually labeled and the various models are compared on this data.

Models: We compare 6 recommendation models. The first 4 models correspond to the 4 basic feature classes. For each such class, the resulting similarity score is used as a final aggregate feature by the model (e.g., *lda* only uses the similarity score based on LDA topics as basic features, etc.). Additionally, a mixed non-social model *non-soc* uses three similarity scores based on the pin descriptions, user-assigned categories and, respectively, latent topics. Finally, a full model *all* uses all 4 similarity scores. SVM classification is used in the first evaluation and SVM regression in the second.

Evaluation: Automatically Derived Gold Standard We start with an intrinsic evaluation using the automatically constructed balanced gold standard. We use SVM classification and the standard metrics F_1 and AUC. We also define another metric called *User Interest Match* (UIM) score which measures the match between a board labeled as *relevant* and the set of explicit user interests for u:

$$\text{UIM} = \frac{1}{|B_u^r|} \sum_{b \in B_u^r} \% \text{ of boards with category } \mathcal{C}(b) \text{ in the account of user } u$$

where, B_u^r is the set of boards recommended to a user u. Higher UIM values correspond to recommended boards from categories of particular interest to the target user. We used Student's t-test to determine stat. significance for reported improvements.

Table 1 summarizes our results. In addition to assessing each method separately, we compare it with another relevant method (indicated in last column). Single feature class methods are compared against the *cat* baseline, *non-soc* against the best single class method (*lda*) and the final *all* method against *non-soc*. The first value in the parenthesis is the % improvement w.r. to the reference method

Table 2. Results: Board recommendation evaluation with human judgments. Combining feature classes leads to better recommendations.

Method	Precision@5	Precision@10	NDCG@5	NDCG@10
Board category (cat)	0.68	0.56	0.89	0.86
Pin description (pin)	0.62	0.60	0.92	0.90
Social metrics (soc)	0.37	0.40	0.77	0.66
LDA-200 topics (lda)	0.78	0.72	0.94	0.92
Non-social (non-soc)	**0.90**	0.81	**0.98**	**0.97**
All features (all)	**0.90**	**0.82**	0.97	0.96

and the second value is the p-value from t-test. We find that: (i) *lda* performs best among single feature class methods; (ii) combining feature classes leads to better performance than using single feature types; and (iii) social interaction information improves recommendation results.

Evaluation: Human judgments In a second experiment, we evaluate the recommendation models learned on automatically generated training data using manual judgments. We set aside a subset of our labeled dataset for testing purposes. We learned 6 SVM regression models on a balanced subset of the remaining data and then used them to make board recommendations for 12 users in the test set. Specifically, we retained the $top-10$ recommendations for each of the 12 users. 2 annotators independently labeled them as *relevant* or *not relevant* (with 70.5% agreement). After resolving disagreements, we used the manual judgments to evaluate the 6 models using precision (% of good recommendations) and the normalized discounted cumulative gain (NDCG), which takes into account the rank of the recommendation as well. Table 2 summarizes the results for the 6 models using $top-5$ and $top-10$ recommended boards.
Based the results in Table 2 and the human judgments, we find that: (i) Board category labels are helpful when accurate, but if absent or wrong they can hurt the similarity score relying on this feature. The latent topics discovered by LDA lead to better performance. (ii) Not surprisingly, using social signals by themselves leads to poor performance, as they do not contribute any topical relevance information. A user who liked a popular Wedding board may not like a popular Technology board. (iii) Methods which combine features perform best - the impact of social features was muted in this smaller-scope evaluation leading to small differences between the two.

6 Conclusion

This paper investigates content-based recommendation of Pinterest boards, with a focus on 4 classes of features user for board representation. Our initial experimental results show that latent topics discovered by LDA correspond to the most valuable single feature class, but combining different feature classes leads to best overall results. Our current work focuses on better ways of incorporating direct and indirect social network information in the recommendation model.

References

1. Adomavicius, G. and Tuzhilin,A.:Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions.In: IEEE Transactions on Knowledge and Data Engineering, 2005
2. Das, A.. , Datar M, Gard A., Rajaram S.: Google News Personalization: Scalable Online Collaborative Filtering. In: Proceedings of WWW 2007.
3. Gilbert, E., Bakhshi, S., Chang, S., Terveen, L.: I Need to Try This: A Statistical Overview of Pinterest. In: Proceedings of CHI-2013.
4. Hagberg A., Schult D., Swart P.: Exploring network structure, dynamics and function using NetworkX. In: Proceedings of SciPy2008, 2008.
5. Hannon, J., Bennet, M., Smyth, B., Recommending Twitter users to follow using content and collaborative filtering approaches. In: Proceedings of RecSys10.
6. Kamath, K., Popescu, A., Caverlee, J. : Board Coherence: Non-visual Aspects of a Visual Site. In: Proceedings of WWW 2013 (poster).
7. Kyew, S.M., Lim, E., Zhu, F. : A survey of recommender systems in Twitter. In: Proceedings of SocInfo 2012
8. Liang, H., Xu, Y., Tjondronegoro, D., Christen P.: Time-aware topic recommendation based on micro-blogs. In: Proceedings of CIKM 2012
9. Melville, P., Mooney, R., Nagarajan, R., Content-Boosted Collaborative Filtering for Improved Recommendations. In Proceedings of AAAI 2002
10. Liu, Lu., Tang, J., Han, J., Jiang, M., Yang, S.: Mining Topic-level influence in heterogeneous networks. In: Proceedings of CIKM 2010.
11. Pazzani, M.J., Billsus, D.: Content-based Recommendation Systems. In: The Adaptive Web: Methods and Strategies of Web Personalization., 2007
12. Pennacchiotti, M., Silvestri, F., Vahabi, H., Venturini, R.: "Making your interests follow you on Twitter". In: Proceedings of CIKM 2012.
13. Ottoni, R., Pesce, J.P., Las Casas, D., Franciscani, G., Kumaruguru, P., Almeida, V.: Ladies First: Analyzing Gender Roles and Behaviors in Pinterest. In: Proceedings of ICWSM 2013
14. Rae, A., Sigurbjrornsson, N., van Zwol, R., Improving Tag Recommendation using Social Networks. In: Proceedings of RIAO 2010.
15. Sharma, A.: Network-centric Recommendation: Personalization with and in social networks. In: SocialCom/PASSAT, 2011.
16. Wikipedia: http://en.wikipedia.org/wiki/Pinterest
17. Yan R., Lapata, M., Li, X.: Tweet recommendation with graph co-ranking. In: Proceedings of ACL 2012
18. Hong L., and Davison B. D.: Empirical study of topic modeling in Twitter. In: Proceedings of the First Workshop on Social Media Analytics
19. Phan X.: GibbsLDA++. https://gibbslda.sourceforge.net/

Crowdsourced Evaluation of Semantic Patterns for Recommendations

Valentina Maccatrozzo, Lora Aroyo and Willem Robert van Hage

The Network Institute
Department of Computer Science
VU University Amsterdam, The Netherlands
v.maccatrozzo@vu.nl, l.m.aroyo@vu.nl, w.r.van.hage@vu.nl

Abstract. In this paper we explore the use of semantics to improve diversity in recommendations. We use semantic patterns extracted from Linked Data sources to surface new connections between items to provide diverse recommendations to the end users. We evaluate this methodology by adopting a bottom-up approach, i.e. we ask users of a crowdsourcing platform to choose a movie recommendation from among five options. We evaluate the results in terms of a diversity measure based on the semantic distance of topics and genres of the result list. The results of the experiment indicate that there are features of semantic patterns that can be used as an indicator of its suitability for the recommendation process.

1 Introduction

Recommender systems help people cope with the amount of information available on the Internet. Widely used are collaborative filtering and content-based recommender systems. The first requires a high availability of ratings spread over the collection, otherwise it tends to suggest only rated items, preventing diversity. Content-based algorithms are based on the characteristics of the items, making less rated items more accessible, but still lacks diversity [4]. We extend the existing approaches with semantic patterns to improve diversity in recommendation results. Linked Data enables us to discover connections between items that otherwise would not surface. We use pattern frequency statistics in the linked datasets as indicators of the ability of patterns to produce recommendations. The goal of this experiment is to find the correlation between the objective statistical measures of patterns in linked data sources and the subjective user perception of their usefulness in order to define user-centered measures of relevance of the recommendations. We do this by performing the following steps: (1) identify relevant patterns in datasets, (2) define recommendation algorithms using these patterns, (3) evaluate with the crowd. This paper reports about the initial results on these contributions.

2 Related Work

Recommender systems developed upon Semantic Web Technologies were developed by Di Noia et al., who present a content-based recommender based only on Linked Data sources, showing its potentiality [5]. Their approach do not make use of content patterns. Oufaida and Nouali [10] propose a multi-view recommendation engine that integrates collaborative filtering with social and semantic recommendation. Our approach aligns more with the work of Aroyo et al. on a content-based semantic art recommender, where [1] explores a number of semantic relationships and patterns.

Semantic patterns as we define them share some similarities with the approach proposed by Sun et al. in [16] to define a path-based semantic similarity. However, our definition of patterns relies more on the work of Gangemi and Presutti [6], who introduce knowledge patterns to deal with the semantic heterogeneity of ontologies. Presutti et al. [12] used such patterns to analyze Linked Data, as a new level of abstraction. In this work, we define such semantic patterns for the purpose of diversity in recommendations.

The use of crowdsourcing for collecting users' contributions has been explored by different works. For instance, Kittur et al. [8] present an exploratory study to show how the experimental design influence the quality of the contributions, we follow their best practices. Crowdsourcing has been used also to build ground truth data by Aroyo and Welty in [2]. Also Sarasua et al. [13] make an interesting use of crowdsourcing for ontology alignment.

3 Semantic Patterns in Recommendations

In ontologies, patterns can emerge in the combination of data instances, the types of these instances, and the links created by the properties. A semantic pattern connects a source type T_1 with a target type T_{l+1} through steps consisting of property-type pairs. This can be formulated as an ordered set: $\{T_1, P_1, T_2, P_2, ..., T_l, P_l, T_{l+1}\}$. The length of the pattern is given by l. The type of the pattern depends on the instantiation of type T_2 to T_l, e.g. people pattern, etc. Patterns are called homogenous when T_2 to T_l are of the same type and heterogenous when the types are different. The workflow we define utilizes such patterns for recommendation purposes: we extract and select patterns suitable for recommendations, performing specific analysis, and we produce recommendations ranked by the diversity measure we define.

Extraction & Selection of Patterns The sources where patterns can be discovered provide numerous candidates, hence it is critical to develop strategies to select relevant patterns. We perform a statistical analysis on the relation occurrences to select candidate patterns on the basis of their frequency, e.g. how many times the pattern is instantiated. Frequencies are calculated in two ways: considering only the properties involved in the pattern (*property frequency*), and considering also the types involved (*type frequency*). The property frequency is considered *global*,

when calculated on the whole source, and *local*, when calculated in relation to an instance. For this experiment, we select patterns using different combinations of frequencies in order to test the correlation between frequencies and users' evaluations. We order the patterns on the basis of the property frequencies and we selected 6 patterns per frequency type: the two most frequent, the two less frequent and the two in the middle.

Diversity measure Diversity in recommendations is usually defined to be applied to list of items, aiming at reducing the number of similar items in the result set [14,17,7]. On the contrary, we designed a measure that does not require a list of recommendations because it is calculated with respect to the items in the user profile, hence, it can be applied also to single recommendations. This measure is defined upon the concept of semantic similarity, in a similar fashion of Middleton et al. [9] and Bogdanov et al. [3]. It allows us to suggest movies which are not exactly the users' favorites, but that are still related to them. We can consider all the metadata about a movie which consists of nouns (i.e. genre, topic, synopsis). Using relevance feedback we can identify the right value of diversity per metadata up to the right balance. Given two programs, p_1 and p_2, to calculate the measure we (1) extract genre and topic of p_1 and p_2; (2) calculate the semantic similarity between genres and topics; (3) calculate the diversity as one minus the semantic similarity; (4) calculate the diversity measure as the average of the previous ones. We use the Wu & Palmer measure [18], but other measures are possible as well.

$$Div(p_1, p_2) = \frac{(1 - sim(genre(p_1), genre(p_2))) + (1 - sim(topic(p_1), topic(p_2)))}{2}$$

4 Experimental Design

The experiment was performed on the platform CrowdFlower[1] to collect user feedback about recommendations generated using a selection of semantic patterns extracted from DBpedia[2]. We ask the users to select a match for a given movie from among five options, providing poster and synopsis. We proposed the following context: "*You are buying a movie for a friend and you want to get the"buy one, get two" promotion. Which of the following movies would you match with the starting one in order to surprise your friend with something not trivially related?*". Four options are defined with semantic patterns and ranked with IMDB ratings. We used IMDB to improve the probability of users knowing the movie to test different values of our diversity measure, as shown in Fig. 1. The fifth option is chosen from the Amazon[3] recommendations as a baseline to compare our performances. The options are in randomized order to avoid bias effects. We also ask the users to explain their choice, to obtain an indication on how they made it and to identify potential spammers. Additionally, we ask

[1] `http://crowdflower.com`
[2] `http://dbpedia.org`
[3] `http://amazon.com`

the users to type the third word in the synopsis of the movie they chose, as an additional spammer detecting question, following the best practices suggested by [8]. In particular spammers are supposed not to put any effort in the task, hence open questions are filled in with nonsense lists of characters. We use a bottom-up approach, i.e. instead of asking users to evaluate a recommendation, we ask them to choose it. In this way we try to be less intrusive as possible in affecting the users' choice.

Table 1: Generic example of options with related patterns.

Starting movie	Pattern	Selected Movie
The Devil Wears Prada	Amazon	Confessions of a Shopaholic
	Starring - Narrator	The Living Sea
	Writer	We Bought a Zoo
	Producer	Forrest Gump
	Set Location	The Bourne Ultimatum

Table 1 shows an example of the five options, starting with the Amazon recommendation, followed by the pattern *starring-narrator*, i.e. an actor in the starting movie performs as the narrator in the suggested movie. The last three options are movies that share the same properties with the starting movie: the writer, the producer, and the set location.

5 Results

We chose 12 movies of three different genres (thriller, history and crime), and selected 12 people patterns (i.e. patterns which involves only types "person") per movie. We built 36 tests and we collected 720 contributions (one contribution per user). 28 spammers were identified and eliminated from the results.

By comparing the results with the Amazon recommendation, all those suggestions that received an high number of votes (on average 27) are also reachable through semantic patterns, namely the starring pattern and the director patterns of length 2. The other Amazon recommendations received a low number of votes (on average 5.3) and performed clearly worse than the semantic patterns ones. This is an interesting result: our method can provide recommendations that can satisfy multiple needs. In order to evaluate the performances in these terms, we consider the explanation for the choices provided by the users. Although we asked the users to address diversity, the explanations show that this was not always what drove them. So, we clustered the choices on the basis of the users' comments into three categories: similar, different and not applicable (i.e. difficult to assess). Three patterns resulted peculiar for recommendations in the category '*different*': cinematographer-director, cinematographer-child-cinematographer and director-editing.

In Fig. 1, we can see the distribution of the diversity values over the movies used in the experiment. In the top right corner there are the movies that are more

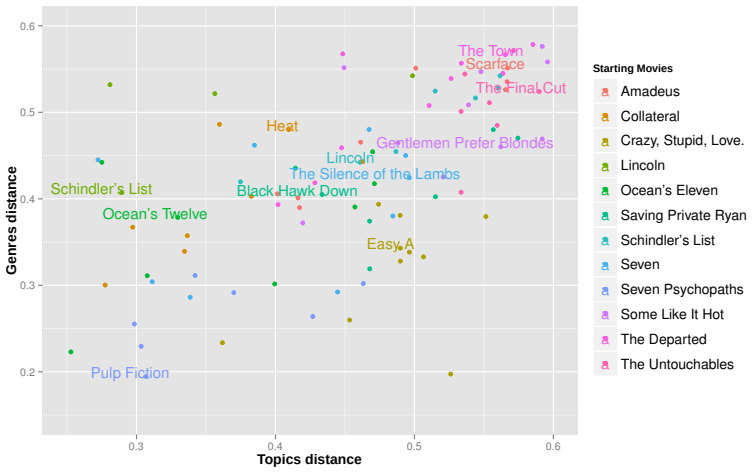


Fig. 1: For every starting movie, the diversity of recommendations is shown. In (0:0) there are the genre and topic of the starting movie. The labelled movies are the most chosen ones.

different from the starting one. Users that chose those movies did not always perceive this diversity, and they often disagree. For instance, the comments from two users who chose the pair Amadeus - Scarface were quite different. One user says: "Both movies are about the life of times of the lead characters.", hinting at similarity. The other user says: "Pairing Scarface with Amadeus would be a surprise. Both films are American classics and contain amazing performances. Both films are biographical in nature as well. However, Amadeus is a "period" film set in Austria and features classic works by Mozart. It's joyous and moving. Scarface is a crime drama focused on the dark underbelly of the drug cartels. It's big moments and shocks come not from musical masterpieces, but brutal violence.", hinting at diversity. This suggests that the perception of the diversity is highly correlated with the users' knowledge of the movies, and attitude towards the task as well. However, this topic requires more investigation, which will be addressed in the future.

6 Analysis and Discussion

Our aim is to determine the most important features of a pattern to deliver meaningful recommendations. We consider local and global property frequencies, type frequencies, and length of the patterns. We perform correlation tests between the features and the users' feedback, using Spearman rank correlation test [15]. The results of this preliminary analysis show that there is a correlation between features of the semantic patterns and users' feedback. In particular, the global property frequency is positively correlated (0.32) to the users' feedback,

Table 2: Correlations between pattern features and users' feedback.

Feature	Correlation	p-value	Significance
Global property frequency	**0.32**	**7.921e-08**	**99% confidence level**
Local property frequency	0.19	0.001326	99% confidence level
Type frequency	0.23	0.0001292	99% confidence level
Length	**-0.35**	**1.616e-09**	**99% confidence level**
All features	-0.35	3.649e-09	99% confidence level
Global & Local property frequencies	-0.29	6.409e-07	99% confidence level
Global property & Type frequencies	-0.34	1.073e-08	99% confidence level
Global property frequency & Length	**-0.40**	**9.629e-12**	**99% confidence level**
Local property & Type frequencies	-0.20	0.0007238	99% confidence level
Local property frequency & Length	-0.36	1.08e-09	99% confidence level
Type frequency & Length	0.39	3.799e-11	99% confidence level
Global & Local property &Type frequencies	-0.28	1.977e-06	99% confidence level
Global & Local property frequencies & Length	-0.38	1.119e-10	99% confidence level
Local property & Type frequencies & Length	-0.30	3.171e-07	99% confidence level

i.e. the more frequent the pattern in the source, the more suitable it is for recommendations. The length of the pattern is, instead, negatively correlated (-0.35) to the users' feedback, i.e. longer patterns introduce too vague links between items, which seems not relevant for users. We performed the Principal Component Analysis [11] on the results to test different combination of the features. A combination of the global property frequency and the length of the pattern increased the correlation up to 0.40, confirming the prominence of these features in the prediction of the pattern usefulness in the recommendation process. These numbers represent a moderate correlation, however, given the limited size of the experiment, both in terms of patterns and users, and the fact that we do not take into consideration users' profile, these numbers are indicators for further research. In Table 2 we report the correlations, the p-value of the tests and their significance. In all cases we can reject the null hypothesis, i.e. all the correlation coefficients are significantly different from zero.

7 Future Work

We aim at improving our results by exploring other patterns features, as well as other sources, e.g. IMDB. We plan to perform larger scale experiments in order to compare general and domain specific vocabularies and analyze their differences in terms of patterns and coverage of items. Further, we will study the user perceived importance of each of the candidate patterns for the recommendation relevance and diversity, taking into consideration users' profiles.

Acknowledgments. This research is supported by the FP7 STREP "ViSTA-TV" project. We would also like to thank our colleague Chris Dijkshoorn for the valuable contribution for the design of the experiment.

References

1. L. Aroyo, N. Stash, Y. Wang, P. Gorgels, and L. Rutledge. CHIP Demonstrator: Semantics-Driven Recommendations and Museum Tour Generation. In *ISWC2007*, pages 879–886, 2007.
2. L. Aroyo and C. Welty. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *WebSci2013*. ACM, 2013.
3. D. Bogdanov, M. Haro, F. Fuhrmann, E. Gómez, and P. Herrera. Content-based music recommendation based on user preference examples. In *Womrad 2010*, 2010.
4. K. Bradley and B. Smyth. Improving Recommendation Diversity. In *The 12th Irish Conf. on Artificial Intelligence and Cognitive Science*, pages 85–94, 2001.
5. T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker. Linked open data to support content-based recommender systems. In *I-SEMANTICS '12*, pages 1–8. ACM, 2012.
6. A. Gangemi and V. Presutti. Towards a pattern science for the Semantic Web. *Semantic Web - Interoperability Usability Applicability*, 1:61–68, 2010.
7. N. Hurley and M. Zhang. Novelty and diversity in top-n recommendation – analysis and evaluation. *ACM Trans. Internet Technol.*, 10(4):14:1–14:30, 2011.
8. A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *CHI*, pages 453–456. ACM, 2008.
9. S. E. Middleton, N. R. Shadbolt, and D. C. De Roure. Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.*, 22(1):54–88, January 2004.
10. H. Oufaida and O. Nouali. Exploiting Semantic Web Technologies for Recommender Systems: A Multi View Recommendation Engine. In *ITWP 2009*, 2009.
11. K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11):559–572, 1901.
12. V. Presutti, L. Aroyo, A. Adamou, B. Schopman, A. Gangemi, and G. Schreiber. Extracting Core Knowledge from Linked Data. In *COLD2011*, 2011.
13. C. Sarasua, E. Simperl, and N. Noy. CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. In *ISWC*, pages 525–541. Springer, 2012.
14. B. Smyth and P. McClave. Similarity vs. diversity. In *Case-Based Reasoning Research and Development*. Springer, 2001.
15. C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15:72–101, 1904.
16. Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003, 2011.
17. S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys '11*, pages 109–116. ACM, 2011.
18. Wu, Z. and Palmer, M. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, 1994.

Utilizing Social Networks for User Model Priming: User Attitudes

Adam Moore[1], Gudrun Wesiak[2], Christina M. Steiner[2], Claudia Hauff[3], Declan Dagger[4], Gary Donohoe[5] & Owen Conlan[1]

[1]KDEG, School of Computer Science and Statistics, Trinity College, Dublin, Ireland
{mooread,owen.conlan}@scss.tcd.ie
[2]Knowledge Technologies Institute, Graz University of Technology, Austria
{gudrun.wesiak,christina.steiner}@tugraz.at
[3]Delft University of Technology, the Netherlands
c.hauff@tudelft.nl
[4]EmpowerTheUser, Trinity Technology & Enterprise Campus, The Tower, Dublin, Ireland
declan.dagger@empowertheuser.com
[5]Department of Psychiatry, School of Medicine, Trinity College, Dublin, Ireland
DONOGHUG@tcd.ie

Abstract. Research on user modeling based on social network information has shown that some user characteristics can be accurately inferred from users' digital traces. This kind of information can be used to inform user models of adaptive systems for personalizing the system. This paper addresses a crucial question for practical application of this approach: *Are users actually willing to provide their social Web profiles and how do they perceive this?* An empirical study conducted with medical students shows that although participants are using social networks, they are reluctant about providing their identities and consider these portals rather private. The outcomes of the study uncover a clear need for further research on enhanced privacy and enhanced trust.

Keywords: user modeling, empirical study, social networks, privacy, user acceptance

1 Introduction

In our increasingly technology driven world, adaptation and personalization technologies make for a more customized, user-centric interaction with often impersonal interfaces. However, the sources of information that drives this adaptation, informing a user model that a system can utilize, are often burdensome or themselves derivative and impersonal. On the other hand, a customized experience can be created by filling in long (often deeply personal) questionnaires. In order to address this, researchers have looked to the open digital traces left on the social Web. Public and semi-public portals such as Twitter and Facebook expose personal details and preferences that can be used to inform underlying models about individuals, harvested and processed automatically and then applied to create a tailored experience.

Researchers have attempted to infer a diverse set of user characteristics, mostly from Twitter streams (due to the open nature of the portal and the ease of data collection), which have often led to algorithms with surprisingly high accuracy. In [1] the political affiliation of users in the United States was predicted with more than 90% accuracy, while in [2] the user's gender could be estimated with a similar success rate. The prediction of higher-level user characteristics, including the user's topical interests from their tweets [3] and the user's personality profile based on their Twitter [4] and Facebook [5] activities have also been investigated, though the prediction of such high-level concepts has proven to be more challenging. In all the studies presented, one important aspect of the research was the identification of the necessary user data (i.e. the user's tweets or the user's photos on Flickr) and the derivation of the ground truth (i.e. gender, political affiliation, etc.). This is usually achieved by collecting the publicly available data of random users and by manually annotating the streams with respect to the wanted characteristic(s). This means, that for such research on public streams, users are usually not explicitly asked about their willingness to participate.

This, however, leaves an open question when these user characteristics are to be employed in practice, i.e. in a working system: Are users, who use the system, actually willing to provide us with their social Web profiles? It is well known that, on the one hand, users appreciate personalized information but, on the other hand, they are very concerned about privacy and that large amounts of personal information may be tracked and made accessible to other users [6]. It has also been shown that social media are deeply integrated into users' daily lives and routines [7]. As a result, privacy attitudes (as indicated in surveys) and privacy behaviors often differ [8]. This so-called "privacy paradox" [9] is evident when comparing social network (SN) users' self-reports on their understanding of caution with regard to privacy settings and their actual lack of utilizing possibilities to change the typically very lax default settings in SNs [7]. Thus, very often the benefit of using SNs for communication or personalized contents (derived from user models) for web queries or commercial ventures outweighs the perceived privacy concerns. However, most commercial personalized web-based systems do not ask users to provide their information, but simply track them from their digital traces. Users themselves are mostly not aware of the comprehensive records search engines capture by integrating different Web 2.0 services such as Flickr, Yahoo, Twitter, etc. [10]. Thus the question arises: "*how does the privacy paradox take effect when users are explicitly asked whether their SN information may be used for personalization purposes?*"

In order to investigate students' attitudes towards providing SN information for personalizing their learning experience, we conducted a survey with medical students that were to be using an adaptive experiential training simulation, requesting a number of pieces of information on their usage and attitudes to social media.

2 Empirical Study

2.1 Method

Data on social network usage was collected in the context of a larger study on the EmpowerTheUser[1] RolePlay Simulation Platform for medical interview training.

Participants. 152 students from Trinity College Dublin participated in the study as part of their third year medical curriculum. They were sent an email requesting their participation in an online survey. 95 students (a response rate of 62.5%) filled out at least one complete section of the survey. They were on average 22.81 years old (SD = 3.79) ranging from 19 to 45 years. Half of the participants were male, half female (47 each), one participant did not indicate his or her gender.

Instruments and Procedure. Data collection was carried out over four weeks during the spring semester of 2013. Students were requested to complete the survey before starting interaction with the simulator.

Besides demographic data, a question concerning students' daily internet usage, and standard questionnaires to cover personality traits (SSP, Swedish Universities Scales of Personality [11]), learning styles (ILS, Felder-Solomon Index of Learning Styles [12]), and metacognitive awareness (MAI, Metacognitive Awareness Inventors [13]) were used. Users' perceptions of privacy, trust, and accuracy of information in Social Networks (SN) were measured by means of 12 questions. The questions differentiated between five SN: Facebook, Twitter, LinkedIn, Flickr, and MySpace.

2.2 Results

Usage. Whereas 81% of the students use Facebook, Twitter is used by only 20%, LinkedIn by 5.3% and Flickr and MySpace by only 1 person each. This basically reflects the general world wide usage of these networks[2]. Considering the usage pattern of our sample, in the following, only results for Facebook, Twitter, and LinkedIn are reported. From the 77 Facebook users only 11 (14.3%) provided their username, Twitter and LinkedIn usernames were provided by 3 (15.8%) and one person (20%), respectively. In total IDs from 13 different persons were provided.

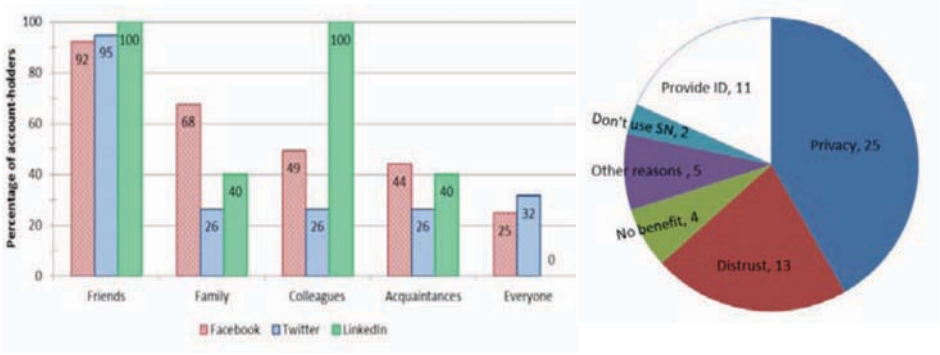
Figure 1(a) shows the kind of people participants intentionally interact with on different social networks. Numbers represent the percentage of account holders selecting an option. Independent of the SN used, almost all of the participants use these networks to communicate with friends (92-100%). Almost half of the Facebook users also interact with colleagues and acquaintances (all ≥ 44%), whereas only 26% of the Twitter users indicated to interact with those groups. On the other hand, all 5 participants with a LinkedIn account said they interact with colleagues.

Interestingly, participants who indicated using their SN accounts to interact with colleagues, acquaintances, and even everyone, were still reluctant to provide their

[1] http://www.etu.ie/

[2] http://en.wikipedia.org/wiki/{Facebook,Twitter,LinkedIn,Flickr,Myspace}

social network identities (SN-IDs) for research purposes. In the training scenario that followed the survey, none of the participants provided their social ID.



(a) (b)

Figure 1. (a) Person-groups participants interact with in different social networks (N = 95) and (b) reasons for (not) providing SN-IDs (incl. frequencies of entries).

Privacy and Trust. Questions on the perception of SNs were answered by 75 to 77 students for Facebook and 18 to 29 for Twitter. In the following medians (Md) are reported as a measure for central tendency. Although students perceive their postings on social networks as rather open (*Md* = 3 on a 4-pt. scale), most of them are either 'nervous' about providing their username or simply state they would not provide it (*Md* = 4 on a 4-pt. scale). They are also rather suspicious of people and companies using their SN postings for research or commercial ventures (*Md* = 4 on a 5-pt. scale). With respect to the representation of their own personality, students think that the portrayal of their personality is partially true and that others would get a medium accurate picture of them based on their posts (both *Md* = 3 on 5-pt. scales).

Finally, students were asked about their feelings towards providing SN-IDs in order to benefit from a more personalized learning experience and whether they trust the ETU operators that their personal information will not be used for any other purpose. With *Md* = 2 (4-pt. scales) participants indicated once more that they don't feel good about providing their SN-IDs, even if it is for their own learning benefit and that they rather distrust the simulation operators. To check whether there are gender differences in the perception of SN, independent samples t-tests have been calculated. Summarizing, male participants evaluate SN-postings as less secure and private, and rate the accuracy of deducing gender as less accurate than their female colleagues.

A last question in open answer format prompted students to relate their reasoning behind how they feel about providing their SN-ID. Open responses from 60 participants were collected, of which 49 or 83% explained why they did not want to provide their SN-ID, whereas the remaining persons gave a reason for providing their ID. All answers were analyzed and sorted into 15 categories (aggregated to 6 categories in Figure 1(b)). Most comments (overall 25 entries) concerned the privacy of SN accounts, i.e. participants use them mainly to connect to friends and family (10 entries), view SNs as something private (9), and want to separate their private life from educational or business life (6). Another group of 13 students stated that they

don't know and don't trust the people behind the survey (6), that they are insecure about what happens with the information from their SN accounts (4), and that they don't want strangers going through their personal information, postings, or pictures (3). Furthermore, students commented that they don't see any benefit in using their SN information (4 entries), especially because they believe that is not related to their true or their "educational" personality, that they want to remain anonymous (2 entries), or simply that they don't see any reason to provide their ID (3). On the other hand, students who did provide their SN-IDs stated that they don't have anything to hide and that their information on the respective networks is not too personal (3 entries), or that they are simply fine with providing it (3 entries). Other users stated that they wanted to help (3) or that they hope to benefit from providing it (2). Two participants provided their Twitter or LinkedIn but not their Facebook ID (since it contains more personal information).

In order to find out how participants' attitudes are related among each other and whether there are any connections to their personality, learning style or metacognitive awareness, responses on the relevant scales were correlated by means of Spearman's Rho coefficient (for ordinal data)[3]. Summarized, the data show that more comfort in providing SN-ID relates to a higher perception that networks are open, more comfort with the use of information for research or commercial ventures, a better feeling of providing one's ID for a benefit regarding learning experience, and more trust that the simulator operators will not use the gained information for any other purposes. Furthermore, participants who think that SNs are very open also believe that SNs do not give a realistic, complete or accurate picture of them and have more trust in the simulator operators. Users who think that the picture derived from their posts is accurate are less comfortable with the use of their SN information for research or commercial purposes. On the other hand, participants who are comfortable with giving away information for research or commercial venture also feel good providing their ID to benefit from more personalized learning experiences and have also more trust that simulator operators will not use their information for other purposes.

A look at daily internet usage, personality traits, or learning styles did not reveal any meaningful relationships. However, students' metacognitive awareness is closely related to their trust in the simulator operators. More specifically, students who have high scores on the monitor and evaluation scales, as well as a high overall regulation of cognition score, indicated a stronger distrust in simulator operators.

Perceptions of Information Inferred from Social Network Posts. Participants were also asked to indicate how accurately they think 10 different traits can be deduced from their social network activities. Most students believe that gender, university degree course, and highest educational degree can be very accurately deduced from their SN ($Md = 5$), age, nationality, and personality somewhat accurately ($Md = 4$), whereas political convictions, income, car model, and music taste cannot be inferred from their SN ($Md \leq 2$).

[3] *Note:* results are reported only for significant correlations derived from Spearman's Rho (with $p < .05$); correlations are either for Facebook, Twitter, or both.

3 Discussion and Conclusion

From a sample of nearly 100 respondents, it became clear that, although they are active on social networks, they do not consider them a place for information to be gathered that could be useful in tailoring training to their individual needs.

With metacognitive awareness being positively correlated with a definitive unwillingness to share this information, there is clearly a need to find ways to increase the trust learners have in the people behind the learning environment they are using.

It is interesting to note the perception of both the privacy and information that can be derived from an active social network account. For the majority of traits, the participants' intuition about how well they can be estimated from the SN is in line with existing research and SN are perceived as rather open. Nevertheless, privacy is a great concern. This, therefore, presents somewhat of a dilemma for researchers and practitioners in adaptation technologies. We can now, with reasonable accuracy, infer and predict many aspects of our systems' users from their traces on open, publically available channels. However, when directly questioned about this approach, our users are reluctant to disclose their identities within these networks (information that can often actually be obtained without their consent), express discomfort and, when asked directly in the training simulator to provide this information for an illustrated educational benefit, exactly zero of our cohort of 152 did so. Thus, in contrast to the privacy paradox concerning users' reported attitudes and behavior [8][9], our sample was very consistent in their reported unwillingness to provide their SN-information and their actual behavior. The paradox, though, lies in the fact that although users are willing to disclose personal information on their SN, they feel uncomfortable providing this information to personalize their learning. Clearly, more work is needed to bridge the gap between perceived usage and audience of these portals and those hoping to use the information contained within to provide benefit to its participants.

In summary, it seems that our study participants view their SN mainly as a means to connect with friends and family and thus as something that should not be linked to their professional development and training. In the same line, willingness to provide their SN-IDs is closely related to the belief, that the networks are very open anyway, and that their true personality cannot be derived from their posts. Personality typing gave entirely normative responses, with no indication of overly cautious or private characteristics. However, it is known that privacy concerns increase with higher age, education, and income [6]. The reluctance about utilizing SN information for user modeling might, to some extent also be correlated with the students' background; thus investigating cohorts from different disciplines, like computer or information science, would be desirable. In line with [6], in order to use information from SN portals, it seems key to explicitly explain to users what kind of information is used, how the information is extracted, and how exactly participants could benefit from providing their network IDs. Also knowledge and control over the used information fosters users' willingness to disclose personal information. In addition some basic information about those making use of the information should help to build up trust into the diligent handling of their information. Future research needs to focus on conclusive ways to convey the benefits for the users and to give them more control and insight on the actually utilized body of information.

Acknowledgments. The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no 257831 (ImREAL project).

References

1. Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of users. In SocialCom/PASSAT, pp. 192-199. IEEE (2011)
2. Burger, J. D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1301-1309. Association for Computational Linguistics (2011)
3. Michelson, M., Macskassy, S. A.: Discovering users' topics of interest on Twitter: a first look. Proc 4th w/s on Analytics for noisy unstructured text data, pp. 73-80. ACM (2010)
4. Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our Twitter profiles, our selves: Predicting personality with Twitter. In SocialCom/PASSAT, pp. 180-185. IEEE (2011)
5. Wald, R., Khoshgoftaar, T., Sumner, C.: Machine prediction of personality from Facebook profiles. In 13th International Conference on Information Reuse and Integration (IRI), pp. 109-115. IEEE (2012)
6. Kobsa, A.: Privacy-Enhanced Web Personalization. In P. Bruslovsky, A. Kobsa, W. Nejdl (eds.). The Adaptive Web: Methods and Strategies of Web Personalization. Berlin, Heidelberg, New York: Springer Verlag, pp. 628-670 (2007).
7. Debatin, B., Lovejoy, J.P., Horn, A.-K., Hughes, B.N.: Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences. Journal of Computer-Mediated Communication 15, pp 83-108 (2009).
8. Stutzman, F., Kramer-Duffield, J.: Friends Only: Examining a Privacy-Enhancing Behavior in Facebook. CHI, pp. 1553-1562 (2010)
9. Barnes, S.B.: A privacy paradox: Social networking in the United States. First Monday 11 (2006). http://firstmonday.org/ojs/index.php/fm/article/view/1394/1312#b1.
10. Zimmer, M.: The Externalities of Search 2.0: the Emergin Privacy Threats when the Drive for the Perfect Search Engine meets Web 2.0. First Monday 13 (2008). http://firstmonday.org/ojs/index.php/fm/article/view/2136/1944
11. Gustavsson , JP., Bergman H., Edman, G., Ekselius, L., von Knorring, L., Linder, J.: Swedish universities Scales of Personality (SSP): construction, internal consistency and normative data. Acta Psychiatrica Scandinavica 102, 217-225 (2000)
12. Felder, R.M., Spurlin, J.: Applications, Reliability and Validity of the Index of Learning Styles. International Journal of Engineering Education 21, 103-112 (2005)
13. Schraw, G., Dennison, R.S.: Assessing metacognitive awareness. Contemporary Educational Psychology 19, 460-475 (1994)

Mining Potential Domain Expertise in Pinterest

Ana-Maria Popescu[1], Krishna Y. Kamath[2] and James Caverlee[2]

[1] Research Consulting
anamariapopescug@gmail.com
[2] Texas A&M University, College Station TX 77840, USA,
krishna.kamath@gmail.com, caverlee@cse.tamu.edu

Abstract. This paper describes a first investigation of *potential domain expertise* in Pinterest. We introduce measures for characterizing the volume and coherence of Pinterest users' pinning activity in a given category, their perceived and declared category-specific expertise and the response from the social network. We use such signals in the context of a supervised ML framework and report encouraging preliminary results on the task of mining *potential experts* for 4 popular content categories.

Keywords: experts, social network, interest network

1 Introduction

Pinterest is an image-based social platform which has seen rapid growth [16] in 2012. The site allows users to curate image collections (*boards*) as well as interact with other users and their content. Pinterest employs a set of >30 content categories to help in curation, search and discovery; when a board is created, the user can select a category label (e.g., "Home Decor"). The site also showcases category-specific time-sensitive feeds which expose users to the newest content and encourage pinning. Given a category, some pinners have more relevant real-life experience or sustained, deep interest in it than others; their collections can be used for high-quality recommendations and search results.
This paper describes a preliminary investigation of *mining potential experts* for Pinterest categories. First, we describe a set of signals used to capture potential expertise. Second, we report on encouraging initial experiments for identifying highly knowledgeable (*potential expert*) users for 4 popular Pinterest categories. Finally, we outline our ongoing work on the topic.

2 Related Work

Pinterest is starting to attract the attention of the research community: recent studies have focused on generic site analysis [4], gender roles and behaviors [14] and initial content quality measures [9]. Our focus is on identifying top users for given Pinterest categories. Extensive previous work has been done on identifying *global* and *topic-sensitive* authorities in other social networks or QA communities, using a variety of approaches (link analysis, text-based methods, etc.) [2, 1, 5, 15,

12, 7, 10, 11, 3]. We leverage insights from previous research for mining potential experts in a new network with a blend of interest-driven and socially motivated activities.

3 User Features

Given a category c (e.g., Design), we characterize a user u's pinning activity, interest, declared and perceived expertise for c using the features in the top 5 rows of Table 1. To start, we rely on the user-supplied category labels to find category-specific boards. Each final feature reflects a signal of potential expertise (see Table 2). In the following, we describe these features in more detail.

Table 1. User-level feature space for category c and user u. $f_{domExpert}$ relies on two automatically acquired lexicons, LexGenExpert (3.1) and LexCat(c) (3.2).

Final features: f(u,c)	**Description**
$f_{domExpert}$	$f_{catRel} * (f_{genExpert} * w_{gen} + f_{selfProm} * w_{prom})$
f_{vol}	$f_{\%boardsCat} * (f_{numBoards} * w_{ct} + f_{boardSize} * w_s)$
f_{coh}	$f_{\%boardsCat} * (f_{semCoh} * w_d + f_{linkCoh} * w_l)$
$f_{socDirect}$	$f_{\%boardsCat} * (f_{repins} * w_r + f_{cumEFR} * w_{efr})$
f_{socNet}	1 if u is "authority" in repin graph for c; 0 otherwise
Basic features: f(u, c)	**Description**
f_{catRel}	$\alpha_0 * \frac{\lvert T(u) \rvert \cap LexCat(c) \rvert}{\lvert T(u)} + \alpha_1 * \frac{\lvert T(u) \rvert \cap LexCat(c) \rvert}{\lvert LexCat(c) \rvert}$ $T(u)$ = tokens in u's profile description
$f_{genExpert}$	$\alpha_0 * \frac{\lvert T(u) \rvert \cap LexGenExpert \rvert}{T(u) \rvert} + \alpha_1 * \frac{\lvert T(u) \cap LexGenExpert \rvert}{\lvert LexGenExpert \rvert}$
$f_{selfProm}$	$f_{url} * w_{url} + f_{accts} * w_{ac} + f_{desc} * w_d + f_{urlProm} * w_{up}$
$f_{url}, f_{accts}, f_{desc}$	binary features indicating the presence/absence of a url, Twitter or FB account, populated description field
$f_{urlProm}$	binary feat.: 1 if user pins from URL in profile; 0 otherwise
$f_{\%boardsCat}$	$\lvert B_{u,c} \rvert / \lvert B_u \rvert$, B_u = set of u's boards; $B_{u,c}$ =u's boards in cat. c
$f_{numBoards}$	$1 - e^{(-(\alpha * \lvert B_{u,c} \rvert^2))}$ ($B_{u,c}$ = set of u's boards in cat. c)
$f_{boardSize}$	$1 - e^{(-(\beta * meanBoardSize^2))}$, $meanBoardSize$ = mean. num. pins for all $b \in B_{u,c}$
f_{semCoh}	mean semantic coherence for $B_{u,c}$ board set (based on [9])
$f_{linkCoh}$	$\frac{\sum_{b \in B_{u_c}} linkCoh(b)}{\lvert B_{u,c} \rvert}$, where $linkCoh(b) = 1 - \frac{\lvert uniqueOriginUrls(pins(b)) \rvert}{\lvert pins(b) \rvert}$
$f_{repins}; f_{cumEFR}$	$1 - e^{(-(\gamma * f^2_{catRepins}))}$; $1 - e^{(-(\delta * f^2_{cumCatEFR}))}$
$f_{catRepins}$	$\frac{\sum_{b \in B_c} repinsStat(b)}{\lvert B_c \rvert}$, $repinsStat$ = avg. num. of repins for b's pins
$f_{cumCatEFR}$	$f_{\%boardsCat} * \sum_{b \in B_c} efr(b)$, where $efr(b) = 1 - \frac{followers(b)}{followers(u)}$

Profile Expertise Clues: Users may claim expertise in a category c by using generic expertise terms (e.g., "expert", "maker", "author") in the context of c (e.g., for $c = Food$: "nutrition expert", "cookbook author", etc.). Some include links to their Facebook/Twitter/Instagram accounts, blogs, shops on Etsy and more. Users may also pin from their websites (linked in the profile) in order to

Table 2. Signals for potential expertise in *Design* category (Data: section 4)

Volume f_{vol}	Coherence f_{coh}	Social(direct) $f_{socDirect}$	Social(network) f_{socNet}	DomExpert $f_{domExpert}$	Potential Expert Model(M)
karyna	rodhunt	karyna	zsazsabellagio	111creative	itscloudcuckoo
dilekarisoy	vtloc1989	plentyofcolour	vickiah	itscloudcuckoo	satsukishibuya
1000pin	woodbridgebuild	2dstudio	tristan50	brittanysharp22	luxe
vtloc1989	tinycastlecrtv	designlovefest	stacier	vbroussard	plentyofcolour
beverbal	HighPointMarket	psimadethis	shellytgregory	nyclq	howaboutorange

better "self-promote" [14]. The $f_{domExpert}$ feature seeks to capture these factors and identify users whose profile suggests potential category-specific expertise, knowledge or experience (e.g., for the Design category, the top users are *111creative*, *itscloudcuckoo* or *brittanysharp22*, professional graphic designers with their own design studios).

Volume: Users with significant category pin or board volume are of interest, especially if the relative volume for the target category with respect to others is large. f_{vol} takes these factors into account: for Design, top users based on volume include *karyna* and *dilekarisoy*, active users who focus on design content.

Coherence: We hypothesize that users with significant knowledge of a given category c tend to have better organized, more coherent content than beginners or users with a passing interest in c. f_{coh} reflects the combined semantic and URL-based coherence of the boards in c. *Semantic board coherence* uses the topic diversity measure in [9] while *URL-based coherence* checks if category boards feature content from a focused set of urls. Users with coherent category-specific content are more likely to be professional (as can be seen from the profiles of example users in Table 2).

Direct Social Feedback: The direct response to a user's boards in the form of *repins*, *likes*, *comments* or board-specific *followers* is a good indicator of audience interest and can help find high quality users (see Table 2). We found that *repins* are the most common form of social feedback and correlate with *likes* (comments are sparse). We leverage repins together with board followers who are not followers of the target user (this suggests a strong interest in the particular board).

Global Social Authority: We also make use of global authority measures for a user and a category-specific repin graph by checking if u is among the top $k = 250$ authorities/hubs (we used the NetworkX package [8]).

3.1 LexGenExpert: Category-independent Expertise Terms

The $f_{domExpert}$ feature checks if the generic expertise terms in a mined lexicon (LexGenExpert - see Table 3) are used in profiles in the context of the target category (e.g. "nutrition expert", "founder of a design firm"). Terms are mined as described in the following. For a sample $k = 10$ of Pinterest categories, users are ranked with respect to how consistently their category-specific content is *repinned* (we use $f_{catRepins}$ in Table 1). The top 150 users per category are

retained and their profile descriptions are tokenized. Resulting tokens t are scored using $s(t,c)) = \frac{freq(c,t)}{totalFreq(t)}$, a measure of term frequency in the top 150 user descriptions for c versus total frequency in description corpus for all Pinterest users in our dataset. Terms directly reflecting category names (e.g., designer) are automatically removed and added to a category-specific lexicon (3.2). For each category c, the top $n = 50$ terms according to $s(t,c)$ are retained. A final score $exp(t)$ is computed for all $t : exp(t) = \sum_{0<i<=k} topN(c_k,t)$, where $topN(c_k,t)$ is 1 if t is in the top n terms for category c_k and 0 otherwise. The top $m = 50$ terms based on $exp(t)$ are retained as *potentially* indicating category-agnostic expertise.

Table 3. Examples of automatically mined *generic potential expertise* terms.

lover	founder	blogspot	blogger	owner	graphic	vintage
write	market	enthusiast	addict	content	entertain	southern
author	writer	official	lifestyle	director	shop	brand

3.2 LexCat(c): Category-specific Interest Terms

Given a category c, we look for profile terms indicating interest in or expertise for c. We start with the terms ranked based on $s(t,c)$ in 3.1 above and remove LexGenExpert entries and terms with $freq(c,t) = 1$. We also leverage *twellow.com*, a public directory where users "list" themselves under category labels and which was helpful in other user modeling research [13]. For the $k = 10$ Pinterest categories, we obtain the 200 top users (w.r. to follower count) for related Twellow category labels and tokenize their profile descriptions. Terms are ranked using $s(t,c)$ limited to the available Twitter user profiles; the top 50 terms are retained for each c (e.g., DIY& Crafts: "beading", "crochet"; Food & Drink: "vegetarian", "produce").

4 Experiments: Identifying Potential Experts

In the following, we describe preliminary results for identifying *potential experts* in Pinterest categories.

Dataset: Our dataset contained 12,543 Pinterest users whose boards and pins were crawled in Dec 2012. The median number of boards per user is 16 and the median user follower count is 82. The >12,000 users are a fully-crawled sample of a larger set of Pinterest accounts mined from the feeds "pinterest.com/popular", "pinterest.com/everything" and pinterest.com/source" (in combination with example domains - e.g., "tumblr.com").[3]

Potential Experts: Given a subsample of 400 users and 4 *popular* Pinterest categories (Food/Drink, DIY/Crafts, Home Decor and Design), users were annotated with respect to their experience in and knowledge of each category after inspecting their pins and boards, social media accounts, website, and all other publicly available information. Users with *relevant experience* (e.g., chefs for Food/Drink, interior designers for Home Decor) were considered potential

[3] We also used BFS-style crawling to augment the user set. As a note, April 2013 experiments found Pinterest has become difficult to crawl due to site changes.

experts. Creators of *category-relevant content* recognized publicly (e.g, by means of awards, etc.) were also labeled as potential experts. In addition to this *strict* expertise definition, we used a more *relaxed* criterion: users with experience or recognized contributions in *closely related* categories were also labeled potential experts (e.g., for DIY/Crafts, a graphic designer whose DIY/Crafts boards cover invitation design, etc.). Remaining users were not considered potential experts[4]. Table 4 summarizes the labeling results for the *relaxed* potential expertise annotation - a larger-scale study and more in-depth guidelines are needed before providing general % numbers. As a note, ongoing work shows that other domains (e.g., Travel) have a drastically lower percentage of potential experts.

Table 4. Potential category expertise (*relaxed* version). Dataset: 400 users

Category	**Potential expert** (example)	**Not expert** (example)	**Potential expert** (%)	**Not expert** (%)
DIY & Crafts	vintagerevivals	boulderlocavore	**20.5%**	79.5%
Home Decor	dbohemia	elle_tea	**21.5%**	78.5%
Food & Drink	30aeats	nellieo	**9%**	91%
Design	980ds	1059alexandra	**22.75%**	77.25 %

Table 5. Results: Category-specific models (balanced GS). Notation: $M \setminus F$ = model using all features but F, B_F = baseline using only F

Cat.	Method	Avg. P	Avg. R	Avg. F1	Cat.	Method	Avg. P	Avg. R	Avg F_1
Food	$B_{domExpert}$	96.7	70.2	79.02	Design	M	89.5	80.3	83.95
Food	M	87.7	68.8	74.9	Design	$M \setminus f_{vol}$	89.3	76.8	81.6
Food	$M \setminus f_{vol}$	89.2	69	74.5	Design	$M \setminus f_{socNet}$	88	75.2	80.5
Food	$B_{socDirect}$	97.5	65.5	74.1	Design	$M \setminus f_{domExpert}$	86.5	72.5	77.5
Home Decor	$B_{domExpert}$	90.6	64.3	74.4	DIY/Crafts	M	71.7	70	68.9
Home Decor	M	81	67.7	72.1	DIY/Crafts	$M \setminus f_{vol}$	74.8	62.4	66.1
Home Decor	$M \setminus f_{coh}$	72.4	71.6	71.2	DIY/Crafts	$M \setminus f_{socNet}$	76.7	55.9	63.7
Home Decor	$M \setminus f_{vol}$	82.2	62.9	69.4	DIY/Crafts	$M \setminus f_{coh}$	75.9	57.7	62.5

4.1 Results

Category-specific models We used the manually labeled data to test if potential experts can be automatically identified. First, each category was targeted separately, with the relevant labeled data subset used for gold standard creation. We experimented with both a *balanced* gold standard set and an *unbalanced* set (entire labeled set). We used Generalized Additive Models (GAM) [6] as our ML framework in a 10-fold cross-validation setting. The full model (M) was compared to models using feature set subsets, including single-feature baselines. We focused on *precision*, *recall* and F_1 values for the *potentialExpert* class (averaged over 10 folds). Tables 5 and 6 show the best performing model versions ranked

[4] Cohen's kappa coefficient for 2 annotators was 0.59, in the "fair-to-good" range. We are devising more specific guidelines for ongoing work

by average F_1. We find that : a) potential experts can be identified with encouraging results; b) *potential expertise* profile clues and *direct social feedback* are particularly useful.

Generic Models We then recast the *potential expert* mining task as follows: given $e(u, c)$, where u is a user and c a category, can $e(u, c)$ can be automatically labeled as "potential expert" or not? A balanced gold standard (482 examples) was derived by combining the category-level balanced gold standards. The unbalanced gold standard corresponded to the entire labeled dataset (1600 examples). Table 7 summarizes the relevant results. The full model M and the model using all but the *volume* information perform best. Among the baselines, the *profile-based expertise clues*, the *coherence* and the *direct social feedback* ones perform best.

Table 6. Results: Category-specific models (unbalanced GS). Notation: $M \setminus F$ = model using all features but F, B_F = baseline using only F

Cat.	Method	Avg P	Avg R	Avg. F_1	Cat.	Method	Avg P	Avg R	Avg. F_1
Food	M	54.2	35.2	39.8	Design	$M \setminus f_{coh}$	64.3	48.04	54.3
Food	$M \setminus f_{socDirect}$	70	29.2	37.7	Design	$M \setminus f_{vol}$	73.1	45.08	53.5
Food	$M \setminus f_{socNet}$	77.7	34.8	37.3	Design	$M \setminus f_{socDirect}$	73.3	44.6	52.9
Food	$M \setminus f_{domExpert}$	66.7	29.3	36.2	Design	$M \setminus f_{socNet}$	75.2	40.8	50.4
Food	$B_{socDirect}$	96.7	22.5	29	Design	M	71.1	39.7	50.2
Home Decor	M	65.4	27.9	32.8	DIY/Crafts	$M \setminus f_{coh}$	60	34.7	38.4
Home Decor	$M \setminus f_{socNet}$	62.3	24.8	32.5	DIY/Crafts	$M \setminus f_{socNet}$	52.2	22.4	30.5
Home Decor	$M \setminus f_{vol}$	53.5	23.9	31.6	DIY/Crafts	M	55	22.3	30
Home Decor	$M \setminus f_{socDirect}$	62.1	24.8	31.5	DIY/Crafts	$M \setminus f_{vol}$	75	23.9	28.1
Home Decor	$M \setminus f_{coh}$	67.7	19	27.6	DIY/Crafts	$M \setminus f_{domExpert}$	54.2	18.8	25.7

Table 7. Results: Generic models. M, $M \setminus f_{vol}$, $M \setminus f_{coh}$ outperform top baselines on Avg. F1 (stat. significance: $*$:$p < 0.05$; $**$:$p < 0.1$).

Dataset	Method	Avg. P	Avg. R	Avg. F_1	Dataset	Method	Avg. P	Avg. R	Avg. F_1
Balanced	M	79.5	65.08	71.04^{*}	Unbalanced	$M \setminus f_{vol}$	70.1	23	34.1^{**}
Balanced	$M \setminus f_{coh}$	80.1	61.3	69^{*}	Unbalanced	M	61.6	23.4	33.7^{**}
Balanced	$B_{domExpert}$	82.7	48.3	60.2	Unbalanced	$B_{domExpert}$	59.5	16.4	25.3
Balanced	B_{coh}	67.6	48.4	55.6	Unbalanced	B_{coh}	54	10.5	17
Balanced	$B_{socDirect}$	84.6	38.4	52.5	Unbalanced	$B_{socDirect}$	72.3	9.3	15.9

4.2 Conclusions and Ongoing Work

This paper presented preliminary results showing that potential experts can be identified for specific Pinterest categories. Our ongoing work is focused on testing on larger gold standard sets and additional categories (both necessary for robust conclusions), improve our models and finally, integrate potential experts and their content in other Pinterest-related tasks.

References

1. Aral, S., Walker, D. Identifying influential and susceptible members of social networks . In: Science, Vol.337 no.6092 pp.337-341, 2012
2. Bakshy, E., Mason, W., Hofman, J., Watts, D.: Everyone's an Influencer: Quantifying Influence on Twitter. In: Proceedings of WSDM 2011
3. Cataldi, M., Mittal, N., Aufaure, M-A.: Estimating Domain-based User Influence in Social Networks. In: Proceedings of SAC, 2013
4. Gilbert, E., Bakhshi, S., Chang, S., Terveen, L.: I Need to Try This: A Statistical Overview of Pinterest. In: Proceedings of CHI-2013.
5. Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., Gummadi, K.: Crowdsourcing search for topic experts in microblogs. In: Proceedings of SIGIR 2012.
6. Hastie, T.J., Tibshirani, R.J.: Generalized Additive Models. In: Chapman & Hall/CRC, 1990
7. Jurczyk, P., Agichtein, E.: Discovering authorities in QA communities by using link analysis. In: Proceedings of CIKM 2007.
8. Hagberg A., Schult D., Swart P.: Exploring network structure, dynamics and function using NetworkX. In: Proceedings of SciPy2008, 2008.
9. Kamath, K., Popescu, A., Caverlee, J. : Board Coherence: Non-visual Aspects of a Visual Site. In: Proceedings of WWW 2013 (Companion Volume).
10. Liu, Lu., Tang, J., Han, J., Jiang, M., Yang, S.: Mining Topic-level influence in heterogeneous networks. In: Proceedings of CIKM 2010.
11. Luiten, M., Kosters, W., Takes, F. Topical influence on Twitter: A feature construction approach. 2012
12. Pal, A., Counts, S.: Identifying topical authorities in microblogs. In: Proceedings of WSDM 2011
13. Pennacchiotti, M., Popescu, A.: Democrats, Republicans and Starbucks Afficionados. In: Proceedings of KDD 2011.
14. Ottoni, R., Pesce, J.P., Las Casas, D., Franciscani, G., Kumaruguru, P., Almeida, V.: Ladies First: Analyzing Gender Roles and Behaviors in Pinterest. In: Proceedings of ICWSM 2013
15. Sharma, N.: Discovering topical experts in Twitter social network. In: Master's Thesis, IIT - Kharagpur, 2012.
16. Wikipedia: http://en.wikipedia.org/wiki/Pinterest

A Quantitative Approach for Modeling and Personalizing Player Experience in First-Person Shooter Games

Noor Shaker[1], Mohammad Shaker[2], Ismaeel Abuabdallah[2], Mehdi Zonjy[2], and Mhd Hasan Sarhan[2]

[1] IT University of Copenhagen, Rued Langaards Vej 7, 2300 Copenhagen, Denmark
[2] University of Damascus, Damascus, Syria
nosh@itu.dk, {mohammadshakergtr,pro.jaeger}@gmail.com,
mehdizonjy@hotmail.com,mhdhasansarhan@gmail.com

Abstract. In this paper, we describe a methodology for capturing player experience while interacting with a game and we present a data-driven approach for modeling this interaction. We believe the best way to adapt games to a specific player is to use quantitative models of player experience derived from the in-game interaction. Therefore, we rely on crowd-sourced data collected about game context, players behavior and players self-reports of different affective states. Based on this information, we construct estimators of player experience using neuroevolutionary preference learning. We present the experimental setup and the results obtained from a recent case study where accurate estimators were constructed based on information collected from players playing a first-person shooter game. The framework presented is part of a bigger picture where the generated models are utilized to tailor content generation to particular player's needs and playing characteristics.

Keywords: Player Experience Modeling, Affect Recognition, Procedural Content Generation, Adaptive Games

1 Introduction

Understanding players' interaction with a game has been the focus of many research studies. Several theoretical attempts have been proposed that aim at identifying patterns of player behaviors and building qualitative theories that relates aspects of game design to key concepts of gameplay experiences [9, 8, 2, 3]. While these theories constitute much of our understanding of the in-game interaction, they lack the necessary details to be implemented in computational models. Moreover, most of these theories are based on general high-level observations which makes them unsuitable for the personalization of content. Having an algorithm that, given information about the player style, can predict the appeal of the game content to this specific player is useful for many reasons: first, this would help us better understand the game-player relationship; second, such an algorithm could allow us to identify the aspects of the game content that

contribute to player entertainment and finally, this would allow us to achieve the ultimate aim of most of the studies in the field of affective computing, that is being able to adapt the game to the player and thus successfully closing the affective loop in games [14, 6, 10, 1].

An interesting direction that has received increasing attention is Procedural Content Generation (PCG) in which artificial and computational intelligence methods have been utilized to generate different aspects of content with or without human interference [15]. An interesting direction within the automatic content generation is the creation of personalized content [7, 5, 13]. The first step towards achieving such goal is to model the relationship between user experience and content. This can be done by the construction of data-driven models based on data collected from the interaction between the user and the digital content and annotating this data with user experience tags [17]. The *Experience-Driven Procedural Content Generation* (ED-PCG) framework [17] suggests the different components that should be implemented to realize this goal.

In this paper, we advocate the use of an ED-PCG approach to adapt games to players, and present an experiment conducted within this direction. We follow a similar protocol to the one followed in our previous attempts to capture and personalize player experience in a clone of the popular game *Super Mario Bros* [11]. We extend our previous attempts through investigating a whole new game genre, more specifically, we follow similar methodology to model player experience in a First-Person Shooter (FPS) game. This allows us to test the generality of the suggested modeling framework and check how well it scales when applied in a more complex environment and more sophisticated, richer, form of the in-game interaction.

Within this context, the presented work employs a fusion scheme of game-content parameters and game-performance indicators in order to predict player preferences between different game variants. Players' preferences are identified via comparative questionnaires and different game variants are ranked with respect to frustration, engagement and challenge. Automatic feature selection and neuroevolutionary preference learning are employed to select a subset of appropriate features that yield accurate predictors of the reported affects. Results show that accurate player experience models (accuracy higher that 71%) can be constructed.

2 The Testbed Game: Sauerbraten

We used a modified version of the FPS game called *Sauerbraten* as a testbed for our experiment (see Figure 1 for a screenshot of the game). The game is built on a game engine called *Cube*, and both game and game engine are public domain and freely available online[3]. The game can be played in a single player or multiplayer mode. For the experiments presented in this paper, we focus only on the single player mode to eliminate the other effects. The levels employed are

[3] http://cubeengine.com

Fig. 1. A screen shot from the FPS game used as a testbed.

composed of two-layered with Non-Player Characters (NPCs) spawned along the levels. Each game session lasts for two minutes and the goal of the game is for the player to get the highest score possible by killing as many of the enemies as possible.

The player can kill enemies by shooting at them using different types of weapons that differ in their accuracy, damage caused and shooting range. NPCs can also shot at the player causing health lose and eventually death. The amount of health lose depends on the type of the weapon used for shooting. Every time the player is killed, he/she looses one point and he/she is re-spawned again as long as he/she still has time left to play.

3 Player Experience Modeling Framework

The Player Experience Modeling (PEM) framework followed consists of two main steps: crowed-sourcing data from players, and constructing data-driven models of player experience. The ultimate aim of the framework is to construct models that approximate the relationship between features of game content and player behavior and reported affective states.

3.1 Data Collection

Game surveys were conducted to collect information about players' interaction with the games and their affective states. The protocol suggested in [18] was followed to design and solicit the information. According to the protocol, players are presented with a pair of two sessions that differ along one or more aspects of game content. While playing, detailed information about player behavior and actions were recorded. After playing each pair, players were asked to report their emotional/behavioral states following the four-alternative forced choice protocol that asks the players to express their preference of the three states: *engagement*, *frustration* and *challenge*. The selection of these states is based on earlier game survey studies [11, 4] and our intention to capture both affective and cognitive/behavioral components of gameplay experience. Moreover, we want to keep

the self-reporting as minimal as possible so that experience disruption is minimized. Pairwise preferences have been adopted for this study because of their numerous advantages over rating-based questionnaires [16]. The questionnaires presented are of the form: "Which game was more E?" where E is the state under investigation. The possible answers are: (1) game A [B] was more E than game B [A] (2) both equally or (3) neither.

A total number of 115 players participated in the data collection experiment and several features were extracted from the recorded data and used to build models of player experience. The participants were all first to fifth-year students at the Faculty of Information Technology Engineering at the University of Damascus.

3.2 Feature Extraction

Several features about the content of the game presented to the players as well as gameplay features capturing different aspects of player behavior and the in-game interaction were extracted from the game sessions recorded. The game engine was modified to allow recording the gameplay features while the game is being played. A complete log is also saved permitting the extraction of additional features after data collection. Table 1 presents a subset of the features extracted. The context features presented are the ones used to construct the variations of the game content presented to the players.

Table 1. Gameplay and expressivity features extracted from the data recorded.

Category	Feature	Description
		GamePlay Features
Time	t_{life}	Duration of play
	t_{weapon}	Time spent using weapons (%)
	t_{shoot}	Time spent shooting (%)
	t_{still}	Time spent not moving (%)
	t_{jump}	Time spent jumping (%)
Interaction with items	n_{health}	Health items collected (%)
	n_{armour}	Armours collected (%)
Interaction with enemies	e_{kill}	Number of times the player kills an enemy (%)
	p_{hit}	Number of times the player receives a hit from an enemy (%)
	e_{hit}	Number of times the player hits an enemy (%)
Miscellaneous	n_{death}	Number of times the player died
	s_{acc}	Shooting accuracy
		Context Features
	E	Number of enemies
	E_{skill}	Skill level of enemies
	W_{type}	Type of weapons including explosive and non-explosive weapons
	H	Number of health items
	R	Number of resources such as bullets and armors

4 Preference Learning for Modeling Playing Experience

The data collected in the previous step is used to construct accurate estimators of player experience. Models of player experience were built using neuroevolutionary preference learning [18]. The features extracted in the previous step are set as input to a feature selection method to chose a subset of relevant features for predicting each emotional state using forward feature selection method. The selected subset of features are then used to build the neural network models which are trained to adjust the weight so that their output matches the reported preferences. The topologies of the models were also optimized for best prediction accuracies.

Different subsets of features were selected to predict each emotional state pointing out to various roles each feature plays to elicit the different affective states. Some of the features, such as the number of enemies and the their skill, were selected as predictors of engagement and challenge suggesting an implicit relationship between these two states. Accurate estimators of player experience were constructed with average accuracies of 71.26%, 81.42% and 97.27% for engagement, frustration and challenge, respectively. Table 2 presents information about the features selected and the average prediction accuracies obtained over five runs. The results indicate that challenge is the easier to predict while engagement is the hardest with the largest subset of features and the lowest accuracy.

Table 2. Features selected from the set of extracted parameters for predicting engagement, frustration and challenge. The table also presents the corresponding average ($Performance$) values obtained. Context features also appear in bold.

	Engagement	Frustration	Challenge
$Selectedfeatures$	p_{hit}	p_{hit}	t_{life}
	t_{still}	e_{hit}	n_{death}
	$\mathbf{E_{skill}}$	e_{kill}	$\mathbf{E}$
	$\mathbf{E}$	t_{still}	$\mathbf{E_{skill}}$
	$\mathbf{W_{type}}$		$\mathbf{W_{type}}$
	t_{exp}		t_{weapon}
	n_{armour}		
$Performance$	71.26%	81.42%	97.27%

It is worth noticing that none of the content features were selected for predicting frustration which indicates the this emotional state is more directly influenced by player behavior unlike engagement and challenge where three out of the four context features were selected highlighting the impact of context information on these two states.

5 Personalizing Player Experience

The models derived can be used to personalize the game context tailoring the content generation to desired levels of engagement, frustration or challenge for an individual player based on his/her playing style. This can be achieved by first adjusting the model for control —by including the set of context parameters into the input of the models— and then searching the content space for game content that, taken together with player specific gameplay characteristics, can optimize a specific experience. This new player dependent content is then presented to the player closing the affective loop in games.

Depending on the size of the content space, exhaustive search or global stochastic search methods can be employed. This approach has been tested to personalize player experience in our previous work on a platform game [12] with encouraging results and we are in an ongoing effort to investigate the applicability of the method for the FPS game under investigation. The preliminary results show that the models are able to recognize different playing characteristics and generate personalized content accordingly. However, more experiments and evaluation are required if we are to draw robust conclusions.

6 Conclusions and future work

In this paper we presented a scheme for modeling player experience from behavior and context features. Players' reports of three emotional states (engagement, frustration and challenge) were collected along with features from game sessions. Feature extraction, selection and neuroevolutionary preference learning methods were employed to approximate the function between context and behavior features, and reported affective states of players. Different subsets of features were selected to predict each emotional state and accurate estimators were constructed.

A game personalization approach is also presented in which the constructed models can be used to evaluate the content and chose the best fit for each individual needs. The experiments and results presented in this paper are part of an ongoing project that aims at validating the extendibility of the player experience modeling framework by applying it on different game genres and for the purpose of closing the affective loop in games. More experiments are currently undertaken to generate and evaluate the models and the personalized content. Moreover, we are investigating the use of other, more expressive, modeling techniques that could potentially be used to help us better understand the in-game interaction and the effect of context on player behavior. Alternative personalization approaches could also be investigated.

The framework followed was previously tested for a platform game and the current paper show its applicability to FPS games. We believe that the same methodology can scale to other games from the same genre or other game genres and that the models constructed can be generalized to capture player experience in other games.

References

1. Charles, D., McNeill, M., McAlister, M., Black, M., Moore, A., Stringer, K., Kücklich, J., Kerr, A.: Player-centred game design: Player modelling and adaptive digital games. In: Proceedings of the Digital Games Research Conference. vol. 285. Citeseer (2005)
2. Chen, J.: Flow in games (and everything else). Communications of the ACM 50(4), 31–34 (2007)
3. Csikszentmihalyi, M.: Beyond Boredom and Anxiety: Experiencing Flow in Work and Play. Jossey-Bass, 25th anniversary edn. (Apr 2000)
4. Gilleade, K.M., Dix, A.: Using frustration in the design of adaptive videogames. In: Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology. pp. 228–232. ACM (2004)
5. Hastings, E.J., Guha, R.K., Stanley, K.O.: Evolving content in the galactic arms race video game. In: Proceedings of the 5th international conference on Computational Intelligence and Games. pp. 241–248. CIG'09, IEEE Press, Piscataway, NJ, USA (2009)
6. Höök, K.: Affective loop experiences - what are they? In: Lecture Notes in Computer Science. vol. 5033, pp. 1–12. Springer (2008)
7. Kazmi, S., Palmer, I.: Action recognition for support of adaptive gameplay: A case study of a first person shooter. International Journal of Computer Games Technology p. 1 (2010)
8. Koster, R.: A theory of fun for game design. Paraglyph press (2004)
9. Malone, T.: What makes computer games fun? ACM, New York, NY, USA (1981)
10. Pagulayan, R.J., Keeker, K., Wixon, D., Romero, R.L., Fuller, T.: User-centered design in games. The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications pp. 883–906 (2003)
11. Shaker, N., Yannakakis, G.N., Togelius, J.: Crowd-sourcing the aesthetics of platform games. IEEE Transactions on Computational Intelligence and Games, Special Issue on Computational Aesthetics in Games (2012)
12. Shaker, N., Yannakakis, G., Togelius, J., Nicolau, M., ONeill, M.: Evolving personalized content for super mario bros using grammatical evolution (2012)
13. Shaker, N., Togelius, J., Yannakakis, G.N., Weber, B., Shimizu, T., Hashiyama, T., Sorenson, N., Pasquier, P., Mawhorter, P., Takahashi, G., Smith, G., Baumgarten, R.: The 2010 Mario AI championship: Level generation track. IEEE Transactions on Computational Intelligence and Games 3, 332–347 (2011)
14. Sundström, P.: Exploring the affective loop. Ph.D. thesis, Stockholm University (2005)
15. Togelius, J., Preuss, M., Yannakakis, G.: Towards multiobjective procedural map generation. In: Proceedings of the 2010 Workshop on Procedural Content Generation in Games. p. 3. ACM (2010)
16. Yannakakis, G., Hallam, J.: Erratum: Ranking vs. preference: a comparative study of self-reporting. Affective Computing and Intelligent Interaction pp. 1–1 (2011)
17. Yannakakis, G.N., Togelius, J.: Experience-Driven Procedural Content Generation. IEEE Transactions on Affective Computing (2011)
18. Yannakakis, G.N., Maragoudakis, M., Hallam, J.: Preference learning for cognitive modeling: a case study on entertainment preferences. IEEE Transactions on Systems, Man, and Cybernetics. Part A 39, 1165–1175 (November 2009)

Understanding the Temporal Dynamics of Recommendations across different Rating Scales

Paula Cristina Vaz[1], Ricardo Ribeiro[1,2], and David Martins de Matos[1,3]

[1]INESC-ID, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
[2]Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisboa, Portugal
[3]Instituto Superior Técnico (IST), 1049-001 Lisboa, Portugal
{paula.vaz,ricardo.ribeiro,david.matos}@inesc-id.pt

Abstract. Libraries have large growing book collections and users have difficulty in browsing the whole collection when choosing new books to read, particularly when looking for books without a defined goal. In this case, recommendation systems are useful and play an important role in improving library usability. Recommendations are based on ratings and the quality of recommendations depends on the quality of the ratings. Studies show that users rate more items if scales have smaller granularity. In this paper, we propose a different rating scale for the book recommendation scenario in a collaborative filtering set-up and study how time influences rating relevance. Our findings suggest that the collaborative filtering algorithm benefits from a rating scale with smaller granularity. Moreover, if some conditions are met, rating prediction quality can be improved if we give lower weight to older ratings.

Keywords: Book recommendation, Collaborative Filtering, Temporal relevance, Rating scale

1 Introduction

Libraries both physical and digital have large growing book collections. Library users have difficulty in browsing the whole collection when choosing new books to read, particularly when looking for books without a defined goal. In this case, recommendation systems come in hand and play an important role in improving library usability.

Recommendation systems (RS) try to *know* the users observing their rating history. RS learn how the users rate their books and searches for other users with the similar tastes to generate reading recommendations. Two main techniques are used to develop recommendation systems [1]: content-based (CB) techniques in which users will be recommended items similar to those the user liked in the past; and collaborative filtering (CF) in which users will be recommended items that were preferred together. Each technique has limitations when taken individually, such as limited content analysis, the new item problem, sparsity, among others. To address these limitations, hybrid recommender systems have

been proposed where CB and CF techniques are combined in order to overcome the limitations of each technique.

To make suggestions, RS heavily depend on ratings, because only ratings tell the system what was the user opinion about an item. Ratings can be obtained implicitly or explicitly. Implicit ratings do not need any kind of user feedback. On the other hand, explicit ratings require explicit user feedback. Typically, these ratings are expressed on a 1-5- or 1-10-scale. This rating system has the advantage of better express users feelings about the book, but has the disadvantage of depending on user's explicit feedback. To further complicate maters, user preferences and opinions change over time. A user will most probably change the rating given to a book if that user is asked to rate it again.

This paper aims to (a) compare 1-5-scale rating to a like/neutral/dislike-scale in the rating prediction task; and (b) study the influence of rating age in predictions in the book recommendation scenario.

This paper is structured as follows: Section 2 gives an overview of the collaborative filtering approach. In section 3 we describe the data-set on which we based our experiments and the evaluation protocol. Section 4 describes and discusses the experiments. Section 5 describes related work. Finally, section 6 draws the conclusions and points to future directions.

2 Collaborative filtering

Following the work of [6], where the author proposes an user-based evolutionary kNN CF algorithm, in which ratings are weighted according to their age, we adapted the item-based CF algorithm in [7] by incorporating temporal information. Our temporal item-based CF algorithm (TICF) implements temporal decay through the use of the function in equation 1.

$$f_{u,i}^{\alpha}(t) = e^{-\alpha(t-t_{u,i})} \tag{1}$$

where u and i are the user and item relative to the rating, t is a time-stamp, and α controls the decaying rate. When α is set to 0, the time influence is ignored. $f_{u,i}^{\alpha}(t)$ measures the relevance of each observed rating $r_{u,i}$ in recommendation making, at time t based on the parameter α.

Temporal relevance has two dimensions: the age of the ratings given by the active user u, i.e., the user for whom recommendations are being made and the ratings given by the community, i.e., other users in the data-set. The age of the user ratings is controlled by parameter α. This parameter affects the rating prediction in equation 2, where $s_{i,j}$ is the adapted Pearson similarity (equation 3) between item i and item j and $r_{u,j}$ is the rating given by the active user u to item j.

$$P_{u,i} = \frac{\sum_j^k \; s_{i,j} * f_{u,j}^{\alpha}(t) * r_{u,j}}{\sum_j^k \; s_{i,j} * f_{u,j}^{\alpha}(t)} \tag{2}$$

The age of community ratings is controlled by parameter β that affects the item similarity calculation, as shown in equation 3, where $\overline{r^{\beta}}_i$ is the avearge

rating given by users to item i and each rating is affected by the time weight. If α and β are 0, the algorithm works as the usual item-based CF.

$$s(i,j) = \frac{\sum_{u=1}^{n}(f_{u,i}^{\beta}(t) * r_{u,i} - \overline{r^{\beta}}_i)(f_{u,j}^{\beta}(t) * r_{u,j} - \overline{r^{\beta}}_j)}{\sqrt{\sum_{u=1}^{n}(f_{u,i}^{\beta}(t) * r_{u,i} - \overline{r^{\beta}}_i)^2}\sqrt{\sum_{u=1}^{n}(f_{u,j}^{\beta}(t) * r_{u,j} - \overline{r^{\beta}}_j)^2}} \quad (3)$$

3 Evaluation protocol

For our experiments, we used the LitRec [9] data-set, from which we selected the 943 users with more than 10 ratings. This user selection left the data-set with 1,679 books and 34,156 ratings. LitRec was collected over a period of 5 years from *GoodReads.com* and was divided in a 90%-10% train-test-set. For each user in the test set, we selected the 10% most recent rated books. Then, we predicted a rating for each pair $<user, book>$ in the test-set and calculated the mean absolute error (MAE) as shown in equation 4, where p_i is the predicted rating, o_i is the observed rating for book i and N is the number of rating-prediction pairs.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - o_i| \quad (4)$$

4 Experimental set-up

We ran two experiments. First, we ran the TICF algorithm without using time relevance ($\alpha = \beta = 0$) to assess if the algorithm could benefit from a rating scale with smaller granularity. Then, we run the algorithm TICF with different values of α and β to study time influence in rating prediction quality.

4.1 Like/neutral/dislike rating scale

For this experiment, we converted the 1-5-scale of ratings in a 3-value-scale by replacing ratings 1-2 with a "dislike", rating 3 with a "neutral", and ratings 4-5 with a "like". This scale division was based on the reading of a significant number of reviews in the GoodReads.com site, that allowed us to get a sense of how users apply the rating scale in this particular data-set. We wanted to assess if error in rating predictions decreases with a smaller scale. The intuition behind the use of this type of scale is that it is easier for users to remember if they liked or hated a book, then to remember if they liked it with an intensity of 4 or 5. Moreover, according to the work presented by [8], users give more feedback to the system if the granularity of the rating scale is smaller. Then, we run the TICF algorithm on the data-set with $\alpha = \beta = 0$, varying the neighborhood size.

Figure 1 shows the error evolution. As can be observed, results are better for the 3-value scale. For the 1-5-scale, the MAE decreases until the 11^{th} neighbor and then becomes stabilized. For the 3-value-scale, the MAE decreases until the 4^{th} neighbor, then, increases until the 10^{th} neighbor, becoming steady after that.

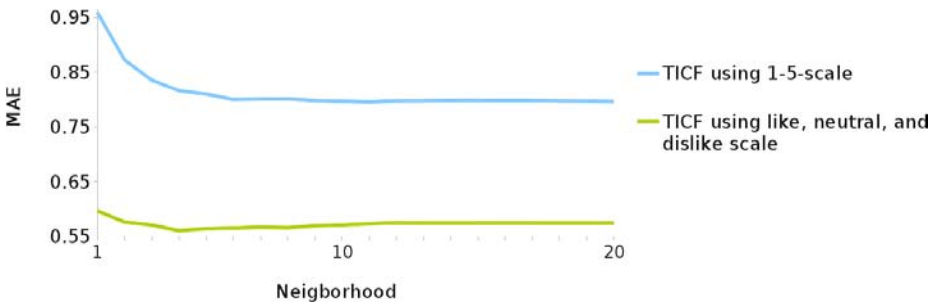


Fig. 1. MAE when using 1-5-scale ratings and a like, neutral, and dislike scale.

4.2 Temporal dynamics

In order to study the influence of rating age in prediction quality we run the TICF rating algorithm for different combinations of α and β with α and $\beta \in \{0, 0.1, ..., 1\}$. Rating age was measured in years and semesters and we ran the experiment for both rating scales. Figures 2 and 3 show the evolution of the MAE according to α and β variations. As can be observed, overall results are better when the rating age is considered in years (figure 2). Moreover, as expected, the MAE is lower for the 3-value rating scale (figures 2 (b) and 3 (b)).

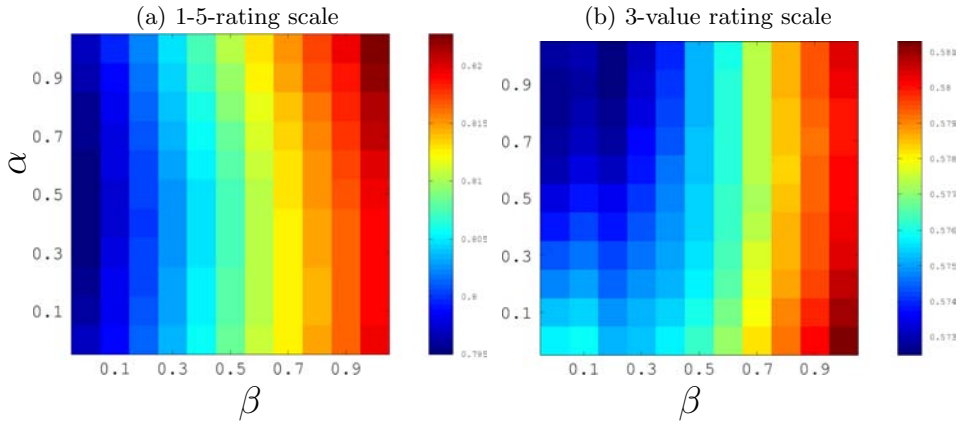


Fig. 2. MAE evolution for α and β variations, considering rating age in years.

Figures 2 and 3 show prediction quality changes with α and β variations. When the 1-5-scale is used, the MAE increases with the increment in the value of β, but when $\beta = 0$, the MAE decreases for $\alpha \in [0.3..0.8]$ (figure 2 (a)) and for $\alpha \in [0.1..0.3]$ (figure 3 (a)). These results show that rating prediction quality is affected both by recent and older ratings, regarding the community rating age. Regarding the active user rating age, rating prediction quality can be improved

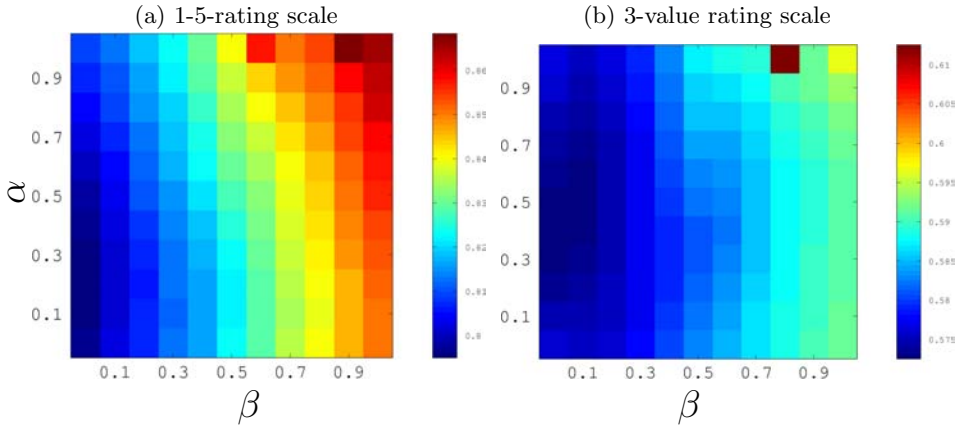


Fig. 3. MAE evolution for α and β variations, considering rating age in semesters.

if only the last two years of ratings are considered ($\alpha = 0.8$ for years and $\alpha = 0.6$ for semesters). The active user preferences are closer to his recent rated books than to older rated books.

When the 3-value rating scale is used, the MAE has the lowest values for $\alpha \in [0.8..1]$ and $\beta = 0.2$ (figure 2 (b)) and for $\alpha \in [0.3..0.6]$ and $\beta \in [0..0.1]$ (figure 3 (b)). results show that rating prediction quality can be improved if we consider ratings from the most recent four years ($\beta = 0.2$ for years and $\beta = 0.1$ for semesters). Regarding the active user rating age, rating prediction quality can be improved if only the present year of ratings is considered ($\alpha = 0.8$ for years and $\alpha = 0.6$ for semesters). The active user preferences are closer to the most recent rated books than to older ones. This scale is more sensitive to time relevance changes.

Results are consistent for both years and semesters and for both rating scales. Recall that the α parameter weights the active user ratings and that the β parameter weights the ratings of other users.

5 Related Work

Several approaches have been proposed to incorporate time relevance in the recommendation process. In [2] the authors adapt the item-based approach by incorporating time-based weights in the score prediction stage, but did not adapt similarity computation. [5] varies neighborhood size considering temporal information. [4] use matrix factorization to model changes in user and items over time. [6] adapted the a user-based CF algorithm by incorporating weights that give more relevance to recent ratings. Their approach affects the active user ratings and the community ratings. Nevertheless, the authors did not experiment with an item-based CF algorithm.

Rating scales used by recommendation systems have also been studied. In [3], the authors study the effect of different rating scales in user ratings, in a

cooking recipes recommendation system. The study shows that different users use rating scales differently and that wider scales are prone to more variability. In [8], the authors study the how often users rate across scales and conclude that as rating scale grows in granularity, users rate fewer items.

6 Conclusions & Future Work

Recalling the goals proposed at the beginning of the paper, we explored the TICF algorithm performance in predicting user tastes with a different rating scale. We converted the 1-5-scale in a like/neutral/dislike scale and run the TICF algorithm. Results shown that the TICF improves rating prediction quality when scale granularity decreases. From our study of the temporal relevance of rating age in the TICF algorithm, we were able to concluded that the active user preferences are closer to more recent ratings than older ones, especially considering a rating scale with lower granularity.

For future work, we want to confirm these results using other available datasets. We also want to explore if by giving less relevance to older rated books when using content-based recommendation, results confirm the ones obtained using a neighborhood-based CF algorithm.

Acknowledgments This work was supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2011.

References

1. G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.
2. Y. Ding and X. Li. Time weight collaborative filtering. In *Proceedings of the 14th ACM CIKM '05*, pages 485–492, New York, NY, USA, 2005. ACM.
3. C. Gena, R. Brogi, F. Cena, and F. Vernero. The impact of rating scales on user's rating behavior. In *Proceedings of the 19th UMAP'11*. Springer-Verlag, 2011.
4. Y. Koren and R. M. Bell. Advances in collaborative filtering. In *Recommender Systems Handbook*, pages 145–186. Springer, 2011.
5. N. Lathia, S. Hailes, and L. Capra. Temporal collaborative filtering with adaptive neighbourhoods. In *Proceedings of the 32nd ACM SIGIR '09*. ACM, 2009.
6. N. N. Liu, M. Zhao, E. Xiang, and Q. Yang. Online evolutionary collaborative filtering. In *Proceedings of the 4th ACM RecSys '10*, pages 95–102. ACM, 2010.
7. B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th WWW'01*. ACM, 2001.
8. E. I. Sparling and S. Sen. Rating: how difficult is it? In *Proceedings of the 5th ACM RecSys '11*, pages 149–156. ACM, 2011.
9. P. C. Vaz, D. Martins de Matos, B. Martins, and P. Calado. Improving a hybrid literary book recommendation system through author ranking. In *Proceedings of the 12th ACM/IEEE-CS JCDL '12*, pages 387–388. ACM, 2012.

Term extraction for user profiling: evaluation by the user

Suzan Verberne[1], Maya Sappelli[1,2], Wessel Kraaij[1,2]

[1] Institute for Computing and Information Sciences, Radboud University Nijmegen
[2] TNO, Delft
s.verberne@cs.ru.nl

Abstract. We compared three term scoring methods in their ability to extract descriptive terms from a knowledge worker's document collection. We compared the methods in two different evaluation scenarios, both from the perspective of the user: a per-term evaluation, and a holistic (term cloud) evaluation. We found that users tend to prefer a term scoring method that gives a higher score to multi-word terms than to single-word terms. In addition, users are not always consistent in their judgements of term profiles, if they are presented in different forms (as list or as cloud).

1 Introduction

In our project we aim to develop smart tools that support knowledge workers in their daily life. One of our objectives is a tool for personalized information filtering. We focus on two information filtering tasks: e-mail organization (which messages are important, which messages are related to a specific project) and professional search. In order to help the user to select relevant and important information in the large body of incoming e-mails and online search results, we need to create a model of the user. In the current work, we focus on the content-based part of the user profile: the user-specific terminology.

We aim to develop a user term profile that serves two purposes: (1) it will be used by our filtering tool for estimating the relevance of incoming information, and (2) it should give the user insight in his or her profile: which terminology is important in which context, and which terminology is shared with co-workers? Thus, the user profile should not only be effective in a system context but also valued by the user.

In the current paper, we evaluate three term scoring methods for the purpose of user profiling. We compared the methods in two different evaluation scenarios, both from the perspective of the user: a per-term evaluation, and a holistic (term cloud) evaluation. In Section 2 we describe three methods for collecting the descriptive terms from a user's self-authored document collection, and the evaluation setup. In Section 3, we present the results from the three methods and compare the two evaluation scenarios. Section 4 describes our conclusions and plans for future work.

2 Methodology

The input for our term extraction technology is a document collection provided by the user. First, the document collection is preprocessed: Each document is converted to plain text and the documents are split in sentences. Then candidate terms are extracted: Given a document collection, we consider as candidate terms all occurring n-grams (sequences of n words) that contain no stop words and no numbers. We used $n = [1, 2, 3]$ for the candidate terms. All candidate terms are saved with their term counts.

2.1 Term scoring methods

Until now, term scoring methods have mainly been evaluated in the context of Information Retrieval [8] and text classification [5, 10]. In Information Retrieval, term weighting is used to estimate the relevance of a document to a query. Therefore, term weighting in IR is generally based on TF-IDF (term frequency–inverse document frequency): a term is weighted by its frequency in the document, and by the number of documents in the corpus in which it occurs. Terms that occur in more documents are less informative for the documents in which they occur. In text classification, term weighting is used to select the terms that are the most informative for a specific category. Chi-square for example measures the lack of independence between a term and a category and Information Gain measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document [11].

The goal of term scoring for user profiling is to find the terms that are the most descriptive for a user's corpus. A way to select informative terms that are distinctive for the user's corpus compared to general English, we use a background corpus. We chose the Corpus of Contemporary American English as background corpus, which is free to use and is easy to process because the developers provide a word frequency list and n-gram frequency lists. We implemented three different term scoring functions from the literature:

1. Parsimonuous language model based (PLM): A method based on [2] where term frequency in the personal collection is weighted with the frequency of the term in the background corpus.

$$e_t = tf(t, D) * \frac{\lambda P(t|D)}{(1-\lambda)P(t|C) + \lambda P(t|D)} \tag{1}$$

$$P(t|D) = \frac{e_t}{\sum_t e_t} \tag{2}$$

Here, $P(t|D)$ is the probability of the term t in the personal document collection, $P(t|C)$ is the probability of the term in the background corpus and λ is a parameter that determines the strength of the contrast between foreground and background probabilities.

2. Cooccurence based (CB): A method based on [6] where term relevance is determined by the distribution of co-occurences of the term with frequent terms in the collection. The rationale is of this method is that no background corpus is needed because the most frequent terms from the foreground collection serve as background corpus.

$$\chi^2(t) = \sum_{g \in G} \frac{freq(t,g) - n_t p_g)^2}{n_w p_g} \tag{3}$$

$$\chi^{'2}(t) = \chi^2(t) - \max_{g \in G} \{ \frac{freq(t,g) - n_t p_g)^2}{n_t p_g} \} \tag{4}$$

   Here, G is the set of frequent terms (the size of which is determined by the parameter $topfreq$), $freq(t,g)$ is the co-occurrence frequency (in sentences) of t and g, n_t is the total number of co-occurrences of term t and G, and p_g is the expected probability of g.
3. Kullback-Leibler divergence for informativeness and phraseness (KLIP): A method based on [9] where the term relevance is based on the expected loss between two language models, measured with point-wise Kullback-Leibler divergence:

$$P(p||q) = \sum_x p(x) log \frac{p(x)}{q(x)} \tag{5}$$

   Tomokiyo and Hurst propose to mix two models for term scoring: phraseness (how tight are the words in the sequence) and informativeness (how informative is the term for the foreground corpus). The parameter γ is the weight of the informativeness score relative to the phraseness score.

The result of each of the term scoring methods is a list of terms for a document collection, with scores. We used the following parameter settings in our experiments: $\lambda = 0.5$ in the PLM method, $topfreq = 10$ in the CB method. In the KLIP method, we decided to give more weight to the phraseness component than to the informativeness component, because this is the only method that has a phraseness component. We set gamma to 0.1, which leads KLIP to generate more multi-word terms than the other methods.[3] We note here that the parameters should be optimized in future work.

2.2 Evaluation set-up

We asked five colleagues to provide us with a collection of at least 20 documents that are representative for their work. On average, we received 22 English-language documents per user (mainly scientific articles) with an average total of around 537.000 words per collecton. For each of these document collections, we generated three lists with 300 terms each using the PLM, CB and KLIP methods. Then we created a pool of terms per collection by first normalizing the

[3] In [9], informativeness and phraseness are weighted equally.

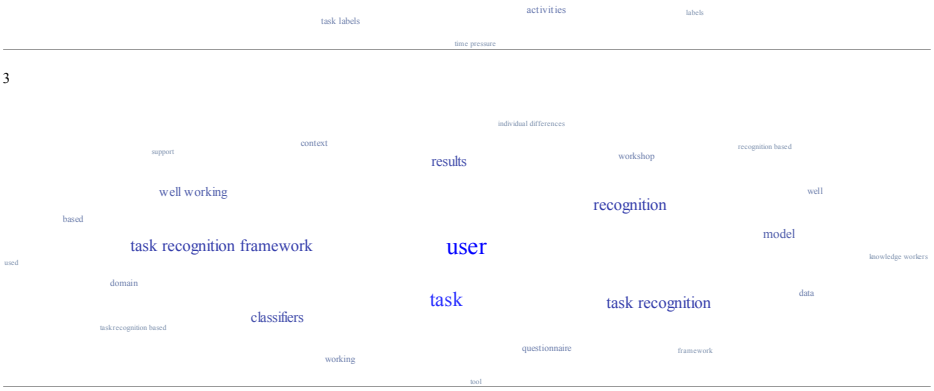


Fig. 1. Example of a tag cloud as it was shown to the user.

scores in each of the three lists relative to the maximum and minimum scores. We then calculated for each term the average of the three normalized scores. We ordered the terms by the combined scores and extracted the top-150. These terms were judged in alphabetical order by the owners of the document collections. We asked them to indicate which of the terms are relevant for their work. There was a large deviation in how many terms were judged as relevant by the users (between 24% and 51%), but on average, around one third of the generated terms (36%) was perceived as relevant.

In a second experiment, we evaluated the terms using term clouds. Instead of evaluating terms one by one, the profiles extracted from the documents were evaluated as a whole. Kaptein et al. [3, 4] and Gottron [1] show that using a term cloud as method to summarize a document can help the user in determining the topic of the document. For each user's document collection, we generated term clouds using the three term scoring methods. We chose a term cloud visualization where the biggest term is in the center of the cloud and the 25 subsequent terms are added in a spiral form, ending with the smallest terms in the outer ring. An example is shown in figure 1. We showed the term clouds in random order to the owners of the document collections, and asked them to rank the three clouds from the best to the worst representation of their work. They were allowed to give the same rank to two clouds, if they judged them equal in quality.

3 Results

We ordered the term lists by term scores from high to low and then used the term assessments to evaluate the ranked term lists for the three scoring methods. As evaluation measure we used Average Precision [12]:

$$\text{Average Precision} = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{n_c}, \tag{6}$$

where $P(k)$ is the precision at rank k, n is the total number of terms in the list, n_c is the total number of relevant terms and $rel(k)$ is a function that equals 1 if

the term at rank k is a relevant theme, and zero if it is not relevant. The results are presented in the upper half of Table 1. The lower half of the table shows the results for the ranking of the term clouds by the users.

Table 1. Results for the evaluation of the term lists and term clouds, per user A–E and overall. TF scores is a baseline ranking based on simple term frequency.

	A	B	C	D	E	Average	Stddev
% of pooled terms judged relevant	49%	30%	29%	51%	24%	36%	13%
	Average precision of ranked list						
TF scores	0.388	0.299	0.213	0.448	0.166	0.303	0.118
PLM scores ($\lambda = 0.5$)	0.407	0.312	0.221	0.461	0.177	0.316	0.120
CB scores ($toprank = 10$)	**0.424**	0.319	0.217	0.441	0.207	0.322	0.111
KLIP scores ($\gamma = 0.1$)	0.409	**0.438**	**0.409**	**0.599**	**0.293**	**0.430**	0.110
	Ranks of the term clouds. 1=best; 3=worst						
PLM scores ($\lambda = 0.5$)	2	2	**1**	2	2	1.8	0.4
CB scores ($toprank = 10$)	**1**	3	2	3	**1**	2	1.0
KLIP scores ($\gamma = 0.1$)	2	**1**	**1**	**1**	2	1.4	0.5

The table shows a large variation in the evaluation scores for the five knowledge workers. All three term extraction methods give better results than the plain TF scores. For all users except one, the KLIP method generates the best ranked list. As explained in Section 2, KLIP extracts more multi-word terms than the other two methods because it has a phraseness component. In fact, in the top-100 of most descriptive terms, KLIP has 64 multi-word terms on average, compared to 5 for Hiemstra and 4 for Matsuo. The finding that KLIP is judged the most positive suggests that users tend to find multi-word terms better descriptors of their work than single-word terms. The ranking of the term clouds for the three methods is significantly correlated to the ranking of the methods based on the Average Precision scores (Kendall $\tau = 0.67, P = 0.008$) but there are some differences. For example, to user E, the KLIP method generated the best ranking, but she judged the CB cloud as best visual reresentation of her work domain. This suggests that the visualisation of a term profile can play a role in how the user perceives the profile.

We also asked the users to label the terms that they judged as irrelevant with a reason why the term was not relevant. The results of this categorization are in Figure 2. The figure shows that there are no big differences between the types of irrelevant terms selected by the term scoring methods.

4 Conclusions and future work

We compared three term scoring methods in their ability to extract descriptive terms from a knowledge worker's document collection. On a small group of five users, we found that Kullback-Leibler divergence incorporating not only infor-

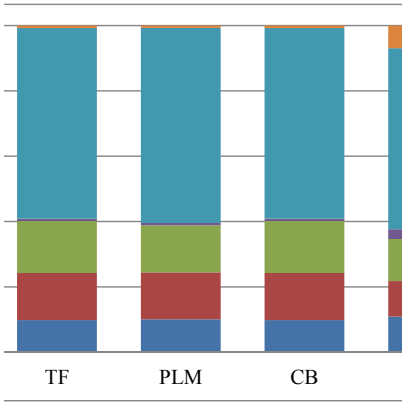


Fig. 2. Reasons that the users provided for irrelevant terms being irrelevant, per term scoring method. The counts have been summed over the users. 'Incomplete' denotes a partial term, e.g. care professional instead of health care professional; 'noise' are words in a different language, a PDF conversion error, parts of the document structure; 'not a term' are n-grams such as 'using' and 'million queries'.

mativeness but also phraseness of the terms gives the best results (Mean Average Precision is 0.43).

Since this work is still in an early stage, we can only draw preliminary conclusions. First, our results suggest that users tend to prefer a term scoring method that gives a higher score to multi-word terms than to single-word terms. It could be that multi-word terms are considered better descriptors because they are more specific than single-word terms. Second, users are not always consistent in their judgements of term profiles, if they are presented in different forms (as list or as cloud).

In the near future, we want to focus more on the best visualization of term profiles. For example, Rivadeneira et al. [7] found that tagclouds presented as an ordered list were easiest to comprehend. We will also study the possibilities of term clustering in order to visualize the multiple topics of projects that a knowledge worker is involved in, and investigate the differences between self-assessment of the profiles and judgments by colleagues. In addition, we will experiment with improving our term scoring methods by (1) optimization of the parameters λ (PLM), $topfreq$ (CB) and γ (KLIP), (2) finding an optimal combination of the three methods, (3) adding features to the terms such as position in the document, giving a higher preference to title words and (4) experimenting with a more specific background corpus. For example, in the Artificial Intelligence field, terms such as 'data' or 'user' are more frequent than in general English, but they might be considered too general to describe the work domain of one specific researcher in Artificial Intelligence.

References

1. Gottron, T.: Document word clouds: Visualising web documents as tag clouds to aid users in relevance decisions. In: Research and Advanced Technology for Digital Libraries, pp. 94–105. Springer (2009)
2. Hiemstra, D., Robertson, S., Zaragoza, H.: Parsimonious language models for information retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 178–185. ACM (2004)
3. Kaptein, R., Hiemstra, D., Kamps, J.: How different are language models andword clouds? In: Advances in Information Retrieval, pp. 556–568. Springer (2010)
4. Kaptein, R., Marx, M.: Focused retrieval and result aggregation with political data. Information retrieval 13(5), 412–433 (2010)
5. Lan, M., Tan, C.L., Su, J., Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(4), 721–735 (2009)
6. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13(01), 157–169 (2004)
7. Rivadeneira, A., Gruen, D.M., Muller, M.J., Millen, D.R.: Getting our head in the clouds: toward evaluation studies of tagclouds. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 995–998. ACM (2007)
8. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management 24(5), 513–523 (1988)
9. Tomokiyo, T., Hurst, M.: A language model approach to keyphrase extraction. In: Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18. pp. 33–40. Association for Computational Linguistics (2003)
10. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-. pp. 412–420. MORGAN KAUFMANN PUBLISHERS, INC. (1997)
11. Zheng, Z., Wu, X., Srihari, R.: Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter 6(1), 80–89 (2004)
12. Zhu, M.: Recall, precision and average precision. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo (2004), working paper

Semantic Technologies as Enabler for Distributed Adaptive Hyperlink Generation

Ruben Verborgh, Mathias Verhoeven, Erik Mannens, and Rik Van de Walle

Ghent University – iMinds – Multimedia Lab
Gaston Crommenlaan 8 bus 201, 9050 Ghent, Belgium
ruben.verborgh@ugent.be

Abstract. It is difficult for publishers to include the right links in documents, because they cannot predict all actions their users might want to perform. Existing adaptive navigation systems can generate relevant links, but doing this on a Web scale is non-trivial, especially if the targets are dynamic actions. As a result, adaptation often happens in a centralized way on a limited or closed document and action set. Distributed affordance is a technology to automatically generate links from any Web resource to matching actions from an open set of Web services, based on semantic annotations. In this paper, we indicate how this technology can be applied to adaptive navigation. We investigate how the generated links can be represented and how their relevance can be guaranteed. Based on that, we conclude that semantic technologies are an enabler to perform adaptive navigation to dynamic actions in a distributed way.

Keywords: adaptive navigation, adaptive hypermedia, Semantic Web

1 Introduction

The revolutionary concepts of hypertext have profoundly shaped the way we nowadays consume information, make decisions, and perform actions. Adding hyperlinks to documents transforms them into an affordance [7] through which users can select those actions [5]. However, this only helps users to the extent the actions they want to perform are afforded by the hyperlinks present in the document. On the Web, the world's largest hypertext system, publishers of documents are the ones who decide what hyperlinks their document contains, and thus what actions the user can perform through hypertext. Of course, it is impossible for a publisher to foresee *all* actions that *any* of its users would like to perform on a published resource. For instance, if the user wants to see the map of a certain place, but the publisher provides an address without the desired link, the user has to enter this address manually in a mapping application. We have previously called this *affordance coupling* [9]: pure hypertext-driven navigation on the Web would only work if the publisher could predict users' desired actions.

To solve the discrepancy between what publishers afford and what users need, we have previously developed an architecture for *distributed affordance* [9]. Based on semantic annotations extracted from documents, this approach enables

hypermedia clients to automatically create hyperlinks to actions that operate on the resources inside these documents. In contrast to most other approaches, we explicitly focus on *actions* that involve a document's resources *directly*, as opposed to finding contextually related documents. This paper explains how distributed affordance can serve as an adaptive hypermedia technique. In Section 3, we describe the architectural differences from other adaptive systems. Section 4 examines representation methods of the generated action links. In Section 5, we investigate how to determine what actions are relevant to the user.

2 Related Work

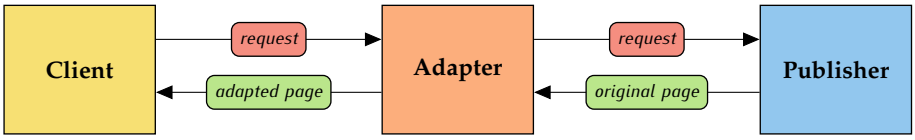
Brusilovsky gives a comprehensive overview of the aspects of adaptive navigation support [3], as well as various systems that existed at the time. He distinguishes five categories of adaptive navigation: *direct guidance*, *link ordering*, *link hiding*, *link annotation*, and *link generation*. The latter category consists of three kinds of approaches: *discovery of new links*, *similarity-based links*, and *dynamic recommendations*. The solution discussed in the present paper falls into the latter group, yet our generation strategy is *open-ended* on both sides of the link, whereas traditional adaptive navigation techniques mostly consider closed corpora. Furthermore, whereas adaptation techniques are traditionally characterized by a specific kind of knowledge representation [2], our technique decouples the information needed for adaptation from a specific representation format.

Dolog and Nejdl discuss the use of Semantic Web technologies for personalized link generation on the Web's open corpus [4]. The present paper differs in two aspects from the work they describe. First, they identify *ontologies* and *reasoning* as the corner stones of Semantic Web-based personalisation techniques. Our method is instead based on matching *Linked Data* [1] to *semantic service descriptions* [11]. Second, they focus on linking related pieces of *information*, whereas we are primarily interested in creating personalized *action* links. These links target dynamic information created from Web services, such as a link that connects any photo on the Web to its black-and-white version, which is then generated on-the-fly when the link is activated. In addition, we also target *world-changing* actions, such as sharing, ordering, purchasing, *etc.*

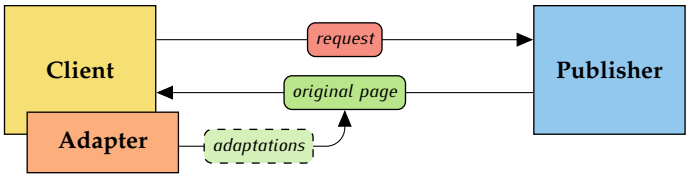
3 Architecture

Distributed affordance involves three parties that each supply a piece of information that allows links to be generated in a distributed way [9]:

- The **information publisher** adds ***semantic annotations*** to the document. Lightweight annotation mechanisms are sufficient, such as Open Graph or Schema.org, which are possibly already present for other purposes [10].
- The **action provider** offers semantically described ***Web services*** [11].
- The **user** indicates ***preferences*** for certain actions and providers, either implicitly or explicitly (see Section 5).



(a) centralized adaptive hyperlink generation



(b) distributed adaptive hyperlink generation

Fig. 1: Distributed adaptive navigation systems are highly scalable because adaptation happens at the client. This is enabled by semantic annotations in the original page that make it machine-interpretable.

These three pieces of information combined enable automated link generation in a distributed way, wherein the word "distributed" serves a double purpose. First, the affordance provided by the generated action links is distributed over multiple action providers, each of which can offer a specific action on the resource in the document. Second, adaptation does not need to happen at a centralized adapter, as is the case with most traditional adaptive systems [Fig. 1a]. Instead, because of the semantics in the document, adaptation can happen in a distributed way at the client [Fig. 1b], either through a browser extension or through a *shim* script that dynamically transcludes the generated links in the document [9].

The enabler of our distributed approach is machine-interpretable semantics, as it allows the on-the-fly combination of documents and services to create the actions the user needs. For instance, if the document indicates the page contains a postal address, the adapter will search for services that *a)* act on a postal address and *b)* have an outcome the user is interested in. Concretely, the user might be interested in viewing a map or adding this address to her personal address book. Semantic matching and subsequent instantiation of the address in the corresponding service descriptions [11] will result in direct links to both actions. Then, these actions have to be presented to the user, which is the topic of the next section.

4 Representation

Brusilovsky identified four categories of links [3]: *contextual links* that are embedded in parts of text or pictures, *local non-contextual links*[1] that reside on the page but are not intertwined with its content, *index links* on overview pages, and

[1] Here, "non-contextual" refers purely to link *placement* and not to relatedness, as generated links should at least be contextually related to the document's contents.

map links that represent a hyperspace or area thereof. Clearly, only the first two categories are relevant here, since the links we generate appear on content pages. This leaves us with two approaches: in-context action links near the resources on which they act, or action links in a separate menu.

In-context links When the document has been marked up with embedded semantic annotations, such as HTML5 microdata or RDFa, the actions generated based on those annotations can be placed close to them. Note, however, that linking from "hotspots" is often not desirable, as the links do not point to merely relevant documents, but to actions on the resources. For instance, it would be confusing if a link on an address directly inserted it into the user's address book. In contrast, a link labeled *"add to address book"* in the vicinity of the address indicates the intent more clearly. To suggest proper link placement, hypertext representations can indicate a placeholder where such links can be inserted [9]. However, this requires the publisher to be aware that adaptation might happen, which is why automated placement is more transparent.

Menu-based links Since we have no control over the page layout—as distributed affordance adaptation works on the full set of all Web pages—we might opt to insert in a separate menu instead. One option are contextual menus that appear when the cursor is hovered over resources that are part of the action. For instance, hovering over an address might reveal a pop-up menu with *"map"* and *"address book"* links. However, this approach will not work well for touch-based devices, which are increasingly gaining popularity. Therefore, we have experimented with a link sidebar that can be shown on demand. An alternate solution, not covered by Brusilovsky's categorization, is to show the action links in the browser window instead of the page itself, which is possible if distributed affordance is supported by the browser or through an extension. That way, the page renders as intended by the publisher, while still affording the user's preferred actions.

The benefit of in-context links is that they are close to information, and this proximity might allow the user to perform the action effortlessly. The drawback is that it can be hard to add them to existing pages in an aesthetically pleasing way—unless the publisher creates a designated affordance placeholder. The advantage of menu-based links is that they are non-obtrusive and offer more flexibility with regard to presentation and emphasis, at the cost of distance from the resources they act on.

5 Relevance

The other challenge in using distributed affordance for adaptive navigation is to find the actions that are relevant for a user. The difficulty lies in the fact that we support an open set of Web services, which, in combination with an open set of resources, result in an unlimited amount of possible actions. So far, the examples in this paper were rather simple, but we aim to support actions such as the following:

– Given a page with a book review, the user might want to buy the book through a preferred online bookstore, download it to her tablet, borrow it from a local library, or check if people in her social graph like it.
– When reading a page about a movie, the user can be interested to obtain tickets for a nearby movie theatre, to stream a digital copy to her portable media player, or to give it as a gift to someone else.

These examples indicate that complex matchmaking takes place. On the one hand, we need to determine the possible desired actions. Both examples show that there are actions tied to the *specific resource type* (books can be borrowed, movies can streamed) and actions tied to a *more general supertype* (books and movies can both be sent as a gift). On the other hand, the same action can be realized through different providers: there are several websites that allow to buy books and/or movies. With this in mind, we envision two possible strategies.

Explicit bookmarking Similar to the current practice of bookmarking, *i.e.,* saving the of a page in the browser for future use, actions can also be "bookmarked". We can imagine for instance, if a user visits an online bookstore, that she is offered to bookmark the *"buy"* action. Underneath the cover, this will add the corresponding Web service description to the user's collection. When the user then visits a document about a book, the description is then instantiated into a direct action link to buy that specific book. In that sense, the user is bookmarking *open-ended links*, the target of which becomes concrete at runtime.

Implicit modeling With bookmarking, the user is responsible for building an explicit model. However, it is far more convenient if the right actions can be suggested without an explicit selection process. Therefore, a user model can be constructed based on data mining [6]. Data sources of interest include previously visited sites of action providers (in combination with service discovery [11] on those sites) and the user's profile on social networking sites and interests from people in the user's social graph [8]. In case this data is missing or incomplete, the system might fall back on a "generic" user model that captures actions a typical user might perform on given resources.

Although implicit modeling is clearly more powerful, explicit bookmarking is more straightforward from an implementation perspective. In practice, both techniques can be combined: relying on an implicit model, but explicitly including bookmarks chosen by the user.

6 Conclusion

In this paper, we examined distributed affordance as an adaptive hypermedia technique. Our method differs from previous adaptive navigation systems, because our goal is to combine *a)* adaptation of the full Web corpus *b)* links to dynamic actions and *c)* fully distributed processing. Semantic technologies are a key enabler for the successful combination of these aspects, as they create the common understanding that eliminates the need for a centralized adaptation component

that must be aware of the full set of documents and actions. The important difference is where the knowledge is concentrated. In centralized systems, this knowledge resides mostly in the adaptation algorithm, whereas distributed affordance uses the knowledge provided by the semantics in the resource description and by the semantic service description.

Currently, we have implemented both the in-context and menu-based representation variants. In the near future, we will conduct user studies to see how both options are perceived and under what circumstances either one is most effective. As far as relevance is concerned, the implementation focuses on explicit bookmarking, but we plan to extend this to implicit user modeling as well.

We believe that distributed affordance can give a significant boost to serendipitous reuse of services, as it dynamically generates inbound links to them. Especially in mobile contexts, where invoking actions is more difficult because of limited controls, direct service action affordances could be a game changer. Demos and documentation are available online at `http://distributedaffordance.org/`.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The story so far. International Journal On Semantic Web and Information Systems 5(3), 1–22 (2009)
2. Brusilovsky, P.: Methods and techniques of adaptive hypermedia. User Modeling and User-Adapted Interaction 6(2–3), 87–129 (1996)
3. Brusilovsky, P.: Adaptive navigation support. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, pp. 263–290. Springer-Verlag (2007)
4. Dolog, P., Nejdl, W.: The adaptive Web. chap. Semantic Web technologies for the adaptive Web, pp. 697–719. Springer-Verlag, Berlin, Heidelberg (2007), `http://dl.acm.org/citation.cfm?id=1768197.1768224`
5. Fielding, R.T.: REST APIs must be hypertext-driven. Untangled – Musings of Roy T. Fielding (Oct 2008), `http://roy.gbiv.com/untangled/2008/rest-apis-must-be-hypertext-driven`
6. Frias-Martinez, E., Chen, S., Liu, X.: Survey of data mining approaches to user modeling for adaptive hypermedia. IEEE Transactions on Systems, Man, and Cybernetics 36(6), 734–749 (Nov 2006)
7. Gibson, J.J.: The theory of affordances. In: Shaw, R., Bransford, J. (eds.) Perceiving, Acting, and Knowing: Toward an Ecological Psychology. Lawrence Erlbaum (1977)
8. Torre, I.: Adaptive systems in the era of the semantic and social web, a survey. User Modeling and User-Adapted Interaction 19(5), 433–486 (2009)
9. Verborgh, R., Hausenblas, M., Steiner, T., Mannens, E., Van de Walle, R.: Distributed affordance: An open-world assumption for hypermedia. In: Proceedings of the Fourth International Workshop on RESTful Design (May 2013), `http://distributedaffordance.org/publications/ws-rest2013.pdf`
10. Verborgh, R., Mannens, E., Van de Walle, R.: The rise of the Web for Agents. In: Proceedings of the First International Conference on Building and Exploring Web Based Environments. pp. 69–74 (Jan 2013), `http://thinkmind.org/download.php?articleid=web_2013_3_30_40070`
11. Verborgh, R., Steiner, T., Van Deursen, D., De Roo, J., Van de Walle, R., Gabarró Vallés, J.: Capturing the functionality of Web services with functional descriptions. Multimedia Tools and Applications 64(2), 365–387 (May 2013), `http://link.springer.com/article/10.1007%2Fs11042-012-1004-5`

Unfolding cultural, educational and scientific long-tail content in the Web

Michael Granitzer[1], Christin Seifert[1], Silvia Russegger[2], and Klaus Tochtermann[3]

[1]University of Passau, Germany
[2]Joanneum Research, Graz, Austria
[3]German National Library for Economics, Kiel/Hamburg, Germany

Abstract. This project poster introduces the recently started EEXCESS project, which aims at Enhancing Europes eXchange in Cultural Educational and Scientific Resource. Europe has digitised vast amounts of cultural, scientific and educational content like for example scientific research, historical sound recordings, images of sculptures, films and sheet music. However, since content dissemination on the Web is driven by a small number of large central hubs like social networks or search engines, this cultural and scientific treasures has hardly been recognized by the general public or utilized in scientific and educational processes. EEXCESS aims to develop personalized and contextualized recommendation technologies to augment existing content dissemination channels (e.g. social media) and content creation process (e.g. blogging) for distributing high-quality educational, scientific and cultural content. In this project poster we present the underlying idea and related work with focus on user modeling and personalized, context-aware recommendation.

Keywords: educational and scientific resources, cultural content, personalized recommendation, long-tail content

1 Motivation

In the past decade, Europe conducted tremendous effort for making cultural, educational and scientific resources publicly available. Based on national aggregators like Collections Trusts Culture Grid, initiatives like Europeana nowadays provide a plethora of cultural resources for people worldwide. Concurrently, the semantic web, particularly Linked Open Data, has been growing exponentially providing semantic-enhanced access to and interchange of interesting scientific and cultural resources. Similarly, young start-ups like Mendeley re-shaped the management of scientific resources in a web-centric manner and several educational platforms started to emerge. Although such massive amounts of culturally rich, educating content are available, the potential of its use for educational and scientific purposes remains largely untapped. The primary reason can be seen in todays Web content distribution mechanisms: content dissemination is dominated by a small number of large central hubs like major search engines (e.g. Google), social networks (e.g. Facebook) or online encyclopedias (e.g.

Wikipedia). However, much valuable content is only available in the long-tail (i.e. a theory arguing that in internet-based markets niche content adds up to a huge body of knowledge, but is hidden from most users). In the long- tail content is maintained and curated by a large number of small to medium-sized professional organizations such as memory organizations (e.g. archives and museums), digital libraries and open educational repositories. However, the few large web hubs hardly support disseminating this long-tail content.

In order to reshape content dissemination mechanisms for highly specialized long-tail content EEXCESS relies on augmenting existing web channels with high-qualitative content through personalized, contextualized and privacy preserving recommendations, as discussed in the following.

2 Approach

In our approach, which is presented in an overview in fig. 1, we differentiate between two content related processes in the web: **content consumption** like reading web pages and surfing the web and **content creation** like authoring web pages or social media content. We aim to inject high-quality, long-tail content into those processes through personalized recommendation techniques by so-called augmentation interfaces. Those interfaces unobtrusively inject recommendation results into existing web pages or browsers. For example, when reading a Wikipedia page or a blog post on a certain topic, users should be given additional background material depending on their level of expertise and task. Similarly, during content creation processes, we aim to support the authors in creating the entry by recommending background material from digital libraries like the ZBW[1] or Europeana[2], digital cultural aggregators like Collection Trust[3] or scientific databases like Mendeley[4].

As in research on knowledge services or just-in-time retrieval [3], personalization along with unobtrusive interfaces will become a corner stone for high user acceptance. Hence, thoughtful user interface design jointly with highly related recommendation will be one major research challenge.

High quality recommendations require personalization and contextualization of the recommendation engines. While machine learning combined with semantic technologies [2] have been successfully applied to this task, privacy considerations remain a crucial task in terms of user acceptance. Hence, EEXCESS aims to retain full user privacy and user control through estimating context mostly on the client and submitting only minimal necessary information to the recommender system. The trade-off between privacy and recommendation accuracy will be the second major research challenge.

Our final research challenge lies in the recommender system itself. Since large-scale recommendation will become cost-intensive, it is unlikely that one

[1] http://www.zbw.eu/
[2] http://www.europeana.eu/
[3] http://www.collectionstrust.org.uk/
[4] https://www.mendeley.com/

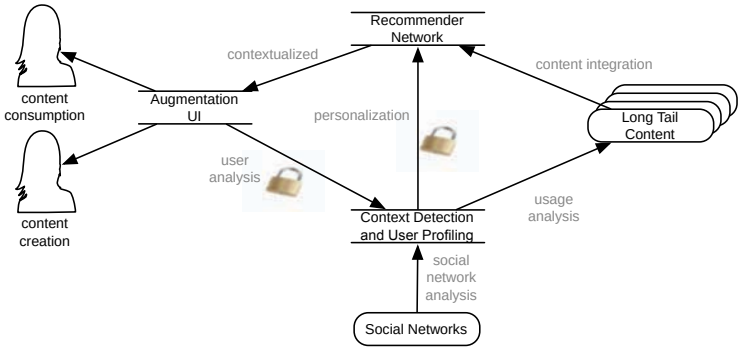


Fig. 1. Overview over EEXCESS envisioned content distribution process

institution will be able to run a single recommender. Therefore we aim to create a recommender network, in which every recommender is specialized on particular content and a subset of a user group. As discussed in the recommender community, like for example in [1], aggregating recommendation result over heterogeneous sources will become challenging in terms of accuracy and timeliness.

3 Impact

Although a large number of challenges have to be solved, very high impact can be expected by achieving the goals of EEXCESS. Particularly, content distribution process will become more open and less driven by big players or commercial interest. Moreover, through improved content creation we also expect to increase the scientific, educational and cultural quality of user generated content.[5]

References

1. Ivan Cantador, Peter Brusilovsky, and Tsvi Kuflik. Second Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec2011). In *Proceedings of the fifth ACM Conference on Recommender systems*, RecSys '11, pages 387–388, New York, NY, USA, 2011. ACM.
2. Michael Granitzer, Mark Kröll, Christin Seifert, Andreas S. Rath, Nicolas Weber, Olivia Dietzel, and Stefanie Lindstaedt. Analysis of machine learning techniques for context extraction. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pages 233–240. IEEE, 2008.
3. Andreas S. Rath, Nicolas Weber, Mark Kröll, Michael Granitzer, Olivia Dietzel, and Stefanie N Lindstaedt. Context-aware knowledge services. *Personal Information Management: PIM*, pages 5–6, 2008.

[5] **Acknowledgments.** The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number 600601.

LinkedUp - Linking Web Data for Adaptive Education

Eelco Herder[1], Stefan Dietze[1] and Mathieu d'Aquin[2]

[1] L3S Research Center, Leibniz University Hannover, Germany
{herder,dietze}@L3S.de
[2] Knowledge Media Institute, The Open University, United Kingdom
m.daquin@open.ac.uk

Abstract. Linked Data principles allow for easy discovery, reference, access and reuse of Web data. The user modeling community already widely exploits Semantic Web technologies, but the Linked Data approach is still not widely adopted. The LinkedUp project aims to advance the exploitation of open data on the Web, particularly for education. In this paper, we discuss the relevance of Linked Data for user modeling and personalization, and how to participate in and profit from the various initiatives of LinkedUp.

Keywords: LinkedUp Project, Linked Data, Learning Analytics, Personalization, Technology-Enhanced Learning

1 Introduction

Adaptive systems typically make use of techniques from the fields of information retrieval, machine learning, data mining and recommender systems to provide information and functionality that matches the user preferences, interests and requirements. The effectiveness of these techniques highly depends on the quality and quantity of available data about the resources and about the users.

Semantic Web technologies are often used in order to solve interoperability issues. However, even though the Linked Data approach [6] has established itself as the de-facto standard for sharing data on the Semantic Web, adoption by the UMAP community has remained limited.

In this paper, we give a brief introduction to Linked Data and its relevance for user modeling and personalization in general, and for adaptive educational systems in particular. Further, we present the LinkedUp project and discuss how the user modeling community can benefit from its activities.

2 Linked Data in a Nutshell

The simplest way to describe Linked Data [6] is that it is about using the Web architecture not only for documents, but also for data. The foundation of Linked Data is that data objects on the Web are identified by Web addresses (URIs),

which can be referenced by a Web link, similarly as one would do with Web documents. This basic principle for easy discovery, reference, access and reuse of Web data is now gaining significant momentum in many different areas.

Governments (most notably in the US[3] and the UK[4]) are leading open data initiatives; they provide information about aspects such as transport, environment, public spending and education. In addition, various more general-purpose datasets are being made available, such as the Geonames initiative[5], which makes it possible to exploit information on geographical places in the world. One of the most referred to sources of open Web data is DBpedia[6], a Linked Data version of Wikipedia. Finally, Linked Data is more and more used by universities and other education institutions (see [1] and [2] for details). These various initiatives make the Web of Linked Data an invaluable resource that connects and gives access to information from an incredibly vast number of domains.

3 Relevance for Personalization

There are various examples of research on adaptive service selection and composition for personalization in the UMAP community. [7] discuss how semantically rich descriptions of available services complement manual composition and AI planning techniques. The Personal Reader [4] is an example system that employs Semantic Web technologies for extending personalization to resources from external repositories. A more recent paper [3] presents a proof of concept on how linked data principles and service-orientation resolve the integration issues for sharing and discovering educational resources.

Linked data is also considered a base technology for the integration of data for (educational) data mining, and gains increasing attention in the Learning Analytics community[7], which focuses on the measurement, collection, analysis and reporting of data about learners and their contexts, for understanding and optimizing learning and the environments in which it occurs [5].

4 The LinkedUp Project

The LinkedUp project[8] is an FP7 Support Action that pushes forward the exploitation and adoption of open data available on the Web, in particular by educational institutions. To address these goals, LinkedUp provides a range of activities, including the LinkedUp Data Challenge[9]. The goal of the Challenge is to identify and promote innovative applications and tools that exploit large-scale

[3] http://www.data.gov/
[4] http://data.gov.uk/
[5] http://www.geonames.org/
[6] http://dbpedia.org/
[7] As illustrated by the Learning Analytics and Linked Data Workshop at LAK 2012
[8] http://linkedup-project.eu/
[9] http://linkedup-challenge.org/

Web data in educational scenarios. An important focus of the challenge will be on adaptive service selection and other personalization techniques.

To support the challenge, the LinkedUp support action collects and catalogs data explicitly related to education, as well as related data that may be relevant, including useful Web media, user-generated content, Web lectures or academic publications. The data is made available through the Linked Education catalog[10] as well as through a data endpoint[11], where a SPARQL endpoint provides access to VoID[12] descriptions of currently included datasets.

5 Summary and Outlook

There is a wealth of useful material on the Web that can be used in education, ranging from slides, tutorials and online courses to Wikipedia articles and YouTube videos. The LinkedUp challenge aims to find ways to link and mash up educational and cross-domain linked and open data to provide novel applications for education. LinkedUp will catalog and curate open Web data, and create a reusable evaluation framework for Open Web Data applications, in particular in the educational domain. In addition, LinkedUp collects applications and use-cases that will help the education sector to capitalize on open Web data.

Acknowledgments This work is partly funded by the European Union under FP7 Grant Agreement No 317620 (LinkedUp).

References

1. d'Aquin, M., Adamou, A., Dietze, S.: Assessing the educational linked data landscape. In: Proceedings Web Science (2013)
2. Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H.Q., Giordano, D., Marenzi, I., Nunes, B.P.: Interlinking educational resources and the web of data a survey of challenges and approaches (2013)
3. Dietze, S., Yu, H.Q., Giordano, D., Kaldoudi, E., Dovrolis, N., Taibi, D.: Linked education: interlinking educational resources and the web of data (2012)
4. Dolog, P., Henze, N., Nejdl, W., Sintek, M.: The personal reader: Personalizing and enriching learning resources using semantic web technologies. In: Adaptive Hypermedia and Adaptive Web-Based Systems. pp. 85–94. Springer (2004)
5. Ferguson, R.: The state of learning analytics in 2012: A review and future challenges. Knowledge Media Institute, Technical Report KMI-2012-01 (2012)
6. Heath, T., Bizer, C.: Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology 1(1), 1–136 (2011)
7. O'Keeffe, I., Conlan, O., Wade, V.: A unified approach to adaptive hypermedia personalisation and adaptive service composition. In: Adaptive Hypermedia and Adaptive Web-Based Systems. pp. 303–307. Springer (2006)

[10] `http://datahub.io/group/linked-education`
[11] `http://data.linkededucation.org/linkedup/catalog/`
[12] `http://www.w3.org/TR/void/`

Information Retrieval and User-Centric Recommender System Evaluation

Alan Said[†], Alejandro Bellogín[†], Arjen De Vries[†], Benjamin Kille[⋆]

CWI, The Netherlands[†], TU Berlin, Germany[⋆]
{alan, alejandro.bellogin, arjen.de.vries}@cwi.nl[†], kille@dai-lab.de[⋆]

Abstract. Traditional recommender system evaluation focuses on raising the accuracy, or lowering the rating prediction error of the recommendation algorithm. Recently, however, discrepancies between commonly used metrics (e.g. precision, recall, root-mean-square error) and the experienced quality from the users' have been brought to light. This project aims to address these discrepancies by attempting to develop novel means of recommender systems evaluation which encompasses qualities identified through traditional evaluation metrics and user-centric factors, e.g. diversity, serendipity, novelty, etc., as well as bringing further insights in the topic by analyzing and translating the problem of evaluation from an Information Retrieval perspective.

Keywords: Recommender Systems, Evaluation, Information Retrieval

1 Introduction

The project is framed in the Recommender Systems (RS) field. The aim of RSs is to assist users in finding their way through huge databases and catalogues, by filtering and suggesting relevant items, taking into account the users' preferences (i.e., tastes, priorities, etc.).

2 Novel Methods for Recommender System Evaluation

Over the last two decades, a vast amount of research in RS has lead to great progress in terms of prediction accuracy [1]. Today, the majority of the work on RS is based on top-n recommendation or rating prediction; the former requires bi/unary interaction data between users and items, whereas the latter requires a dataset with ratings. This type of evaluation is also common in information retrieval (IR) systems [3].

2.1 User-centric Evaluation

Both top-n and rating prediction-based evaluation build on several assumptions which could potentially have negative effects on recommendation algorithms tuned solely on these evaluation metrics [4, 1, 5, 6]. These are:

– there is an absolute ground truth which the RS should attempt to identify,
– users are primarily interested in the items which have received the highest ratings,
– higher top-n accuracy or lower rating error levels translate to a higher perceived usefulness from the users.

Recent work has, however, shown that these assumptions are not always true in the RS context, e.g. [7–9]. In specific cases, the assumptions are detrimental to the users' perceived quality.

The main focus of this sub-project is to analyze the discrepancies between *offline* and *online* evaluation, and to gain insights into the subjective qualities of various recommendation algorithms and their specific qualities. With this in mind, there are several goals we will strive to achieve:

– Analyze the correlation of IR-related evaluation metrics and user-centric concepts such as diversity, novelty and other non-quantifiable RS aspects.
– Identify whether there exists a correlation between the properties of items that are regarded as false recommendations in traditional (offline) IR and RS evaluation settings and true (high quality) recommendations in user-centric (online) evaluation, taking factors such as serendipity and usefulness into consideration.
– Evaluate whether metrics from research areas outside of IR and RS (e.g. signal processing, economics) can estimate sought for qualities better than the currently used RS and IR metrics.
– Improve the understanding of which evaluation metrics should be applied to RSs in different contexts, using different algorithms, data sets, and other system-specific features.

For these purposes, we will use a variety of data sets from the multimedia domain, ranging from publicly available, e.g. Movielens,Last.fm, as well as proprietary from other related services, e.g. Filmtipset, Moviepilot, etc. The variety of data will ensure the general applicability of the research results. Additional data will be collected through user studies and surveys.

2.2 Information Retrieval-based Evaluation

RS are usually considered as a special case of IR systems, specifically, one where no query is given and the information to be retrieved has to be inferred from previous user experiences. For this reason, some of the models and theories developed in IR have already been translated to RS, such as the Vector Space Model and the Probability Ranking Principle [10].

There are, however, several gaps in the understanding of RS as personalized IR systems, such as the need of formal methods to introduce implicit and contextual feedback in the recommendations and the lack of a proper evaluation framework.

A strong link between IR and RS has already been shown in our previous research (see [11, 12]), where we adapted to RSs different techniques proposed

in IR to predict the performance of a system. A natural next step is to explore how the evaluation in RS may benefit from extending this analogy between IR and recommendation, and applying more retrieval methodologies to recommendation.

More specifically, in this sub-project, we aim to exploit IR concepts, algorithms, and methodologies for recommendation .RS has a well known tradition in integrating contextual information which could be, in turn, transferred from RS to IR and investigate how such methods could be integrated in contextual IR and whether some benefits could be found by creating such links.

With this goal in mind, there are several objectives we aim to achieve:

- Analyze how evaluation in recommendation should be performed to obtain a general methodology that would result in interpretable and comparable results for the community, mainly by adapting and integrating models and metrics from IR.
- Identify if there is any correlation between the metrics typically used in offline experiments (e.g., precision) and those pervasive in real applications, more useful from a business point of view (such as the click through and conversion rates).
- Bring the models and theories used in context-aware recommendation to contextual IR and vice versa, in such a way that further interactions between these two areas could be found.
- Exploit implicit information for RS in novel ways based on current research from the use of search logs and other implicit sources of information in IR.

3 Current Results

At the moment of writing, we have obtained positive results regarding some of the aspects of this research project. Specifically, a workshop on reproducibility and replication in recommender evaluation has been accepted at the 2013 ACM RecSys conference. We are also working on revisiting several classic recommendation techniques in order to evaluate them under a common framework based on an IR-inspired recommendation method previously proposed in [14]. A novel evaluation protocol has also been researched specifically for RS. Additionally, we are researching different features that could help in the understanding of why some similarity functions perform better when applied within a recommendation strategy. Finally, we have analyzed different sources of implicit and explicit popularity scores in order to exploit them in a recommendation context; this study was recently accepted for publication as [15].

4 Acknowledgements

This work was carried out during the tenure of an ERCIM “Alain Bensoussan” Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no.246016.

References

1. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1) (January 2004) 5–53
2. Amatriain, X., Basilico, J.: Netflix recommendations: Beyond the 5 stars (part 1) – the netflix tech blog. http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html (retrieved May 12, 2012) (April 2012)
3. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1999)
4. Hill, W., Stead, L., Rosenstein, M., Furnas, G.: Recommending and evaluating choices in a virtual community of use. In: Proceedings of the SIGCHI conference on Human factors in computing systems. CHI '95, New York, NY, USA, ACM Press/Addison-Wesley Publishing Co. (1995) 194–201
5. Amatriain, X., Pujol, J.M., Oliver, N.: I like it... i like it not: Evaluating user ratings noise in recommender systems. In: Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization: formerly UM and AH. UMAP '09, Berlin, Heidelberg, Springer-Verlag (2009) 247–258
6. Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A., Turrin, R.: Comparative evaluation of recommender system quality. In: Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems. CHI EA '11, New York, NY, USA, ACM (2011) 1927–1932
7. Said, A., Fields, B., Jain, B.J.: User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In: Proceedings of the ACM 2013 conference on Computer Supported Cooperative Work, New York, NY, USA, ACM (2013)
8. Said, A., Jain, B., Narr, S., Plumbaum, T.: Users and noise: The magic barrier of recommender systems. In Masthoff, J., Mobasher, B., Desmarais, M., Nkambou, R., eds.: User Modeling, Adaptation, and Personalization. Volume 7379 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2012) 237–248 10.1007/978-3-642-31454-4_20.
9. Said, A., Jain, B.J., Narr, S., Plumbaum, T., Albayrak, S., Scheel, C.: Estimating the magic barrier of recommender systems: a user study. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. SIGIR '12, New York, NY, USA, ACM (2012) 1061–1062
10. Wang, J., Robertson, S., de Vries, A., Reinders, M.: Probabilistic relevance ranking for collaborative filtering. Information Retrieval **11**(6) (December 2008) 477–497
11. Bellogín, A.: Recommender System Performance Evaluation and Prediction: an Information Retrieval Perspective. PhD thesis, Universidad Autonóma de Madrid, Spain (November 2012)
12. Bellogín, A.: Performance Prediction in Recommender Systems. In Konstan, J., Conejo, R., Marzo, J., Oliver, N., eds.: User Modeling, Adaption and Personalization. Volume 6787 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, Berlin, Heidelberg (2011) 401–404
13. Said, A.: Researching Novel Methods for Recommender System Evaluation. In: UMAP (extended proceedings), Rome (2013)
14. Bellogín, A., Wang, J., Castells, P.: Bridging memory-based collaborative filtering and text retrieval. (2012) 1–28
15. Bellogín, A., de Vries, A., He, J.: Artist popularity: do web and social music services agree? In: Int. Conf. on Weblogs and Social Media (ICWSM), Boston (2013)

techplay.mobi: A Technological Framework for Developing Affective Inclusive Personalized Mobile Serious Games to Enrich Learning Competences

Olga C. Santos[1], Mar Saneiro[1], Emmanuelle Gutiérrez y Restrepo[1], Jesus G. Boticario[1], Elena Del Campo[2], Raúl Cabestrero[3], Pilar Quirós[3], Sergio Salmeron-Majadas[1], Emmanuelle Raffenne[1], Emanuela Mazzone[1]

[1]Artificial Intelligence Department, Computer Science School, UNED, Spain
[2] Developmental and Educational Department, Faculty of Psychology, UNED, Spain
[3] Basic Psychology II Department, Faculty of Psychology, UNED, Spain
{ocsantos, marsaneiro, egutierrez, jgb}@dia.uned.es, mcampo@psi.uned.es, {rcabestrero, pquiros}@psi.uned.es, ssalmeron@bec.uned.es, {eraffenne, emazzone}@dia.uned.es,

Abstract. *techplay.mobi* is a research project aimed at building a platform for developing mobile serious games. The novelty is to support the design of games that consider affective features during the game interaction, provide personalized responses according to users' interactions, comply with accessibility requirements and focus on improving psycho-educational competences and on promoting critical thinking.

Keywords: Serious games, psycho-educational competences, affective states.

1 Introduction

Mobile devices are integrated into users' daily life, including their learning activities. Serious games (usually defined as 'games for purposes other than entertainment' [1, 2] embracing all aspects of education –teaching, training, advertising and informing– and at all ages [3]) follow suit. This type of games can support users to get a more motivating and efficient learning experience, delivering multimodal information and disseminating knowledge in a socially complex environment [4] by impacting on the users' affective, cognitive and metacognitive capabilities [5] (e.g. selecting and codifying relevant information, decision making, problem solving, communicating, monitoring). Often, users face difficulties in acquiring and applying knowledge because their learning, affective and critical thinking strategies are functioning in a poor or inadequate way. In these cases, learners face a gap between their skills and their performance. This distance can affect any user in any learning context (e.g. traditional, virtual...) and involve diverse cognitive and meta-cognitive processes (e.g. attention, reasoning, memory, communication, reading comprehension, planning, social abilities, emotional management).

From our experiences in past research projects (e.g. EU4ALL) effective and efficient virtual learning environments need to be adapted according with the user' physi-

cal, cognitive and perceptual abilities and limitations by integrating different types of adaptation, which allow the user to accomplish the tasks posed to them [6]. Serious games, like any educational resource, should consider those issues. Thus, *techplay.mobi* focuses on modeling users' needs and preferences, types of educational scenarios and device characteristics to support educators in designing serious games that can improve user's competences through their interaction with an inclusive and personalized multimodal mobile learning environment.

2 About *techplay.mobi* and UNED

Technological Framework for Developing Affective Personalized Serious Games to Enrich the Integral Human Development (IPT-430000-2011-1721) research project involves four partners (creativ IT, oneclick, AIJU and UNED) and aims at developing a platform to facilitate and support the design, creation and packaging of serious games with integrated applications, promoting the integral human development by improving users psycho-educational strategies, affective competences and critical thinking abilities [7].

There are few authoring platforms that enable non-expert users to develop place-based or narrative gaming activities designed for teaching and learning for mobile context (e.g. *ARIS* [8], *e-adventure* [9], *Collage* [10], *Infantium* [11]). However, these platforms do not deal with the integral human development considered by *techplay.mobi* (i.e. accessibility and personalization regarding user affective state, learning needs and competences), which is to be integrated into the serious games learning flow, activities and resources.

techplay.mobi covers both technological and educational goals. At UNED, we mainly focus on the educational goals and address the following objectives: 1) develop a support framework to create serious games by combining technology and education involving the user in a fun formative learning process, 2) define the appropriate instructional design for serious games, where selected contents, activities and resource are linked to foster the user affective, learning and critical thinking strategies when functioning in a poor or inadequate way, 3) facilitate inclusive personalized serious games adaptations taking into account the user functional diversity (cognitive, physical, perceptual) by providing usable and accessible interfaces, personalized learning flows, alternative contents, resources and activities, and 4) validate the benefits of this support framework integrated in inclusive personalized and affective serious games on the users' learning experience (e.g. achievement, satisfaction, motivation, engagement, generalization, etc. of acquired knowledge in real life contexts).

3 Ongoing work and expected results

Serious games should combine learning strategies guiding students' exploration of learning content with entertainment enhancing learning [12]. Specifically learning strategies integrated into serious games must facilitate active knowledge construction, practicing key learning skills such as problem solving, decision-making and collabo-

ration [13]. From lessons learned at EU4LL [14], MAMIPEC [15] and ALTERNATIVA [16] projects on learning personalization taking into account the user profile (learning style, learning and affective level of competences, accessibility preferences, functional needs) and specific educational context features, at UNED we have focused on understanding how psycho-educational competences, emotional states management and critical thinking processes can be improved through mobile accessible videogames where formal and informal education and entertainment need to be matched. To illustrate this, Table 1 shows how learning competences, achieved affective states and critical thinking abilities can be integrated into a serious game learning flow for problem solving.

Table 1. Learning competences, affective states achieved and critical thinking abilities.

Serious games activity: Problem solving			
Step	**Learning competences**	**Affective state**	**Critical thinking abilities**
Identifying problems	Establishing relations, how, when and why. Identifying concepts. Grouping information according to target criteria	Interested	Development of a realistic view of the problem.
Selecting key information	Representing a structure, establishing main ideas, secondary information	Excited	Avoidance of ambiguous and useless information, Additional information search.
Developing an answer	Activating previous knowledge, focusing attention on problem issues, establishing order in solving process	Confident, relaxed	Initiation of positive actions
Evaluating the answer	Collaborating with others, asking orientations, reformulating problem from other perspective, facing critics	Pleased, satisfied	Improvement of believes about one-self and others, Improvement of personal and social adjustment, Acceptance of challenges

On-going work focuses on 1) studying the state of the art on which and how learning and affective strategies can be integrated into serious games, 2) determining how the EU4ALL outcomes [14] can be reoriented and adapted to mobile accessible videogames to take advantage of mobile devices capabilities (e.g. spatial inclination provided by internal accelerometer), 3) identifying, through varied sources of information (e.g. questionnaires, interviews, data mined from sensors such as eye trackers, biofeedback devices and interaction patterns [15]), relevant users' needs and preferences (educational, affective and accessibility) to build an open standards-based user model extending existing specifications (e.g. IMS LIP [17], IMS AfA [18] and W3C Emotion ML [19]) which informs the instructional design of serious games.

The expected result will be the definition of a synergy matrix that maps a) data collected about psycho-educational competences, affective states and accessibility needs and preferences detected, with b) technological requirements to design, create and package mobile accessible multiplatform videogames for the most relevant mobile operating systems (Android, iOS). Thus, this matrix will guide the framework design that support the creation of serious games that promote the integral human development by improving psycho-educational strategies, affective competences and critical thinking abilities. The resulting developing platform for mobile serious games will be evaluated in several scenarios involving educators and serious game players.

Acknowledgements

The authors would like to thank the European Regional Development Fund and the Spanish Ministry of Economy and Competence for funding *techplay.mobi* project.

References

1. Corti, K. (2006) Games-based Learning; a serious business application. PIXE Learning.
2. Zyda, M. (2005) From visual simulation to virtual reality to games. IEEE Computer, 38 (9), 25-32.
3. Michael, D., Chen, S. (2006) Serious games: Games that educate, train, and inform. Boston, MA.: Thomson Course Technology.
4. Gee, J.P. (2009) Deep learning properties of good digital games: How far can they go? In Ritterfeld, U., Cody, M., Vorderer, P.(Eds). Serious games: Mechanisms and effects, 67-82.
5. ELSPA (2006) Unlimited learning: Computer and video games in the learning landscape. Entertainment and Leisure Software Publishers Association, ELSPA.
6. Campo del, E., Saneiro, M., Santos, O.C., Boticario, J.G. (2010) Psycho-educational support for students with disabilities in higher education, applied through a recommender system integrated in a virtual learning environment. International Journal of Developmental and Educational Psychology, 237-247.
7. techplay.mobi Project. Website: http://techplay.creativit.com
8. Augmented Reality and Interactive Storytelling (ARIS). Website: http://arisgames.org/
9. Blanco, A., Torrente, J., Marchiori, J., Martínez-Ortiz, I., Moreno-Ger, P., Fernández-Manjón, B. (2012) A framework for simplifying educator tasks related to the integration of games in the learning flow. Education Technology & Society 15 (4), 305-318. Website: http://e-adventure.e-ucm.es
10. Collaborative Learning Platform Using Game-like Enhancements (COLLAGE). Website: http://www.ea.gr/ep/collage/main.asp
11. Infantium Web Site: http://www.infantium.com/en/developers/
12. Ching-Sheng,Y., Shu-Hua, L.,Kuo-Hua, W. (2012) A comparison of learning effectiveness among serious games with varying degrees of playability. Proceedings of ICEE.
13. Arnab, S., Berta, R., Earp, J., Freitas, S., Popescu, M., Romero, M., Stanescu, I., Usart, M. (2012) Framing the Adoption of Serious Games in Formal Education. Electronic Journal of e-Learning, 10 (2), 159-171.
14. Boticario, J.G., Rodríguez-Ascaso, A., Santos, O.C., Raffenne, E., Montandon, L., Roldán, D., Buendia, F. (2012) Accessible Lifelong Learning at Higher Education: Outcomes and Lessons Learned at two Different Pilot Sites in the EU4ALL Project. Journal of Universal Computer Science, 18 (1), 62-85.
15. Santos, O.C., Salmeron-Majadas, S., Boticario, J.G. (2013) Emotions Detection from Math Exercises by combining several Data Sources. LNAI, 7926, 742–745.
16. Gutiérrez y Restrepo, E., Benavidez, C., Gutiérrez, H. (2012). The Challenge of Teaching to Create Accessible Learning Objects to Higher Education Lecturers. Proceeding of 4th International Conference on Software Development for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2012), 14, 371–381.
17. IMS Learner Information Package. Website: http://www.imsglobal.org/profiles/
18. IMS Access for All. Website: http://www.imsglobal.org/accessibility/
19. W3C Emotion Markup Language 1.0. Website: http://www.w3.org/TR/emotionml/